MÉTODOS NO PARAMÉTRICOS

EN CIENCIAS VETERINARIAS

Aplicaciones prácticas para el análisis de datos sin supuestos paramétricos



Autores/as:

Víctor Alejandro Bosquez Barcenes
Washington Fernando Carrasco Sangache
Verónica Lourdes Carrasco Sangache
Jorge Jagger Segura Ochoa

ISBN: 978-9907-818-40-6

MÉTODOS NO PARAMÉTRICOS EN CIENCIAS VETERINARIAS

Métodos No Paramétricos en Ciencias Veterinarias

Aplicaciones prácticas para el análisis sin supuestos paramétricos

Autores:

Víctor Alejandro Bosquez Barcenes

Orcid*:* https://orcid.org/0000-0001-7679-6023

Correo: abosquez@ueb.edu.ec

Filial: Universidad Estatal de Bolívar

Washington Fernando Carrasco Sangache

Orcid: https://orcid.org/0000-0001-6634-0612

Correo*:* wcarrasco@ueb.edu.ec

Filial: Universidad Estatal de Bolívar

Verónica Lourdes Carrasco Sangache

Orcid: https://orcid.org/0009-0000-2606-3171

Correo: lourdes.carrasco@espoch.edu.ec

Filial: Espoch

Jorge Jagger Segura Ochoa

Orcid: https://orcid.org/0009-0004-2506-4833

Correo: jsegura@ueb.edu.ec

Filial: Universidad Estatal de Bolívar

Editorial “ACACFESA SAS”

La presente obra fue revisada por 2 pares académicos externos ciegos conforme al proceso editorial de ACACFESA SAS.

Los rigurosos procedimientos editoriales de ACACFESA SAS garantizan la selección de manuscritos por sus aportes significativos al conocimiento y cualidades científicas.

Editorial: ACACFESA SAS.

Sello editorial: (978-9907-818)

Teléfono: (+593) 998955446

Web: https://acacfesa.com/editorial/index.php/1

ISBN: 978-9907-818-40-6

Doi:

Año: Agosto 2026

Aviso Legal

El contenido de esta obra, que incluye ilustraciones, textos, tablas, gráficos, cuadros y referencias bibliográficas, es responsabilidad única del autor o autores. Lo que se ha dicho, los criterios y los datos no reflejan necesariamente la posición institucional ni el pensamiento de la Editorial ACACFESA SAS.

Derechos de Autor ©

Este documento se publica conforme los términos y condiciones de la EDITORIAL ACACFESA SAS

Esta obra está bajo una licencia internacional



Creative Commons Atribución-NoComercial-CompartirIgual 4.0

La "Editorial ACACFESA SAS " y todos sus autores tienen la propiedad única de los derechos de autor y de propiedad intelectual e industrial relacionados con el contenido de esta publicación. La reproducción total o parcial de esta obra, su almacenamiento en sistemas informáticos, su tratamiento digital y cualquier tipo de distribución, transmisión o comunicación pública por medios electrónicos, ópticos, mecánicos, químicos, fotográficos o de grabación están prohibidos. Esto se encuentra bajo las sanciones que establece la legislación vigente a menos que se cuente con la autorización previa y por escrito de los titulares del copyright.

Queda excluido únicamente el uso con fines académicos o de investigación científica, siempre que no busque objetivos comerciales y se ejecute sin remuneración, debiendo mencionarse en todo momento a la fuente editorial correspondiente. Los autores son los únicos responsables de las opiniones expresadas en los diferentes capítulos, las cuales no necesariamente representan el punto de vista institucional de la editorial.

Constancia de Arbitraje

La Editorial ACACFESA SAS, certifica que este libro es el resultado de un estudio llevado a cabo por los autores, el cual fue evaluado por jurados expertos bajo el sistema de doble ciego, tanto en contenido como en forma. Asimismo, se llevó a cabo un análisis del método de investigación, paradigma y enfoque; a partir de la matriz epistémica adoptada por los autores, utilizando las normas APA, Séptima Edición y el proceso de antiplagio en línea turniting edu para asegurar que la obra fuera científica.

DEDICATORIA

A mi hijo Miguel, porque tu sonrisa es el motor que impulsa cada uno de mis pasos y la razón más poderosa para no rendirme jamás. A mis padres, quienes con su amor, sacrificio y ejemplo me enseñaron que la educación es el camino más noble para transformar la vida. Gracias por creer en mí incluso cuando el camino parecía incierto, por cada desvelo compartido y por haberme dado las raíces firmes desde las cuales hoy puedo crecer. Este libro es, ante todo, un tributo a ustedes tres, porque sin su presencia nada de esto tendría sentido.

Víctor Alejandro

A mis estudiantes:

Que estas páginas sean el comienzo de su aprendizaje, dándoles la base, la confianza y la curiosidad necesaria para entender el porqué de cada resultado. Espero que encuentren aquí no solo conocimiento, sino una invitación a maravillarse con el arte del análisis y la interpretación de los datos. Con sincero cariño y estima.

Washington Fernando

Dedico este trabajo a mi familia, por ser mi apoyo constante y acompañarme en cada etapa de mi formación personal y profesional. A quienes han confiado en mí y han contribuido, con sus enseñanzas y experiencias, a fortalecer mi vocación. De manera especial, a los animales, razón fundamental de mi profesión, cuya vida, bienestar y cuidado inspiran cada día mi compromiso con la Medicina Veterinaria. A mis estudiantes, colegas y maestros, quienes me motivan a continuar aprendiendo, investigando y compartiendo conocimientos con responsabilidad, ética y vocación de servicio.

Verónica Lourdes

A mi esposa, compañera incansable de este camino, por su paciencia, su amor y su apoyo silencioso en cada noche de desvelo y en cada meta que parecía lejana. Sin tu presencia a mi lado, estas páginas no existirían. A mis padres, por haberme enseñado con su ejemplo que el esfuerzo y la constancia son los pilares de todo logro. Gracias por creer en mí desde el principio, por cada

sacrificio que hicieron para que yo pudiera llegar hasta aquí. Este libro es el reflejo de todo lo que ustedes han sembrado en mí con amor.

Jorge Jagger

PROLOGO

Quien haya dirigido tesis de grado o posgrado en ciencias veterinarias en alguna universidad latinoamericana sabe que hay una escena que se repite con variantes menores, pero con una regularidad casi predecible. El tesista llega con sus datos conteos de huevos de parásitos, puntajes de condición corporal, tiempos de recuperación posanestésica, calificaciones de cojera en una escala de cinco puntos y los introduce en un software estadístico con la intención de aplicar una prueba t o un ANOVA. Alguien le dijo alguna vez que esas eran las pruebas "estándar", y nadie le explicó con suficiente claridad que la palabra estándar no significa universal. Los datos se resisten. El histograma muestra una distribución que se aplasta contra el cero y se arrastra hacia la derecha con una cola larga. La prueba de Shapiro-Wilk arroja un valor p de 0.003. El tesista, presionado por los plazos y quizá por la inercia de hacer lo que siempre se ha hecho, aplica la prueba paramétrica de todos modos, obtiene un valor p, escribe sus conclusiones y entrega el documento.

Ese documento contiene un error que no es de tipeo ni de redacción. Es un error de lógica estadística: la inferencia descansa sobre supuestos que los datos no cumplen, y el valor p que se reporta puede estar lejos del que correspondería. A veces la conclusión coincide con la que se habría obtenido por la vía correcta. Otras veces no. Y el problema es que el investigador no tiene forma de saber en cuál de los dos casos se encuentra.

Este libro nace de esa realidad, no de una inquietud teórica abstracta. Nace de haber visto, a lo largo de años de docencia e investigación en universidades ecuatorianas y latinoamericanas, cómo los métodos no paramétricos se enseñan con frecuencia como una nota al pie, un apéndice menor del curso de bioestadística que se menciona en la última clase del semestre, cuando ya nadie presta demasiada atención. Y nace también de la convicción de que esos métodos merecen un tratamiento serio, extenso y anclado en los problemas reales que enfrenta el veterinario investigador de esta región.

La investigación veterinaria en América Latina trabaja con condiciones que hacen de los métodos no paramétricos no una alternativa de segunda categoría, sino una necesidad de primer orden. Las muestras son pequeñas porque los animales grandes cuestan, porque los ensayos clínicos con bovinos o equinos no pueden replicar los tamaños de los estudios con roedores de laboratorio, y porque las restricciones éticas y presupuestarias limitan el número de unidades experimentales.

Los datos son frecuentemente ordinales: escalas de dolor, grados de lesión histopatológica, clasificaciones del estado sanitario de un hato. Los conteos huevos por gramo de heces, unidades formadoras de colonias, garrapatas por animal producen distribuciones tan alejadas de la campana gaussiana que pretender analizarlas con pruebas paramétricas es forzar una herramienta donde no cabe. Los estudios de campo en fincas de Chimborazo, del altiplano boliviano, de los valles interandinos de Colombia o de las llanuras chaqueñas argentinas generan datos con ceros inflados, valores extremos y asimetrías que no desaparecen ni con transformaciones logarítmicas.

Frente a ese panorama, los autores de esta obra han tomado decisiones que conviene hacer explícitas porque definen su carácter.

La primera es que este no es un libro de recetas. No se limita a decir "si sus datos no son normales, use Kruskal-Wallis". Antes de cada prueba, se explica por qué funciona: qué hace con los datos, qué información extrae de los rangos o de los signos, cuál es la lógica que sustenta el estadístico de prueba. Un investigador que entiende la mecánica de una prueba no solo la aplica mejor, sino que sabe reconocer cuándo no debe aplicarla.

La segunda decisión es que los métodos no paramétricos se presentan siempre en diálogo con sus equivalentes paramétricos, no en oposición a ellos. Cada capítulo incluye una comparación explícita: qué se gana y qué se pierde en potencia estadística al elegir la vía no paramétrica, en qué condiciones ambas pruebas dan conclusiones idénticas, y en qué situaciones divergen de manera que la elección del método cambia la interpretación del estudio. Esa honestidad es necesaria porque la estadística no tiene herramientas perfectas; tiene herramientas más o menos adecuadas según la estructura de los datos y la pregunta que se formula.

La tercera decisión es que cada concepto se ancla en un problema veterinario concreto. Los ejemplos no son genéricos ni inventados para facilitar el cálculo. Son situaciones que un veterinario reconoce: la evaluación de un protocolo antiparasitario en ovinos Corriedale de la sierra ecuatoriana, la comparación de protocolos analgésicos en perras sometidas a ovario histerectomía, la infestación por garrapatas en bovinos de distintos genotipos, la sobrevivencia de vacas con anaplasmosis tratadas con imidocarb frente a oxitetraciclina, la composición de comunidades parasitarias en sistemas productivos de páramo versus valle. Los datos simulados se basan en

rangos publicados en la literatura, con razas, fármacos, patologías y localidades que pertenecen a la geografía productiva de la región.

La cuarta decisión, quizá la que más distingue a este libro de otros textos de estadística no paramétrica disponibles en español, es la integración completa con R. No se trata de un apéndice con fragmentos de código sueltos. Cada prueba se implementa con código completo, comentado línea por línea, con la salida de consola reproducida e interpretada valor por valor. Cuando existen varias funciones en R para la misma prueba y las hay, con diferencias reales en el cálculo del valor p, se muestran todas y se explica por qué dan resultados distintos. El estudiante puede copiar el bloque de código, ejecutarlo en su computadora y obtener exactamente lo que el libro describe. Esa reproducibilidad no es un detalle cosmético; es la base de la práctica estadística responsable.

Los ocho capítulos del libro cubren un arco que va desde los fundamentos y la verificación de supuestos hasta las tendencias actuales en métodos bayesianos no paramétricos y aprendizaje automático aplicado a datos veterinarios. El primer capítulo establece cuándo y por qué los métodos paramétricos fallan, y proporciona las herramientas de diagnóstico gráficas y formales para tomar esa decisión con evidencia. Los capítulos segundo, tercero y cuarto desarrollan las pruebas clásicas para una muestra, dos muestras, múltiples grupos, tablas de contingencia y correlación, con sus respectivas comparaciones post hoc y medidas de tamaño del efecto. El quinto capítulo aborda el análisis de sobrevivencia y los métodos de remuestreo, dos familias de técnicas que comparten la filosofía de extraer información sin imponer supuestos distribucionales. El sexto capítulo se dedica a los datos ordinales, las escalas de evaluación clínica y la potencia estadística, un tema que la investigación veterinaria latinoamericana necesita discutir con más franqueza. El séptimo capítulo presenta los métodos multivariados basados en distancias PERMANOVA, ANOSIM, NMDS, la regresión no paramétrica y los modelos aditivos generalizados. El octavo cierra con las normas de reporte según APA séptima edición y un panorama de hacia dónde se mueven los métodos analíticos en la disciplina.

Cada capítulo sigue una estructura uniforme que incluye objetivos de aprendizaje con verbos operativos, introducción con un problema veterinario real, desarrollo teórico-práctico progresivo, implementación en R, un estudio de caso integrador con análisis completo, una sección de errores frecuentes escrita desde la experiencia directa de quienes han corregido esos errores en tesis y

manuscritos, y ejercicios de autoevaluación en tres niveles de dificultad con sus respuestas o rúbricas. Esa uniformidad no es rigidez: es una decisión pedagógica para que el lector sepa siempre dónde buscar lo que necesita, ya sea la fórmula con sus símbolos definidos, el código en R para replicar el ejemplo, o la advertencia sobre el error que no debe cometer.

Una aclaración necesaria sobre las referencias: los autores han priorizado artículos publicados en revistas veterinarias indexadas en Scopus, Web of Science y SciELO de los últimos diez años, complementados con los textos clásicos de Conover, Siegel y Castellan, Hollander y Wolfe, y Daniel, que siguen siendo fuentes insustituibles para la fundamentación teórica.

Este libro está dirigido a estudiantes de semestres avanzados de Medicina Veterinaria y Zootecnia, a maestrantes y doctorandos que enfrentan el análisis de sus datos de tesis, y a investigadores en ejercicio que necesitan un manual de consulta con respaldo teórico y aplicación práctica inmediata. No exige formación previa en métodos no paramétricos ni experiencia sólida en R, aunque sí asume conocimientos de estadística descriptiva y nociones básicas de pruebas de hipótesis. Es, en suma, el libro que los autores habrían querido tener a mano cuando empezaron a enseñar bioestadística y a dirigir tesis en las que los datos nunca se portaban como la teoría prometía.

Que cumpla su propósito.

Víctor Alejandro Bosquez Barcenes

Washington Fernando Carrasco Sangache

Verónica Lourdes Carrasco Sangache

Jagger Segura

ÍNDICE

DEDICATORIA vi

PROLOGO viii

Índice de tablas xx

Índice de figuras xxiv

INTRODUCCIÓN 1

Capítulo 1 6

Fundamentos de la estadística no paramétrica y verificación de supuestos 6

Introducción 7

Objetivos de aprendizaje 7

1.1. Qué asume un método paramétrico y por qué esos supuestos importan 8

1.2. Qué significa “no paramétrico” y qué no significa 11

1.3. Ventajas reales y limitaciones honestas 14

1.4. Evaluación de normalidad: métodos gráficos y pruebas formales 16

1.5. Evaluación de homocedasticidad 19

1.6. Qué hacer cuando los supuestos no se cumplen: transformaciones, métodos robustos, GLM y vía no paramétrica 21

1.7. El árbol de decisión: ¿paramétrico o no paramétrico? 24

1.8. Implementación en R: exploración de supuestos con datos veterinarios 26

1.9. Estudio de caso integrador: conteo de HPG en ovinos y ganancia de peso en pollos de engorde 33

1.10. Errores frecuentes y advertencias 37

1.11. Ejercicios de autoevaluación 39

1.12. Conclusiones del capítulo 43

Capítulo 2 45

Pruebas para una muestra, datos pareados y dos muestras independientes 45

Introducción 46

Objetivos de aprendizaje 46

2.1. Prueba de signos 47

2.2. Prueba de rangos con signo de Wilcoxon 50

2.3. Prueba de Wilcoxon para una muestra contra una mediana teórica 54

2.4. Comparación con la prueba t pareada 57

2.5. Prueba de Mann-Whitney U (Wilcoxon rank-sum) 59

2.6. El debate: ¿prueba de medianas o de dominancia estocástica? 62

2.7. Estimador de Hodges-Lehmann 64

2.8. Prueba de Brunner-Munzel para distribuciones con forma diferente 66

2.9. Comparación con la prueba t para muestras independientes 67

2.10. Tamaño del efecto: r de rango biserial, A de Vargha y Delaney 69

2.11. Implementación en R: wilcox.test(), wilcox_test() de coin, wilcox.exact() 71

2.12. Estudio de caso integrador 82

2.13. Errores frecuentes y advertencias 92

2.14. Ejercicios de autoevaluación 94

2.15. Conclusiones del capítulo 100

Capítulo 3 102

Comparación de múltiples grupos: Kruskal-Wallis, Friedman y comparaciones post hoc 102

Introducción 103

Objetivos de aprendizaje 103

Introducción con problema real 104

3.1. Prueba de Kruskal-Wallis: lógica, estadístico H y corrección por empates 105

3.2. Comparaciones post hoc para Kruskal-Wallis 115

3.2.2. Prueba de Conover-Iman 118

3.3. Comparación con ANOVA de un factor 121

3.4. Prueba de Friedman para diseños con bloques y medidas repetidas 124

3.5. Comparaciones post hoc para Friedman 130

3.6. Prueba de Quade como alternativa a Friedman 131

3.7. Coeficiente de concordancia de Kendall (W) 133

3.8. Comparación con ANOVA de medidas repetidas 135

3.9. Tamaño del efecto 136

3.10. Implementación en R 139

3.11. Estudio de caso integrador 156

3.12. Errores frecuentes y advertencias 160

3.13. Ejercicios de autoevaluación 162

3.14. Conclusiones del capítulo 168

Capítulo 4 169

Tablas de contingencia, asociación y correlación no paramétrica 169

Introducción 170

Objetivos de aprendizaje 170

Introducción con problema real 171

4.1. Chi-cuadrado de independencia y de bondad de ajuste 172

4.2. Corrección de Yates: cuándo aplicarla y controversia 175

4.3. Prueba exacta de Fisher 177

4.4. Prueba de McNemar para datos pareados categóricos 179

4.5. Prueba Q de Cochran 182

4.6. Medidas de asociación: odds ratio, riesgo relativo, V de Cramér, phi 184

4.7. Correlación de Spearman: lógica, cálculo y manejo de empates 187

4.8. Tau de Kendall: variantes τ-a, τ-b, τ-c y cuándo preferirlo sobre Spearman 191

4.9. Comparación con la correlación de Pearson 194

4.10. Correlación parcial no paramétrica 196

4.11. Implementación en R 197

4.12. Estudio de caso integrador 207

4.13. Errores frecuentes y advertencias 216

4.14. Ejercicios de autoevaluación 217

4.15. Conclusiones del capítulo 224

Capítulo 5 226

Análisis de Sobrevivencia y Métodos de Remuestreo 226

Introducción 227

Objetivos de Aprendizaje 227

5.1. Introducción: Cuando el Tiempo Hasta el Evento es la Variable de Interés 228

5.2. Conceptos Fundamentales del Análisis de Sobrevivencia 229

5.3. Estimador de Kaplan-Meier 233

5.4. Prueba de Log-Rank y Variantes 238

5.5. Supuestos del Análisis de Sobrevivencia No Paramétrico 241

5.6. Mención del Modelo de Cox como Extensión Semiparamétrica 243

5.7. Pruebas de Permutación 244

5.8. Bootstrap No Paramétrico 247

5.9. Supuestos y Limitaciones del Remuestreo 249

5.10. Implementación en R 250

5.11. Estudio de Caso Integrador 265

5.12. Errores Frecuentes y Advertencias 272

5.13. Ejercicios de Autoevaluación 273

5.14. Conclusiones del Capítulo 278

Capítulo 6 280

Datos ordinales, escalas de evaluación clínica y tamaño del efecto 280

Introducción 281

Objetivos de aprendizaje 281

6.1. Escalas ordinales en veterinaria: condición corporal, claudicación, dolor, diarrea neonatal, endometritis 282

6.2. Escala ordinal vs. escala de intervalo: la diferencia que cambia el análisis 284

6.3. Regresión ordinal: logit acumulativo y supuesto de odds proporcionales 286

6.4. Acuerdo entre evaluadores: kappa de Cohen, kappa ponderada, kappa de Fleiss 290

6.5. Tamaño del efecto: qué es, por qué reportarlo y medidas para cada prueba no paramétrica 294

6.6. Potencia estadística en métodos no paramétricos y eficiencia relativa asintótica 297

6.7. Determinación del tamaño muestral para estudios con análisis no paramétrico 299

6.8. Estimación de potencia por simulación 301

6.9. Implementación en R 303

6.10. Estudio de caso integrador 319

6.11. Errores frecuentes y advertencias 323

6.12. Ejercicios de autoevaluación 325

6.13. Conclusiones del capítulo 330

Capítulo 7 332

Métodos multivariados, regresión no paramétrica y modelos aditivos generalizados 332

Introducción 333

Objetivos de aprendizaje 333

Introducción: cuando un solo número no alcanza 334

7.1. Matrices de distancia y disimilitud: Bray-Curtis, Jaccard, euclidiana 335

7.2. PERMANOVA: lógica, pseudo-F de Anderson y verificación de dispersiones 338

7.3. ANOSIM: estadístico R global y comparación con PERMANOVA 342

7.4. NMDS: ordenación no paramétrica y estrés como medida de ajuste 344

7.5. Regresión local: LOESS/LOWESS 346

7.6. Modelos aditivos generalizados (GAM): de GLM a funciones suaves 348

7.7. GAM para distribuciones no normales: Poisson, binomial negativa, binomial 350

7.8. Diagnósticos del GAM: residuos, selección de k, concurvidad 352

7.9. Regresión cuantílica como alternativa para medianas y percentiles 353

7.10. Implementación en R 355

7.11. Estudio de caso integrador 370

7.12. Errores frecuentes y advertencias 382

7.13. Ejercicios de autoevaluación 383

7.14. Conclusiones del capítulo 388

Capítulo 8 390

Reporte de resultados y tendencias en la investigación veterinaria 390

Introducción 391

Objetivos de aprendizaje 391

8.1. Normas de reporte estadístico según APA 7.ª edición aplicadas a ciencias veterinarias . 392

8.2. Plantillas de reporte para cada prueba no paramétrica 395

8.3. Diseño de tablas de resultados: contenido mínimo y buenas prácticas 399

8.4. Figuras para métodos no paramétricos 402

8.5. Errores de reporte frecuentes en la literatura veterinaria 407

8.6. Introducción a métodos bayesianos no paramétricos 409

8.7. Machine learning como extensión no paramétrica 411

8.8. Datos masivos en veterinaria 414

8.9. La estadística no paramétrica clásica en la era del machine learning: complemento, no obsolescencia 416

8.10. Implementación en R 418

8.11. Estudio de caso integrador 429

8.12. Errores frecuentes y advertencias 438

8.13. Ejercicios de autoevaluación 439

8.14. Conclusiones del capítulo 443

Referencias Bibliográficas 444

Índice de tablas

Tabla 1 Comparación de la evaluación de supuestos entre dos conjuntos de datos veterinarios. 33

Tabla 2 Puntajes de dolor (CMPS-F) en 9 gatas antes de la cirugía y 6 horas después de ovario histerectomía con protocolo analgésico multimodal. 49

Tabla 3 Conteo de huevos por gramo (HPG) en 12 corderos Corriedale antes y 14 días después de la administración de albendazol. 53

Tabla 4 Puntajes de condición corporal (escala 1–5) en 15 vacas Holstein mestizas, primer tercio de lactancia. Finca en Cochabamba, Bolivia. 55

Tabla 5 Tiempo de cicatrización (días) de heridas quirúrgicas en perros, según tipo de apósito. 61

Tabla 6 Comparación entre la prueba t para muestras independientes y la prueba de Mann-Whitney U. 68

Tabla 7 Conteo de huevos por gramo de heces (HPG) de Haemonchus contortus a los 14 días posdosificación en cabritos tratados con tres antihelmínticos 109

Tabla 8 Asignación de rangos al conjunto combinado de observaciones. 110

Tabla 9 Comparación de tres procedimientos post hoc para Kruskal-Wallis. 120

Tabla 10 Comparación entre ANOVA de un factor y prueba de Kruskal-Wallis. 121

Tabla 11 Puntajes de calidad espermática asignados por tres andrólogos a 8 toros. 126

Tabla 12 Rangos dentro de cada bloque (toro). 127

Tabla 13 Comparación entre ANOVA de medidas repetidas y prueba de Friedman. 135

Tabla 14 Distribución de 150 gatos según estado reproductivo y presencia de enfermedad del tracto urinario inferior (ETUI). Clínicas veterinarias de Lima, Perú 173

Tabla 15 Eficacia del tratamiento antihelmíntico en ovejas Corriedale. Riobamba, Ecuador. 178

Tabla 16 Resultados serológicos pre y posvacunación contra leptospirosis en 60 vacas. Llanos Orientales, Colombia. 181

Tabla 17 Resultados de tres pruebas diagnósticas para brucelosis en 20 vacas. Altiplano boliviano. 183

Tabla 18 Peso corporal, longitud femoral, rangos y cálculo de r_s en 8 cachorros Labrador Retriever. 188

Tabla 19 Condición corporal y producción láctea en 10 vacas Holstein. Cajamarca, Perú. 189

Tabla 20 Escala de dolor de Glasgow y dosis de tramadol en 6 perras posquirúrgicas. Medellín, Colombia. 192

Tabla 21 Distribución de 87 perros según raza y gravedad de dysplasia de cadera. Hospital Veterinario Docente, Quito 207

Tabla 22 Distribución de 87 perros según sexo y presencia de dysplasia grave. Hospital Veterinario Docente, Quito. 208

Tabla 23 Estructura de datos de sobrevivencia en un estudio de anaplasmosis bovina (n = 12, datos parciales)..... 231

Tabla 24 Tiempos de sobrevivencia (semanas) de 10 gatos con linfoma alimentario tratados con protocolo COP..... 234

Tabla 25 Cálculo del estimador de Kaplan-Meier para los datos de linfoma felino. 235

Tabla 26 Tiempos de sobrevivencia (días) de perros con osteosarcoma apendicular..... 237

Tabla 27 Selección de la variante de la prueba de log-rank según la estructura temporal de las diferencias..... 240

Tabla 28 Cálculo paso a paso de la prueba de log-rank para los datos de anaplasmosis bovina. 240

Tabla 29 Distribución del grado de endometritis según retención de placenta (n = 120) .. 287

Tabla 30 Coeficientes del modelo de regresión ordinal para grado de endometritis 288

Tabla 31 Distribución del puntaje de claudicación según tipo de piso (n = 30 por grupo) 290

Tabla 32 Clasificación de 50 microcitomas caninos por dos patólogos 292

Tabla 33 Medidas de tamaño del efecto para pruebas no paramétricass 295

Tabla 34 Eficiencia relativa asintótica de pruebas no paramétricas respecto a sus equivalentes paramétricos bajo normalidad 298

Tabla 35 Potencia estimada de Mann-Whitney para comparar conteos de HPG según tamaño muestral 302

Tabla 36 Distribución del grado de lesión de tarso según tipo de cama (Evaluador 1) 319

Tabla 37 Medidas de distancia y disimilitud comunes en análisis multivariado ecológico y veterinario. 338

Tabla 38 Conteos de géneros de garrapatas en 15 bovinos según sistema productivo. 341

Tabla 39 Comparación entre PERMANOVA y ANOSIM. 343

Tabla 40 Interpretación del valor de estrés en NMDS 345

Tabla 41 Familias comunes para-GAM en investigación veterinaria 352

Tabla 42 Conteos de ectoparásitos (géneros de garrapatas y otros artrópodos) en bovinos de tres sistemas productivos 371

Tabla 43 Parámetros hematológicos y parasitológicos en bovinos según raza, finca El Porvenir, Manabí, Ecuador, 2023 400

Tabla 44 Tiempo de recuperación posanestésica (minutos) según protocolo, en 36 perras sometidas a ovario histerectomía, Hospital Veterinario USFQ, Quito, 2024. 400

Tabla 45 Puntaje de dolor (escala analógica visual modificada, 0–4) a las 2, 6 y 12 horas posquirúrgicas en 15 perras del grupo propofol + isoflurano 401

Tabla 46 Incidencia de vómito posanestésico según protocolo anestésico en 36 perras. .. 401

Índice de figuras

Figura 1 Árbol de decisión para la selección del método de análisis estadístico en investigación veterinaria 25

Figura 2 Reducción del conteo de huevos por gramo (HPG) en 14 cabras criollas tras tratamiento con ivermectina 86

Figura 3 Comparación de los puntajes de dolor (CMPS-SF) a las 24 horas post-cirugía entre los dos protocolos analgésicos. 91

Figura 4 Distribución del conteo de garrapatas R. microplus por # genotipo bovino (n = 12 por grupo) 142

Figura 5 Rangos medios del conteo de garrapatas por genotipo bovino(N = 48) 149

Figura 6 Perfil de puntajes de calidad espermática de 15 toros Suizo Pardo evaluados por tres andrólogos. 155

Figura 8 Distribución de puntajes por tres andrólogos 155

Figura 8 Gráficos de barras proporción de ETUI según estado reproductivo 203

Figura 9 Gráficos de dispersión Condición corporal vs. producción láctea 205

Figura 10 Raza vs Gravedad de displacía 210

Figura 11 Proporción de displasia grave según sexo 210

Figura 12 Distribución de HPG en 30 alpacas 212

Figura 13 Carga parasitaria vs. hematocrito en alpacas 213

Figura 14 Correlaciones de Spearman — 30 alpacas (Puno) 215

Figura 15 Gráfico de la curva de Kaplan-Meier de Supervivencia de gatos con linfoma alimentaria 253

Figura 16 Supervivencia en osteosarcoma canino 256

Figura 17 Diagnóstico gráfico del supuesto de riesgos proporcionales 257

Figura 18 Distribución nula de permutación 259

Figura 19 Distribución bootstrap de la mediana de IgG calostral 262

Figura 20 Concentración de IgG calostral por paridad 264

Figura 21 Distribución bootstrap de la diferencia 265

Figura 22 Supervivencia en anaplasmosis bovina 268

Figura 23 Residuos de Schoenfeld escalados 269

Figura 24 Distribución del puntaje de claudicación\nsegún tipo de piso 307

Figura 25 Curva de potencia: Mann-Whitney para conteo de HPG 315

Figura 26 Acuerdo entre patólogos:\nclasificación de mastocitomas 317

Figura 27 Puntaje de claudicación según tipo de piso\nen vacas lecheras 318

Figura 28 PERMDISP: Distancias al centroide por sistema 358

Figura 29 Distancias al centroide 358

Figura 30 ANOSIM: Rangos de disimilitud entre y dentro de grupos 359

Figura 31 NMDS de comunidades de ectoparásitos en bovinos 361

Figura 32 Diagrama de Shepard: ajuste del NMDS 361

Figura 33 Curva de lactancia suavizada con LOESS 363

Figura 34 Efecto suave de días en leche sobre producción 365

Figura 35 gam_lactancia 365

Figura 36 Figura diagnostica 367

Figura 37 Gráfico del efecto suave de DEL 367

Figura 38 Regresión cuantílica: peso al destete 370

Figura 39 PERMDISP: Distancia al centroide por sistema 372

Figura 40 ANOSIM: Rangos de disimilitud 373

Figura 41 NMDS: comunidades de ectoparásitos según sistema productivo 375

Figura 42 Gam Check 378

Figura 43 Efecto suave de los días en leche sobre RCS 379

Figura 44 Efecto suave de DEL sobre RCS 379

Figura 45 GAM: RCS según días en leche y época 380

Figura 46 GAM para-RCS: representación en escala logarítmica 381

Figura 47 Recuento de huevos por gramo (HPG) según raza bovina, finca El Porvenir, Manabí, Ecuador. 403

Figura 48 Evolución del puntaje de dolor posquirúrgico a las 2, 6 y 12 horas en perras del grupo propofol + isoflurano (n = 15). 404

Figura 49 Sobrevivencia libre de recidiva en 48 perros con mastocitoma cutáneo de grado II según protocolo quimioterapéutico: vinblastina-prednisolona (n = 26) vs. observación (n = 22). 405

Figura 50 Ordenación NMDS (Bray-Curtis) del microbiota ruminal de vacas lecheras Holstein suplementadas con extracto de orégano (n = 12) vs. control (n = 12), finca experimental de la Universidad Nacional de Colombia, Bogotá. Stress = 0.09 406

Figura 51 Efecto suavizado de la temperatura ambiental (°C) sobre la producción diaria de leche (L) en vacas Girolando, modelo GAM con k = 10. edf = 4.2. 407

INTRODUCCIÓN

En algún punto de la formación de todo veterinario investigador ocurre un desencuentro que rara vez se discute en voz alta. El estudiante aprende estadística en un curso general a veces dictado por un matemático, a veces por un agrónomo, casi nunca por alguien que haya contado garrapatas en un potrero a las seis de la mañana y sale de ese curso con la convicción de que la prueba t de Student y el ANOVA son las herramientas universales del análisis de datos. Más tarde, cuando enfrenta su tesis de grado o su primer proyecto de investigación, descubre que los datos que recoge no se parecen a los del ejemplo del libro de texto. Los conteos de huevos por gramo de heces tienen valores de cero en la mitad de las muestras y un animal con 3200 HPG que arrastra la media hacia un número que no representa a nadie. Las evaluaciones de dolor posoperatorio son puntajes en una escala de 0 a 4 donde la media aritmética de 2.3 no corresponde a ningún estado clínico reconocible. Los tiempos de sobrevivencia incluyen animales que se vendieron antes de que terminara el estudio y otros que llegaron al final sin presentar el evento. Las muestras tienen 8 o 12 animales por grupo, no los 30 que el teorema central del límite necesita para hacer su trabajo.

Frente a esa situación, el investigador tiene varias opciones. Puede ignorar el problema y aplicar la prueba paramétrica de todos modos, confiando en que "el ANOVA es robusto". Puede intentar una transformación logarítmica o de raíz cuadrada y esperar que los residuos se vuelvan normales, lo cual a veces funciona y a veces no resuelve nada. Puede consultar a un estadístico, si tiene acceso a uno. O puede buscar en la literatura y encontrar que existe una familia completa de métodos estadísticos diseñados para trabajar sin los supuestos que sus datos no cumplen: los métodos no paramétricos.

El problema es que la tercera opción consultar al estadístico no siempre está disponible, especialmente en universidades del interior de Ecuador, Bolivia, Colombia, Perú o Argentina, donde el investigador veterinario trabaja con recursos limitados y acceso restringido a asesoría especializada. Y la cuarta opción aprender métodos no paramétricos por cuenta propia choca contra una barrera concreta: la mayoría de los textos de referencia están en inglés, fueron escritos pensando en ciencias sociales o biomedicina humana, y los ejemplos involucran ensayos clínicos con cientos de pacientes, no estudios de campo con 15 terneros y tres tratamientos. Los textos en español que existen tienden a cubrir el tema de manera superficial, como un capítulo final dentro

de un libro de bioestadística general, sin la profundidad que el investigador necesita para decidir cuándo usar cada prueba, cómo interpretarla correctamente y cómo reportarla en un artículo científico. Este libro existe para cerrar esa brecha.

Qué contiene este libro y cómo está organizado

El libro se estructura en ocho capítulos que siguen un orden lógico, desde los fundamentos hasta las aplicaciones más avanzadas.

El Capítulo 1 parte de una pregunta que debería ser obligatoria antes de cualquier análisis: ¿cumplen mis datos los supuestos que la prueba paramétrica requiere? Se presentan las herramientas gráficas y formales para evaluar normalidad y homocedasticidad, y se discuten con honestidad las cuatro vías posibles cuando los supuestos no se cumplen: transformar, usar métodos robustos, recurrir a modelos lineales generalizados, o tomar la vía no paramétrica. Se incluye un árbol de decisión que el lector puede usar como guía, con la advertencia explícita de que ningún diagrama de flujo reemplaza al juicio informado.

El Capítulo 2 desarrolla las pruebas para una muestra y dos muestras: la prueba de signos, la prueba de rangos con signo de Wilcoxon, la prueba de Mann-Whitney U y la prueba de Brunner-Munzel. Aquí aparece por primera vez el debate sobre si Mann-Whitney compara medianas o distribuciones completas, una confusión que se filtra hasta en revistas indexadas y que este libro aborda presentando ambas posturas con sus autores, sin elegir una y silenciar la otra. Se introduce el estimador de Hodges-Lehmann como herramienta de estimación puntual, y las medidas de tamaño del efecto que el valor p por sí solo no proporciona.

El Capítulo 3 extiende la comparación a tres o más grupos con la prueba de Kruskal-Wallis para diseños independientes y la prueba de Friedman para diseños con bloques o medidas repetidas. Se desarrollan tres procedimientos de comparación post hoc para cada prueba (Dunn, Conover-Iman y DSCF para Kruskal-Wallis; Nemenyi y Conover para Friedman), se incluye la alternativa de Quade, y se presenta el coeficiente de concordancia W de Kendall con la aparente paradoja de que una concordancia alta y un Friedman significativo puedan coexistir.

El Capítulo 4 cubre las tablas de contingencia (chi-cuadrado de independencia, prueba exacta de Fisher, McNemar para datos pareados) y la correlación no paramétrica (Spearman y Kendall),

incluyendo la controversia sobre la corrección de Yates, las medidas de asociación para tablas de distintas dimensiones, y la correlación parcial no paramétrica.

El Capítulo 5 aborda dos familias de métodos que comparten una filosofía común: extraer inferencia sin imponer formas distribucionales. La primera mitad desarrolla el análisis de sobrevivencia estimador de Kaplan-Meier, prueba de log-rank y sus variantes, un territorio donde la censura de datos es la norma y no la excepción en veterinaria clínica. La segunda mitad presenta los métodos de remuestreo: pruebas de permutación y bootstrap no paramétrico, con aplicaciones a intervalos de confianza para medianas, diferencias entre grupos y coeficientes de correlación.

El Capítulo 6 se ocupa de un problema que atraviesa la práctica veterinaria cotidiana: los datos ordinales. Condición corporal, escalas de claudicación, puntajes de dolor, grados de diarrea neonatal, clasificaciones histológicas de tumores: todas son variables donde el orden tiene significado, pero la distancia entre categorías no es constante ni conocida. Se desarrolla la regresión ordinal, las medidas de concordancia entre evaluadores (kappa de Cohen, kappa ponderada, kappa de Fleiss), y un tratamiento extenso del tamaño del efecto y la potencia estadística, con simulaciones en R que permiten al investigador estimar el tamaño de muestra que su diseño realmente necesita.

El Capítulo 7 sube el nivel de complejidad con los métodos multivariados basados en distancias (PERMANOVA, ANOSIM, NMDS), la regresión no paramétrica (LOESS, modelos aditivos generalizados) y la regresión cuantílica. Son herramientas para cuando la pregunta no se reduce a comparar dos medianas sino a evaluar si la composición de una comunidad parasitaria difiere entre sistemas productivos, o si la relación entre días en leche y recuento de células somáticas sigue una curva que ninguna recta captura.

El Capítulo 8 cierra con dos dimensiones que rara vez se enseñan juntas pero que convergen en la práctica del investigador: cómo reportar correctamente los resultados de pruebas no paramétricas según la norma APA séptima edición, y hacia dónde se mueven los métodos analíticos con la irrupción de los métodos bayesianos no paramétricos, el aprendizaje automático y los datos masivos generados por sensores, plataformas genómicas y registros clínicos electrónicos.

Qué hace diferente a este libro

Tres características distinguen a este texto de otros manuales de estadística no paramétrica disponibles en español.

La primera es su anclaje en la investigación veterinaria latinoamericana. Todos los ejemplos, estudios de caso y ejercicios se sitúan en contextos que el lector reconoce: fincas de la sierra ecuatoriana, el altiplano boliviano, los valles interandinos de Colombia, las llanuras ganaderas de Argentina. Las razas son las que se manejan en la región (Corriedale, Brahman, Gyr, Holstein, cruza F1, Suizo Pardo, Ross 308). Los fármacos son los que se prescriben en la práctica local. Los tamaños de muestra son los que un investigador veterinario realmente puede obtener, no los que se necesitarían en un mundo ideal. Esa especificidad no es un adorno; condiciona la elección del método estadístico, porque muchas de las recomendaciones que funcionan con muestras de 50 o 100 sujetos pierden validez con los 8, 12 o 15 animales por grupo que son comunes en ensayos clínicos con bovinos, equinos o caninos.

La segunda es la integración completa con R a lo largo de todo el libro. Cada prueba se implementa con código reproducible, comentado y ejecutable. Se muestran las funciones de distintos paquetes para la misma prueba (wilcox.test del paquete base frente a wilcox_test de coin, por ejemplo), se explica por qué producen valores p diferentes (exacto versus asintótico, con o sin corrección de continuidad), y se guía al lector en la interpretación de cada elemento de la salida de consola. Los gráficos se producen con ggplot2 siguiendo principios de comunicación visual que sirven para una tesis y para un artículo. El estudiante que no tenga experiencia previa en R puede seguir las instrucciones desde la instalación de paquetes; el que ya trabajé con R encontrará funciones y paquetes que quizá no conocía.

La tercera es el tratamiento honesto de las limitaciones y los debates. Los métodos no paramétricos no son pruebas mágicas sin supuestos. Tienen los suyos, menos restrictivos que los paramétricos pero reales. Kruskal-Wallis asume independencia y, si se quiere interpretar como prueba de medianas, requiere que las distribuciones tengan la misma forma. La prueba de rangos con signo de Wilcoxon asume simetría de las diferencias alrededor de la mediana. El bootstrap necesita una muestra representativa y un tamaño suficiente. A lo largo del libro se señalan estos supuestos, se explica cómo verificarlos cuando es posible, y se advierte qué puede salir mal cuando se ignoran. En los puntos donde la literatura no tiene consenso la interpretación de Mann-Whitney, la

corrección de Yates, los umbrales para decidir entre la prueba exacta y la asintótica, se presentan las posiciones de los distintos autores con sus argumentos.

Capítulo 1

Fundamentos de la estadística no paramétrica y verificación de supuestos

Introducción

Este capítulo establece las bases sobre las que se construye el resto del libro. Se trabajan los cuatro supuestos que los métodos paramétricos exigen a los datos normalidad de los residuos, homocedasticidad, independencia de las observaciones y nivel de medición de intervalo o razón y se desarrollan las herramientas para evaluarlos: pruebas formales (Shapiro-Wilk, Anderson-Darling, Lilliefors, Levene, Brown-Forsythe) y métodos gráficos (histogramas, gráficos Q-Q, diagramas de caja). Se examina qué ocurre con el error tipo I y la potencia de las pruebas paramétricas cuando cada supuesto se viola, con simulaciones que permiten ver el efecto en lugar de solo leerlo. Se presentan las cuatro vías de acción disponibles cuando los supuestos no se cumplen transformación de datos, métodos robustos, modelos lineales generalizados y pruebas no paramétricas, con criterios para decidir cuál es la más adecuada según la estructura de los datos y la pregunta de investigación. Se construye un árbol de decisión estadística que guía al investigador desde la inspección inicial hasta la elección del método, y se implementa todo el proceso diagnóstico en R con dos conjuntos de datos veterinarios que representan situaciones opuestas: uno donde los supuestos paramétricos son insostenibles (conteos de HPG en ovinos de la sierra ecuatoriana) y otro donde sí se cumplen (ganancia de peso en pollos de engorde), para que el lector practique tanto la detección del problema como el reconocimiento de cuándo el problema no existe.

Objetivos de aprendizaje

Al finalizar este capítulo, el estudiante será capaz de:

1. **Identificar** los supuestos que exige cada método paramétrico clásico (normalidad, homocedasticidad, independencia, escala de medición) y reconocer las situaciones experimentales en medicina veterinaria donde esos supuestos se violan con frecuencia.

2. **Contrastar** la lógica de inferencia paramétrica con la lógica no paramétrica, diferenciando lo que cada enfoque estima, lo que gana y lo que sacrifica.

3. **Aplicar** pruebas gráficas (histogramas, Q-Q plots, diagramas de caja) y pruebas formales (Shapiro-Wilk, Levene, Bartlett) para evaluar normalidad y homocedasticidad en conjuntos de datos veterinarios reales.

4. **Decidir**, ante un conjunto de datos concreto, si corresponde usar un método paramétrico, transformar los datos, ajustar un modelo generalizado o recurrir a la vía no paramétrica, justificando la elección con criterios técnicos y no por costumbre.

5. **Implementar** en R el flujo completo de exploración de supuestos desde la lectura de datos hasta la generación de gráficos diagnósticos y la ejecución de pruebas formales usando los paquetes *ggplot2*, *car*, *nortest* y funciones del sistema base.

1.1. Qué asume un método paramétrico y por qué esos supuestos importan

El escenario que todo tesista reconoce

Un estudiante de maestría en ciencias veterinarias quiere evaluar el efecto de dos antiparasitarios sobre el conteo de huevos por gramo de heces (HPG) en ovinos Corriedale de una finca en la provincia de Chimborazo, Ecuador. Tiene 12 animales por grupo. Abre su software, aplica una prueba *t* de Student para muestras independientes, obtiene un valor $p = 0.08$ y concluye que "no hay diferencia significativa entre tratamientos". Lo que nunca verificó es que los datos de HPG se ven así:

Grupo A: 0, 0, 50, 100, 100, 150, 200, 300, 450, 1000, 1800, 3200

Grupo B: 0, 0, 0, 50, 50, 100, 100, 150, 200, 250, 400, 600

Esos datos no se distribuyen normalmente. Ni siquiera se acercan. Tienen una asimetría pronunciada a la derecha, varios ceros, y un par de valores extremos que arrastran la media hacia arriba. La prueba *t* estima la diferencia entre medias asumiendo que los datos provienen de poblaciones normales con varianzas iguales. Cuando eso no se cumple, el valor *p* que produce puede estar inflado o deflactado, y la decisión que el tesista toma sobre su hipótesis descansa sobre arena.

Este escenario no es hipotético. Es cotidiano. Y es el punto de partida de este libro.

Los cuatro supuestos clásicos

Los métodos paramétricos más usados en investigación veterinaria la prueba *t* de Student, el análisis de varianza (ANOVA), la regresión lineal comparte un conjunto de supuestos sobre los datos. Cuando un investigador elige una de estas pruebas, acepta implícitamente que sus datos cumplen las siguientes condiciones:

1. Normalidad de las distribuciones (o de los residuos). La variable respuesta, o más precisamente los residuos del modelo, siguen una distribución normal (gaussiana) en cada grupo o condición. Esto implica que la distribución es simétrica alrededor de la media, que la media y la mediana coinciden, y que la proporción de observaciones en cualquier intervalo se puede predecir a partir de la media y la desviación estándar.

¿Por qué importa? Porque la prueba *t* y el ANOVA construyen sus estadísticos de prueba asumiendo que la variabilidad de los datos sigue un patrón específico. La distribución *t* de Student, que se usa para calcular valores *p* e intervalos de confianza, es la distribución de un cociente donde el numerador es una variable normal y el denominador involucra una chi-cuadrado. Si la distribución real de los datos difiere de la normal si es asimétrica, tiene colas pesadas o está truncada, ese cociente ya no sigue una distribución *t*, y los valores *p* calculados dejan de ser confiables.

2. Homocedasticidad (homogeneidad de varianzas). Las varianzas poblacionales de los grupos comparados son iguales. En un ANOVA de un factor con tres tratamientos, esto significa que $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$. La prueba F del ANOVA es sensible a violaciones de este supuesto, especialmente cuando los tamaños de grupo son desiguales.

En investigación veterinaria, la heterocedasticidad aparece con frecuencia. Un tratamiento que reduce la carga parasitaria, por ejemplo, suele reducir también la variabilidad: los animales del grupo tratado se concentran en valores bajos, mientras que los del grupo control muestran un rango amplio de respuestas. Un ensayo de ganancia de peso donde un grupo responde bien al tratamiento y otra muestra respuestas muy heterogéneas genera varianzas dispares.

3. Independencia de las observaciones. Cada dato es independiente de los demás. No hay correlación entre las respuestas de distintos individuos, y la medición de un animal no influye en

la de otro. Este supuesto se viola cuando se toman mediciones repetidas en el mismo animal sin modelar la correlación, cuando los animales comparten corral y se influyen mutuamente (efecto de agrupamiento), o cuando existe una estructura jerárquica no reconocida (animales dentro de fincas, fincas dentro de regiones).

La violación de independencia es posiblemente la más grave de las cuatro, y la que menos se evalúa. Ninguna prueba no paramétrica la resuelve por sí sola: si las observaciones no son independientes, tanto los métodos paramétricos como los no paramétricos producen inferencias incorrectas, a menos que se use un diseño o modelo que contemple la estructura de dependencia.

4. Escala de medición al menos de intervalo. La variable respuesta se mide en una escala donde las distancias entre valores son interpretables y constantes. La diferencia entre 10 kg y 12 kg de peso tiene el mismo significado que la diferencia entre 30 kg y 32 kg. Esto permite calcular medias y desviaciones estándar con sentido.

Este supuesto se viola de manera rutinaria en veterinaria. Cuando un clínico califica el grado de claudicación en una escala de 0 a 5, la condición corporal en una escala de 1 a 9, o la severidad de una lesión histopatológica como leve, moderada o grave, trabaja con datos ordinales. La media de un puntaje de claudicación de 2.7 no tiene interpretación clínica clara. Un análisis paramétrico sobre estos datos trata los intervalos entre categorías como iguales, cuando no hay razón para suponer que la distancia entre “claudicación grado 1” y “grado 2” sea la misma que entre “grado 4” y “grado 5”.

No todos los supuestos pesan igual

Conviene distinguir entre supuestos a los que las pruebas son relativamente tolerantes (supuestos frente a los cuales la prueba es *robusta*) y aquellos cuya violación distorsiona seriamente los resultados.

La prueba *t* y el ANOVA, por ejemplo, toleran desviaciones moderadas de la normalidad cuando los tamaños de muestra son grandes, gracias al teorema central del límite. Con 30 o más observaciones por grupo, la distribución muestral de la media se aproxima a la normal incluso si los datos individuales no lo son. Pero en investigación veterinaria, ¿cuántas veces se trabaja con 30 animales por grupo? En ensayos clínicos con bovinos, equinos o caninos, 8 a 15 animales por

tratamiento es común. Con esas muestras, la robustez del teorema central del límite no alcanza para proteger la inferencia.

La heterocedasticidad combinada con tamaños de muestra desiguales es particularmente dañina. Si el grupo más pequeño tiene la varianza más grande, la prueba *t* clásica tiende a rechazar la hipótesis nula con más frecuencia de la que debería (tasa de error tipo I inflada). Si el grupo más pequeño tiene la varianza más pequeña, la prueba se vuelve conservadora y pierde potencia. Ninguna de las dos situaciones es deseable.

La violación de independencia, como se mencionó, es la más peligrosa y la menos corregible con ajustes simples. En un estudio donde se miden 5 terneros de cada una de 10 vacas, y se tratan los 50 terneros como observaciones independientes, la unidad experimental real es la vaca (o el par vaca-ternero), no el ternero individual. El error estándar se subestima, los valores *p* se inflan y el investigador "encuentra" diferencias que no existen. Los métodos no paramétricos estándar que se presentan en este libro no resuelven este problema: requieren independencia tanto como los paramétricos.

1.2. Qué significa "no paramétrico" y qué no significa

El nombre y su significado técnico

El término "no paramétrico" tiene un significado preciso que a menudo se diluye en la enseñanza introductoria. Un método estadístico paramétrico asume que los datos provienen de una familia conocida de distribuciones de probabilidad (por ejemplo, la familia normal) y estima los parámetros de esa distribución (μ, σ^2). Toda la inferencia se construye alrededor de esos parámetros.

Un método no paramétrico, en su definición más estricta, no asume una forma funcional específica para la distribución de los datos. No necesita estimar μ ni σ^2 porque no trabaja con un modelo distribucional. En su lugar, trabaja con propiedades más generales de los datos: sus rangos (el orden relativo de las observaciones), sus signos (si una diferencia es positiva o negativa), las frecuencias con que caen en categorías, o en el caso de los métodos de remuestreo con la distribución empírica de los propios datos observados.

Tres cosas que "no paramétrico" no significa

1. No significa "sin supuestos". Este es quizá el mito más persistente. Es frecuente leer en tesis veterinarias que "se usó la prueba de Kruskal-Wallis porque no requiere supuestos". Falso. Kruskal-Wallis asume independencia de las observaciones, que la variable respuesta es al menos ordinal, y si se quiere interpretar como prueba de diferencia de medianas que las distribuciones de los grupos tienen la misma forma (lo que incluye la misma varianza). Si las distribuciones difieren en forma, Kruskal-Wallis puede rechazar la hipótesis nula porque las formas son distintas, no porque las medianas lo sean. El capítulo 5 desarrolla esta distinción en profundidad.

Cada prueba no paramétrica tiene sus propios supuestos. Son menos restrictivos que los paramétricos, pero existen. Parte de la responsabilidad del investigador es conocerlos, verificar los que se puedan verificar y reportar los que se asumen.

2. No significa "para datos que no sirven". En algunos programas de estadística veterinaria, los métodos no paramétricos se enseñan como el último recurso, la alternativa inferior que se usa "cuando los datos no cumplen los supuestos y no se puede hacer otra cosa". Esa visión es reduccionista. Los métodos no paramétricos son la herramienta correcta no la alternativa pobre en varias situaciones legítimas:

- Cuando la variable es ordinal por naturaleza (puntajes clínicos, escalas de bienestar animal, grados de lesión histopatológica).
- Cuando se trabaja con medianas porque la media no es un estadístico representativo de la distribución (datos muy asimétricos como conteos de parásitos).
- Cuando la muestra es tan pequeña que evaluar normalidad carece de sentido estadístico (5 animales por grupo, algo habitual en estudios piloto con grandes animales).
- Cuando el interés recae en la distribución completa y no solo en un parámetro de localización.

3. No significa que siempre sean menos potentes. La potencia estadística es la probabilidad de detectar una diferencia real cuando existe. Los métodos no paramétricos pagan un precio en

potencia cuando los datos realmente provienen de una distribución normal: en ese escenario, Mann-Whitney tiene aproximadamente un 95% de la potencia de la prueba *t* (lo que se conoce como eficiencia relativa asintótica de $\pi/3 \approx 0.955$). Esa pérdida del 5% rara vez cambia una conclusión.

Pero cuando los datos provienen de distribuciones con colas pesadas, sesgadas o contaminadas con valores atípicos situaciones comunes en datos veterinarios de campo, los métodos no paramétricos pueden ser *más* potentes que los paramétricos. La media aritmética, que es el corazón de la prueba *t*, es muy sensible a valores extremos. La mediana y los rangos no lo son. Un solo animal con un conteo de HPG de 15,000 en un grupo de 12 puede arrastrar la media del grupo lo suficiente como para que la prueba *t* no detecte una diferencia real. Mann-Whitney, que trabaja con el orden de las observaciones y no con sus magnitudes absolutas, es inmune a esa distorsión.

La clasificación operativa

A efectos prácticos, las pruebas estadísticas que se cubren en este libro se pueden organizar en tres categorías según su relación con los supuestos distribucionales:

Pruebas paramétricas clásicas (*t* de Student, ANOVA, correlación de Pearson, regresión lineal): asumen normalidad y homocedasticidad. Trabajan con medias y varianzas.

Pruebas no paramétricas basadas en rangos y signos (Wilcoxon, Mann-Whitney, Kruskal-Wallis, Friedman, Spearman): reemplazan las observaciones originales por sus rangos o signos. No asumen una distribución específica. Comparan distribuciones, no necesariamente medias.

Métodos basados en la distribución empírica y remuestreo (bootstrap, pruebas de permutación, Kaplan-Meier): usan la distribución observada en los datos como base para la inferencia, sin recurrir a tablas de distribuciones teóricas. El capítulo 10 los desarrolla.

Una cuarta categoría, que se aborda en los capítulos 13 y 16, incluye métodos semiparamétricos y no paramétricos más avanzados (GAM, métodos bayesianos no paramétricos) que no encajan limpiamente en las anteriores.

1.3. Ventajas reales y limitaciones honestas

Ventajas que sí cuentan

Aplicabilidad a datos ordinales. La ventaja más clara y menos debatible. Cuando la variable respuesta es ordinal condición corporal, grado de cojera, escala de dolor, clasificación histopatológica, los métodos no paramétricos son los métodos correctos, no una alternativa. Calcular la media de una escala de dolor de 0 a 4 y aplicar un ANOVA es un error conceptual, no solo estadístico: trata las distancias entre categorías como iguales sin evidencia de que lo sean. Los rangos, en cambio, respetan la ordinalidad sin imponer supuestos sobre los intervalos.

Resistencia a valores atípicos. En datos de campo veterinario, los valores extremos son la norma, no la excepción. Un bovino con un conteo de células somáticas de 8,000,000 cel/mL en un hato donde el promedio es 300,000 cel/mL, un perro con un valor de creatinina quintuplicado respecto al grupo, un pollo con una ganancia de peso negativa porque enfermó durante el ensayo. Los métodos basados en rangos absorben estos valores sin distorsión: un dato atípico extremo recibe el rango más alto, pero ese rango tiene un peso acotado en el estadístico de prueba.

Validez con muestras pequeñas. Con muestras de 5 a 10 observaciones por grupo, evaluar normalidad tiene poca potencia (Shapiro-Wilk con $n = 6$ no rechaza casi nada), y la aproximación del teorema central del límite no es confiable. Las pruebas exactas no paramétricas calculan la distribución del estadístico de prueba bajo la hipótesis nula enumerando todas las permutaciones posibles, sin necesidad de aproximaciones asintóticas. Esto las hace válidas incluso con muestras que a un método paramétrico le resultarían problemáticas.

Menor cantidad de supuestos verificables. No se elimina la necesidad de supuestos, pero se reduce la cantidad de verificaciones que el investigador debe hacer y, en consecuencia, la cantidad de decisiones que pueden salir mal en el camino.

Limitaciones que no se deben ocultar

Menor potencia cuando los supuestos paramétricos sí se cumplen. Si los datos son genuinamente normales, homocedásticos y la muestra es razonablemente grande, la prueba *t* es más potente que Mann-Whitney. No por mucho (la pérdida es de alrededor del 5%

asintóticamente), pero existe. Usar un método no paramétrico "por si acaso" cuando los datos cumplen los supuestos paramétricos es conservador y sacrifica potencia sin necesidad.

Dificultad para construir intervalos de confianza y tamaños de efecto. Los métodos paramétricos producen naturalmente intervalos de confianza para las diferencias de medias, cocientes de varianzas y coeficientes de regresión. Los métodos no paramétricos clásicos no siempre tienen equivalentes directos. La estimación de Hodges-Lehmann para la diferencia de localización entre dos grupos y los intervalos de confianza basados en rangos existen, pero son menos conocidos y menos intuitivos. El capítulo 11 aborda este tema.

Limitaciones en diseños complejos. No existe un equivalente no paramétrico directo para un ANOVA factorial de dos factores con interacción, ni para un modelo mixto con efectos aleatorios anidados. Cuando el diseño es multifactorial o jerárquico, las opciones no paramétricas son más limitadas y a menudo requieren técnicas avanzadas (PERMANOVA, modelos de remuestreo) que se cubren en los capítulos 10 y 12.

Pérdida de información cuando se convierten datos de intervalo a rangos. Si los datos originales están en escala de razón y cumplen los supuestos paramétricos, convertirlos a rangos descarta información sobre las magnitudes reales de las diferencias. La diferencia entre 10 kg y 50 kg se trata igual que la diferencia entre 10 kg y 11 kg si ambas corresponden a rangos consecutivos. En datos que son genuinamente normales, esa pérdida de información es innecesaria.

La tentación de usarlos como solución automática. "Los datos no son normales, entonces uso Kruskal-Wallis" es un razonamiento incompleto. Antes de saltar a la prueba no paramétrica, el investigador debería preguntarse si una transformación logarítmica estabiliza la distribución, si un modelo lineal generalizado (GLM) es más apropiado para el tipo de variable (conteos, proporciones), o si el problema es el diseño experimental y no el método de análisis. La sección 1.6 desarrolla estas alternativas.

1.4. Evaluación de normalidad: métodos gráficos y pruebas formales

La pregunta correcta

No se trata de preguntar "¿mis datos son normales?" como si fuera una pregunta de sí o no. Ningún dato real es exactamente normal. La pregunta operativa es: **¿la desviación de la normalidad es lo suficientemente grande como para que la inferencia paramétrica deje de ser confiable?** La respuesta depende de cuánto se desvían los datos, del tamaño de muestra y de cuánto tolera la prueba paramétrica en cuestión.

Métodos gráficos: ver antes de calcular

Los métodos gráficos son el primer paso y, en muchos casos, el más informativo. Una prueba formal devuelve un valor *p*, pero no dice *cómo* se desvían los datos de la normalidad. Un gráfico sí lo dice: muestra si la distribución es asimétrica, bimodal, tiene colas pesadas o valores atípicos aislados. Cada tipo de desviación tiene implicaciones diferentes para la elección del método de análisis.

Histograma. El más intuitivo. Divide el rango de los datos en intervalos (bins) y muestra la frecuencia de observaciones en cada uno. Con muestras grandes ($n > 50$), un histograma permite apreciar la forma general de la distribución. Con muestras pequeñas, la apariencia del histograma depende mucho del número de intervalos elegidos, lo que lo hace poco confiable como herramienta diagnóstica. En estudios veterinarios con 10 a 20 animales por grupo, el histograma es orientativo, pero no definitivo.

Diagrama de caja (boxplot). Muestra la mediana, los cuartiles, el rango intercuartílico y los valores atípicos. No evalúa normalidad directamente, pero revela asimetría (cuando la mediana no está centrada en la caja) y la presencia de valores extremos. En datos veterinarios es especialmente útil para comparar la forma de la distribución entre grupos tratamiento.

Gráfico Q-Q (quantile-quantile). El método gráfico más potente para evaluar normalidad. Coloca los cuantiles observados de los datos en el eje vertical y los cuantiles teóricos de una distribución normal en el eje horizontal. Si los datos son normales, los puntos caen sobre una línea recta a 45 grados. Las desviaciones de esa línea tienen patrones diagnósticos reconocibles:

- Puntos que se curvan hacia arriba en ambos extremos: distribución con colas pesadas (leptocúrtica).

- Puntos que se curvan hacia abajo en ambos extremos: distribución con colas ligeras (platicúrtica).

- Curva en forma de "S": asimetría. Si la curva sube por encima de la línea a la derecha, la distribución tiene cola derecha larga (asimetría positiva).

- Puntos aislados que se despegan de la línea en un extremo: valores atípicos.

El gráfico Q-Q es la herramienta que este libro recomienda como primer paso en cualquier evaluación de normalidad, complementada con la banda de confianza simulada que produce la función *qqPlot()* del paquete *car* en R.

Gráfico de densidad (density plot). Suaviza el histograma mediante una estimación de densidad kernel. Visualmente atractivo y útil con muestras medianas a grandes. Con muestras pequeñas, la suavización puede crear formas engañosas.

Pruebas formales de normalidad

Las pruebas formales complementan el diagnóstico gráfico. Todas prueban la hipótesis nula H_0: "los datos provienen de una distribución normal" contra la alternativa H_1: "los datos no provienen de una distribución normal". Un valor *p* bajo (convencionalmente < 0.05) lleva a rechazar H_0 y concluir que los datos no son normales.

Prueba de Shapiro-Wilk. Es la prueba más recomendada para muestras pequeñas a moderadas ($n < 50$), que es precisamente el rango habitual en investigación veterinaria. Su estadístico *W* mide qué tan bien los cuantiles observados se alinean con los esperados bajo normalidad:

$$W = \frac{\left(\sum_{i=1}^{n} a_i \; x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Donde:

- $x_{(i)}$ son los datos ordenados de menor a mayor (estadísticos de orden).
- a_i son coeficientes tabulados que dependen del tamaño de muestra *n*.
- $\bar{x}$ es la media muestral.
- *W* toma valores entre 0 y 1; valores cercanos a 1 indican compatibilidad con normalidad.

Ejemplo numérico. Considérese un conjunto de 8 valores de ganancia diaria de peso (g/día) en pollos de engorde Ross 308: 52, 55, 57, 58, 60, 61, 63, 78. A simple vista, el valor 78 se separa del resto. Shapiro-Wilk produce $W = 0.877$ y $p = 0.18$. Con $n = 8$, la prueba no tiene suficiente potencia para rechazar normalidad, aunque gráficamente la distribución luce asimétrica. Esto ilustra un problema recurrente: con muestras pequeñas, Shapiro-Wilk rara vez rechaza.

Prueba de Kolmogorov-Smirnov (con corrección de Lilliefors). Compara la función de distribución empírica con la función de distribución normal teórica. Es menos potente que Shapiro-Wilk para detectar la mayoría de las desviaciones de normalidad, pero se incluye en algunos programas estadísticos como prueba predeterminada. La versión original de Kolmogorov-Smirnov requiere que los parámetros (μ, σ) estén especificados de antemano, no estimados de los datos. La corrección de Lilliefors ajusta la distribución del estadístico para el caso en que μ y σ se estiman de la muestra, que es lo que ocurre en la práctica.

Prueba de Anderson-Darling. Variante de Kolmogorov-Smirnov que da más peso a las colas de la distribución. Es más potente que Kolmogorov-Smirnov-Lilliefors para detectar desviaciones en las colas y está disponible en el paquete *nortest* de R.

La trampa de las muestras pequeñas y grandes

Con las pruebas formales de normalidad se presenta una paradoja práctica:

- Cuando la muestra es pequeña ($n < 15$), que es cuando más necesitaríamos verificar normalidad, las pruebas tienen poca potencia y casi nunca rechazan. El tesista concluye que “los datos son normales” cuando en realidad la prueba no tenía capacidad para detectar la desviación.

- Cuando la muestra es grande ($n > 100$), las pruebas son tan potentes que rechazan normalidad ante desviaciones triviales que no afectan la validez de la prueba *t* ni del ANOVA. El investigador descarta un método paramétrico que habría funcionado bien.

Por esta razón, la evaluación de normalidad no puede descansar solo en el valor *p* de una prueba formal. Debe integrar la inspección gráfica (¿cuánto se desvía la distribución?), el tamaño de muestra (¿alcanza para que el teorema central del límite ayude?) y el conocimiento de la variable (¿la variable es intrínsecamente no normal, como un conteo o una proporción?).

1.5. Evaluación de homocedasticidad

Qué se evalúa y por qué

La homocedasticidad (del griego *homos* = igual, *skedasis* = dispersión) es la igualdad de varianzas entre los grupos que se comparan. En un ANOVA de un factor, la prueba F asume que $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma\_k^2$. Cuando esta condición no se cumple, la distribución F ya no es la referencia correcta, y los valores *p* se distorsionan.

La regla informal y su límite

Una regla empírica frecuente indica que si la razón entre la varianza más grande y la más pequeña es menor que 4 (o, equivalentemente, si la razón entre las desviaciones estándar más grande y la más pequeña es menor que 2), la violación de homocedasticidad es tolerable para la prueba *t* y el ANOVA, siempre que los tamaños de grupo sean similares. Esta regla es orientativa. Con tamaños de grupo muy desiguales, incluso razones de varianzas moderadas producen distorsiones.

Prueba de Levene

La prueba de Levene evalúa la hipótesis nula H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma\_k^2$. Su principio es sencillo: calcula las desviaciones absolutas de cada observación respecto a la medida de tendencia central de su grupo y aplica un ANOVA sobre esas desviaciones. Si un grupo tiene mayor dispersión, las desviaciones absolutas de ese grupo serán sistemáticamente mayores.

El estadístico de Levene basado en la mediana (que es la versión más robusta y la implementada por defecto en la función *leveneTest()* del paquete *car* en R) se calcula como:

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} n_i \, (\tilde{z}_{i\cdot} - \tilde{z}_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(z_{ij} - \tilde{z}_{i\cdot}\right)^2}$$

donde:

- $z_{ij} = \left| y_{ij} - \tilde{y}_i \right|$ es la desviación absoluta de la observación j del grupo i respecto a la mediana $\tilde{y}_i$ de su grupo.
- $\tilde{z}_{i\cdot}$ es la media de las z_{ij} dentro del grupo i.
- $\tilde{z}_{\cdot\cdot}$ es la media general de todas las z_{ij}.
- k es el número de grupos, n_i es el tamaño del grupo i, $N = \sum n_i$ es el total de observaciones.

El estadístico *W* se distribuye aproximadamente como F con (k – 1) y (N – k) grados de libertad bajo H_0.

Prueba de Bartlett

La prueba de Bartlett es más antigua y se basa en la razón de verosimilitudes. Es más potente que Levene cuando los datos son normales, pero extremadamente sensible a la no normalidad: con datos no normales, rechaza homocedasticidad incluso cuando las varianzas son iguales, simplemente porque la prueba detecta la no normalidad. Por esta razón, **no se recomienda Bartlett cuando hay dudas sobre la normalidad de los datos**, que es precisamente la situación más común en la práctica. Levene basada en la mediana es la opción preferida.

Prueba de Fligner-Killeen

Una alternativa no paramétrica para evaluar homocedasticidad. Usa los rangos de las desviaciones absolutas respecto a la mediana de cada grupo y aplica una prueba chi-cuadrado. Es la prueba de elección cuando se sospecha fuertemente de no normalidad y se quiere evaluar homocedasticidad sin que la no normalidad contamine el resultado. Está implementada en R base como *fligner.test()*.

1.6. Qué hacer cuando los supuestos no se cumplen: transformaciones, métodos robustos, GLM y vía no paramétrica

Verificar que los supuestos no se cumplen es solo la mitad del problema. La otra mitad es decidir qué hacer. La tabla de decisión no es binaria ("paramétrico o no paramétrico"), sino un abanico de opciones que el investigador debe evaluar según el tipo de variable, el tamaño de muestra y la pregunta de investigación.

- **Opción 1: Transformación de datos**

La idea es aplicar una función matemática a los datos para que la distribución transformada se acerque más a la normalidad y estabilice las varianzas.

Transformación logarítmica: $y' = \log(y)$ o $y' = \log(y + 1)$ cuando hay ceros. Es la más usada en veterinaria para datos de conteo (HPG, recuento bacteriano, títulos de anticuerpos) y para variables con distribución log-normal (concentraciones séricas, tiempos de coagulación). Comprime los valores altos y expande los bajos, reduciendo la asimetría positiva.

Transformación raíz cuadrada: $y' = \sqrt{y}$ o $y' = \sqrt{y + 0.5}$. Útil para datos de conteo con distribución Poisson, donde la varianza es proporcional a la media.

Transformación arcoseno de la raíz cuadrada: $y' = \arcsin\left(\sqrt{p}\right)$, donde p es una proporción. Usada para datos expresados como porcentajes o proporciones (porcentaje de fertilidad, tasa de supervivencia, porcentaje de eclosión).

Limitaciones de las transformaciones. La transformación cambia la escala en la que se interpretan los resultados. Cuando se aplica una transformación logarítmica a datos de HPG y se obtiene una diferencia significativa entre medias en la escala log, la retrotraducción a la escala original da una razón de medias geométricas, no una diferencia de medias aritméticas. Esto no es un defecto, pero requiere que el investigador sepa interpretar correctamente lo que está reportando. La media geométrica de HPG puede ser un estadístico más relevante que la media aritmética para datos parasitológicos, pero no siempre es intuitivo.

Otro problema: la transformación que normaliza los datos puede no estabilizar las varianzas, y viceversa. No siempre existe una transformación que resuelva ambos problemas simultáneamente.

- **Opción 2: Métodos robustos**

Los métodos robustos mantienen la estructura del análisis paramétrico, pero reemplazan estimadores sensibles a valores atípicos por otros más resistentes. La prueba *t* con corrección de Welch, por ejemplo, no asume homocedasticidad: ajusta los grados de libertad para compensar la diferencia de varianzas. No resuelve la no normalidad, pero elimina la dependencia de la homocedasticidad.

La media recortada (trimmed mean), que descarta un porcentaje fijo de los valores extremos antes de calcular la media, es otra herramienta robusta que se ha propuesto como alternativa a los métodos no paramétricos en algunos contextos.

- **Opción 3: Modelos lineales generalizados (GLM)**

Para ciertos tipos de datos, el problema no es que los supuestos no se cumplan, sino que se está usando el modelo equivocado. Los GLM extienden la regresión lineal a distribuciones no normales y permiten modelar directamente:

- Datos de conteo (HPG, recuento de ectoparásitos, número de lesiones): distribución Poisson o binomial negativa con función de enlace logarítmica.
- Datos binarios (vivo/muerto, enfermo/sano, preñada/no preñada): distribución binomial con función de enlace logit.
- Datos de proporciones: distribución binomial o beta.

El GLM no transforma los datos; modela directamente la relación entre la variable respuesta y los predictores bajo la distribución apropiada. Esto lo convierte en una herramienta poderosa que, en muchos casos, es más adecuada que la prueba no paramétrica correspondiente. Un conteo de HPG, por ejemplo, se modela mejor con un GLM binomial negativo que con un Kruskal-Wallis, porque el GLM puede manejar múltiples factores, covariables y su interacción.

Este libro se centra en métodos no paramétricos, pero el investigador veterinario debe conocer los GLM como alternativa. No es honesto presentar lo no paramétrico como la única salida cuando los supuestos paramétricos fallan.

- **Opción 4: Vía no paramétrica**

La ruta no paramétrica es la elección más directa cuando se cumple alguna de estas condiciones:

- La variable es ordinal (no se puede aplicar GLM ni transformaciones a una escala de condición corporal de 1 a 5).
- La muestra es tan pequeña que ninguna verificación de supuestos es confiable y se prefiere un método que no dependa de ellos.
- Los datos tienen una distribución irregular que no se ajusta a ninguna de las familias usadas por los GLM (ni normal, ni Poisson, ni binomial negativa, ni gamma).
- El interés recae en comparar distribuciones completas, no solo medias.
- Se busca simplicidad y transparencia: la prueba de rangos es más fácil de explicar y replicar que un GLM binomial negativo con dispersión estimada.

¿Cuándo no elegir lo no paramétrico?

Conviene ser concreto sobre las situaciones en que la vía no paramétrica no es la mejor elección:

- Si los datos cumplen razonablemente los supuestos paramétricos (normalidad aproximada, varianzas similares, $n > 15$ por grupo), usar el método paramétrico da más potencia, intervalos de confianza más directos y permite análisis multifactoriales que los no paramétricos no manejan fácilmente.
- Si la variable respuesta es un conteo o una proporción y se cuenta con un tamaño de muestra moderado a grande, un GLM con la distribución apropiada suele ser mejor que la prueba no paramétrica porque modela la relación de manera más eficiente y permite incluir covariables.

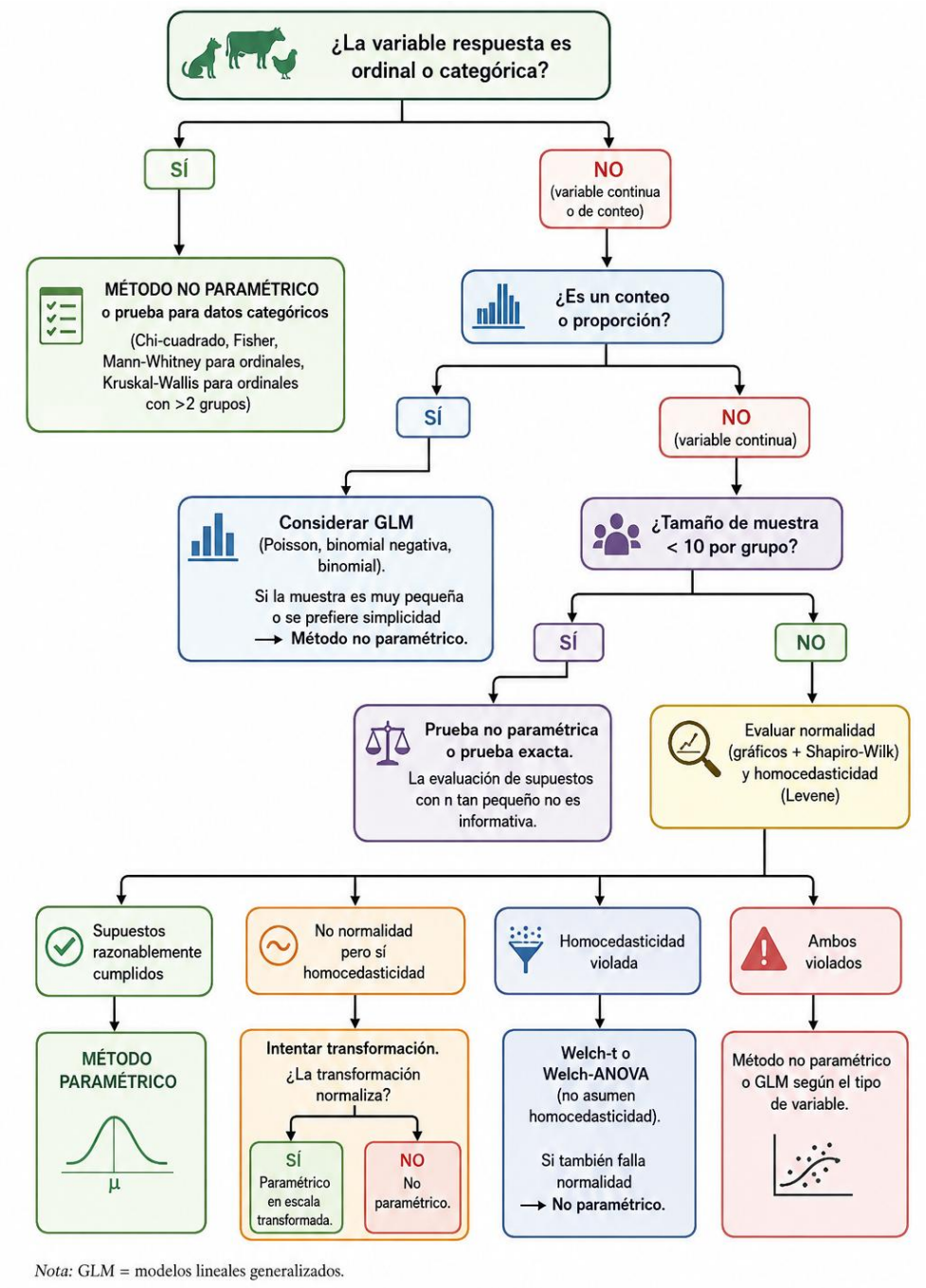
- Si el diseño tiene más de un factor y se necesita evaluar interacciones, los métodos no paramétricos clásicos (Kruskal-Wallis, Friedman) no manejan interacciones directamente.

1.7. El árbol de decisión: ¿paramétrico o no paramétrico?

El diagrama de flujo que sigue resume la lógica de decisión para la elección del método de análisis. No pretende cubrir todos los casos, pero orienta la decisión inicial.

Figura 1

Árbol de decisión para la selección del método de análisis estadístico en investigación veterinaria



Nota: GLM = modelos lineales generalizados.

Este árbol se presenta como orientación, no como algoritmo rígido. La decisión final siempre debe considerar la pregunta de investigación, la experiencia previa con ese tipo de datos y la facilidad de interpretación de los resultados.

1.8. Implementación en R: exploración de supuestos con datos veterinarios

Paquetes requeridos

```
# Instalar solo si no están instalados
# install.packages(c("ggplot2", "car", "nortest", "dplyr", "moments"))

library(ggplot2)    # Gráficos de alta calidad
library(car)        # leveneTest(), qqPlot() con bandas de confianza
library(nortest)    # lillie.test(), ad.test() (Lilliefors, Anderson-Darling)
library(dplyr)      # Manipulación de datos
library(moments)    # skewness(), kurtosis()
```

Ejemplo 1: Conteo de huevos por gramo de heces (HPG) en ovinos

Creación de los datos

```
# Conteo de HPG en ovinos Corriedale tratados con dos antiparasitarios

# Finca en Riobamba, provincia de Chimborazo, Ecuador (datos simulados

# con base en rangos publicados en estudios de resistencia antihelmíntica)

hpg <- data.frame(
  tratamiento = factor(rep(c("Ivermectina", "Albendazol"), each = 12)),
  hpg = c(
    # Grupo Ivermectina
    0, 0, 50, 100, 100, 150, 200, 300, 450, 1000, 1800, 3200,
    # Grupo Albendazol
    0, 0, 0, 50, 50, 100, 100, 150, 200, 250, 400, 600
  )
)

# Inspección rápida
str(hpg)
summary(hpg)
```

Salida esperada:

```
'data.frame':   24 obs. of 2 variables:
 $ tratamiento: Factor w/ 2 levels "Albendazol","Ivermectina": 2 2 2 2 2 2 2 2 2 2 ...
 $ hpg: num  0 0 50 100 100 150 200 300 450 1000 ...

    tratamiento      hpg
 Albendazol :12   Min:   0
 Ivermectina:12   1st Qu.:  50
```

```
Median: 125
Mean: 383
3rd Qu.: 375
Max.   :3200
```

La media (383) es tres veces mayor que la mediana (125). Esta discrepancia es una primera señal de asimetría marcada.

Evaluación gráfica de normalidad

```
# Histograma por grupo
ggplot(hpg, aes(x = hpg)) +
  geom_histogram(binwidth = 200, fill = "#4E79A7", color = "white",

                 alpha = 0.8) +
  facet_wrap(~ tratamiento, ncol = 1) +
  labs(
    x = "Huevos por gramo de heces (HPG)",
    y = "Frecuencia",
    title = "Distribución del conteo de HPG por tratamiento"
  ) +
  theme_minimal(base_size = 12)

# Gráfico Q-Q con banda de confianza simulada (paquete car)
# Se evalúa por grupo
par(mfrow = c(1, 2))  # Dos paneles lado a lado
qqPlot(hpg$hpg[hpg$tratamiento == "Ivermectina"],
       main = "Q-Q Ivermectina", ylab = "HPG",
       pch = 19, col = "#4E79A7", id = FALSE)
qqPlot(hpg$hpg[hpg$tratamiento == "Albendazol"],
       main = "Q-Q Albendazol", ylab = "HPG",
       pch = 19, col = "#E15759", id = FALSE)
par(mfrow = c(1, 1))  # Restaurar diseño de un panel

# Boxplot comparativo
ggplot(hpg, aes(x = tratamiento, y = hpg, fill = tratamiento)) +
  geom_boxplot(alpha = 0.7, outlier.shape = 19, outlier.size = 2) +
  scale_fill_manual(values = c("#E15759", "#4E79A7")) +
  labs(
    x = "Tratamiento antiparasitario",
    y = "HPG",
    title = "Conteo de HPG post-tratamiento en ovinos Corriedale"
  ) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "none")
```

Interpretación gráfica. El histograma muestra una distribución fuertemente sesgada a la derecha en ambos grupos, con la mayoría de las observaciones concentradas en valores bajos y unas pocas observaciones en valores altos. El gráfico Q-Q revela que los puntos se despegan de la línea teórica en el extremo superior derecho, indicando cola derecha pesada. El boxplot confirma la presencia de valores atípicos hacia arriba en el grupo de ivermectina.

Pruebas formales de normalidad

```
# Shapiro-Wilk por grupo
shapiro_iver <- shapiro.test(hpg$hpg[hpg$tratamiento == "Ivermectina"])
shapiro_albe <- shapiro.test(hpg$hpg[hpg$tratamiento == "Albendazol"])

cat("Shapiro-Wilk Ivermectina: W =", round(shapiro_iver$statistic, 4),
    ", p =", round(shapiro_iver$p.value, 4), "\n")
cat("Shapiro-Wilk Albendazol:  W =", round(shapiro_albe$statistic, 4),
    ", p =", round(shapiro_albe$p.value, 4), "\n")
```

Salida esperada:

```
Shapiro-Wilk Ivermectina: W = 0.7962 , p = 0.0078
Shapiro-Wilk Albendazol:  W = 0.8238 , p = 0.0170
```

Ambos valores *p* están por debajo de 0.05, lo que lleva a rechazar la hipótesis nula de normalidad en ambos grupos. Combinado con la evidencia gráfica, la conclusión es clara: estos datos no se distribuyen normalmente.

```
# Medidas de forma: asimetría y curtosis
hpg %>%
  group_by(tratamiento) %>%
  summarise(
    n       = n(),
    media   = mean(hpg),
    mediana = median(hpg),
    DE      = sd(hpg),
    asimetria = moments::skewness(hpg),
    curtosis  = moments::kurtosis(hpg) - 3  # Exceso de curtosis
  )
```

Salida esperada:

```
# A tibble: 2 × 7
  tratamiento     n media mediana    DE asimetria curtosis
```

```
  <fct>       <int> <dbl>   <dbl> <dbl>     <dbl>    <dbl>
1 Albendazol     12 158.    100   186.      1.26    0.616
2 Ivermectina    12 529.    175   925.      1.91    2.78
```

La asimetría positiva (1.26 y 1.91) y el exceso de curtosis confirman lo que los gráficos mostraban. Valores de asimetría mayores a 1 en valor absoluto se consideran indicativos de asimetría fuerte. La media supera a la mediana por un factor de 1.6 en albendazol y 3 en ivermectina.

Evaluación de homocedasticidad

```
# Prueba de Levene (basada en la mediana, más robusta)
levene_hpg <- car::leveneTest(hpg ~ tratamiento, data = hpg)
print(levene_hpg)
```

Salida esperada:

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group 1 2.1048 0.1613
      22
```

La prueba de Levene no rechaza homocedasticidad ($p = 0.16$). Pero atención: esto no significa que las varianzas sean iguales. Con $n = 12$ por grupo, la potencia de Levene es limitada. La desviación estándar del grupo ivermectina (925) es cinco veces mayor que la del grupo albendazol (186). La prueba no detectó esa diferencia porque, con datos tan asimétricos y muestras pequeñas, las desviaciones absolutas respecto a la mediana son ruidosas.

```
# Prueba de Fligner-Killeen (no paramétrica, más apropiada aquí)
fligner.test(hpg ~ tratamiento, data = hpg)
```

Salida esperada:

```
    Fligner-Killeen:med chi-squared test

data:  hpg by tratamiento
Fligner-Killeen:med chi-squared = 2.8703, df = 1, p-value = 0.09025
```

Fligner-Killeen tampoco rechaza ($p = 0.09$), aunque el valor p es más bajo. La diferencia de varianzas existe, pero la prueba no tiene suficiente potencia para declararla significativa con estos tamaños de muestra.

Intento de transformación logarítmica

```
# Transformación log(HPG + 1) para manejar los ceros
hpg$log_hpg <- log(hpg$hpg + 1)

# Shapiro-Wilk sobre los datos transformados
shapiro.test(hpg$log_hpg[hpg$tratamiento == "Ivermectina"])
shapiro.test(hpg$log_hpg[hpg$tratamiento == "Albendazol"])

# Q-Q de los datos transformados
par(mfrow = c(1, 2))
qqPlot(hpg$log_hpg[hpg$tratamiento == "Ivermectina"],
       main = "Q-Q log(HPG+1) Ivermectina", ylab = "log(HPG+1)",
       pch = 19, col = "#4E79A7", id = FALSE)
qqPlot(hpg$log_hpg[hpg$tratamiento == "Albendazol"],
       main = "Q-Q log(HPG+1) Albendazol", ylab = "log(HPG+1)",
       pch = 19, col = "#E15759", id = FALSE)
par(mfrow = c(1, 1))
```

Salida esperada (Shapiro-Wilk de datos transformados):

```
Shapiro-Wilk Ivermectina log: W = 0.9211 , p = 0.2976
Shapiro-Wilk Albendazol log: W = 0.8802 , p = 0.0872
```

La transformación logarítmica mejora la situación: Shapiro-Wilk ya no rechaza normalidad en ninguno de los dos grupos ($p > 0.05$). El investigador tiene ahora dos opciones legítimas: (a) aplicar una prueba *t* sobre los datos transformados e interpretar los resultados en escala logarítmica, o (b) usar Mann-Whitney sobre los datos originales. Ambas son defendibles. La elección depende de la pregunta de investigación: si interesa la razón de medias geométricas de HPG, la opción (a) es natural. Si interesa comparar las distribuciones completas de respuesta a los tratamientos, la opción (b) es más directa.

Ejemplo 2: Ganancia de peso en pollos de engorde

```
# Ganancia diaria de peso (g/día) en pollos Ross 308 alimentados con
# tres dietas: control, dieta con probiótico, dieta con prebiótico.
# Ensayo en una granja avícola de Pereira, Risaralda, Colombia.
# 15 pollos por tratamiento, medidos durante 21 días (datos simulados
# basados en rangos publicados en estudios de aditivos en dietas avícolas).

set.seed(42)  # Reproducibilidad
pollos <- data.frame(
```

```
  dieta = factor(rep(c("Control", "Probiótico", "Prebiótico"), each = 15)),
  gdp = c(
    # Control: distribución aproximadamente normal
    round(rnorm(15, mean = 58, sd = 5), 1),
    # Probiótico: distribución aproximadamente normal, media ligeramente mayor
    round(rnorm(15, mean = 62, sd = 5), 1),
    # Prebiótico: distribución aproximadamente normal
    round(rnorm(15, mean = 60, sd = 5.5), 1)
  )
)

# Estadísticas descriptivas por grupo
pollos %>%
  group_by(dieta) %>%
  summarise(
    n       = n(),
    media   = round(mean(gdp), 2),
    mediana = round(median(gdp), 2),
    DE      = round(sd(gdp), 2),
    min     = min(gdp),
    max     = max(gdp)
  )
```

Salida esperada:

```
# A tibble: 3 × 7
  dieta         n media mediana    DE   min   max
  <fct>     <int> <dbl>   <dbl> <dbl> <dbl> <dbl>
1 Control      15 57.1    57.2   5.2 48.3 65.2
2 Prebiótico   15 60.5    60.6   5.8 50.0 70.1
3 Probiótico   15 61.2    60.8   5.0 52.7 71.4
```

Aquí la media y la mediana son similares en los tres grupos, las desviaciones estándar son comparables y el rango de valores luce razonable.

```
# Evaluación gráfica
# Q-Q plots por grupo
ggplot(pollos, aes(sample = gdp)) +
  stat_qq(color = "#4E79A7", size = 2) +
  stat_qq_line(color = "#E15759", linewidth = 0.8) +
  facet_wrap(~ dieta) +
  labs(
    x = "Cuantiles teóricos (normal)",
    y = "Cuantiles observados (GDP, g/día)",
    title = "Gráficos Q-Q de ganancia diaria de peso por dieta"
  ) +
```

```
  theme_minimal(base_size = 12)

# Boxplot comparativo
ggplot(pollos, aes(x = dieta, y = gdp, fill = dieta)) +
  geom_boxplot(alpha = 0.7) +
  scale_fill_manual(values = c("#4E79A7", "#F28E2B", "#59A14F")) +
  labs(
    x = "Tipo de dieta",
    y = "Ganancia diaria de peso (g/día)",
    title = "GDP en pollos Ross 308 por tipo de suplementación"
  ) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "none")

# Pruebas formales
# Shapiro-Wilk por grupo
pollos %>%
  group_by(dieta) %>%
  summarise(
    W_shapiro = shapiro.test(gdp)$statistic,
    p_shapiro = shapiro.test(gdp)$p.value
  )
```

Salida esperada:

```
# A tibble: 3 × 3
  dieta      W_shapiro p_shapiro
  <fct>          <dbl>     <dbl>
1 Control        0.954     0.592
2 Prebiótico     0.969     0.842
3 Probiótico     0.976     0.932
```

Los tres valores *p* de Shapiro-Wilk son muy superiores a 0.05. Los gráficos Q-Q muestran puntos alineados sobre la recta teórica sin desviaciones sistemáticas. La normalidad no se rechaza en ningún grupo.

```
# Homocedasticidad
car::leveneTest(gdp ~ dieta, data = pollos)
```

Salida esperada:

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group 2 0.2013 0.8184
      42
```

Las varianzas no difieren significativamente entre grupos ($p = 0.82$). Los supuestos paramétricos se cumplen razonablemente. En este caso, un ANOVA de un factor es el método apropiado, y usar Kruskal-Wallis sacrificaría potencia sin necesidad.

Decisión contrastada entre los dos ejemplos

Tabla 1
Comparación de la evaluación de supuestos entre dos conjuntos de datos veterinarios.

Criterio	HPG en ovinos	GDP en pollos
Tipo de variable	Conteo, escala de razón	Continua, escala de razón
Normalidad (gráfica)	Fuertemente asimétrica	Aproximadamente simétrica
Normalidad (Shapiro-Wilk)	Rechazada ($p < 0.02$)	No rechazada ($p > 0.59$)
Homocedasticidad	Cuestionable (DE 5:1)	Cumplida ($p = 0.82$)
Transformación log	Normaliza razonablemente	No necesaria
Decisión recomendada	Mann-Whitney o *t* sobre log (HPG+1), o GLM binomial negativo	ANOVA de un factor

1.9. Estudio de caso integrador: conteo de HPG en ovinos y ganancia de peso en pollos de engorde

- **Caso 1: Eficacia antihelmíntica en ovinos Análisis completo**

Contexto

Un médico veterinario en la provincia de Cotopaxi, Ecuador, quiere comparar la eficacia de ivermectina y albendazol para reducir la carga de nematodos gastrointestinales en ovinos de raza Corriedale en pastoreo. Selecciona 24 ovejas con conteos de HPG previos al tratamiento superiores a 200 y las asigna aleatoriamente a dos grupos de 12. Administra los tratamientos y mide HPG 14 días después.

Los datos post-tratamiento son los presentados en la sección 1.8 (Ejemplo 1).

Justificación de la vía no paramétrica

El investigador evaluó los supuestos y encontró:

1. Distribución fuertemente asimétrica en ambos grupos (asimetría > 1).
2. Shapiro-Wilk rechaza normalidad en ambos grupos.
3. Las varianzas difieren en un factor de 5:1.
4. Hay valores cero (animales donde el tratamiento eliminó completamente la carga parasitaria) que crean una distribución mixta (punto de masa en cero + distribución continua para valores positivos).

La transformación logarítmica mejora la normalidad, pero no resuelve completamente el problema de los ceros (log(0+1) = 0 genera un punto de acumulación artificial). El investigador decide usar la prueba de Mann-Whitney como análisis principal y la prueba *t* sobre datos transformados como análisis complementario.

Análisis con código en R

```
# ---- ANÁLISIS NO PARAMÉTRICO: Mann-Whitney ----

# Los datos ya están cargados como 'hpg' del ejemplo anterior.

# Prueba de Mann-Whitney (equivalente a Wilcoxon rank-sum)
# wilcox.test() del paquete base de R
mw_resultado <- wilcox.test(hpg ~ tratamiento, data = hpg,
                            exact = FALSE,      # Aproximación normal
                            correct = TRUE,     # Corrección de continuidad
                            conf.int = TRUE,    # Intervalo de confianza
                            conf.level = 0.95)

print(mw_resultado)
```

Salida esperada:

```
    Wilcoxon rank sum test with continuity correction

data:  hpg by tratamiento
W = 49.5, p-value = 0.1655
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
```

```
 -50.00002 349.99991
sample estimates:
difference in location
                100
```

Interpretación detallada:

- **W = 49.5:** Es el estadístico de la prueba, que corresponde a la suma de rangos del grupo que R identifica como el primero (Albendazol, en este caso, por orden alfabético). Más precisamente, R reporta el estadístico U de Mann-Whitney.

- **p = 0.1655:** No hay evidencia suficiente para rechazar la hipótesis nula de que las distribuciones de HPG post-tratamiento son iguales entre los dos antiparasitarios (α = 0.05).

- **Diferencia de localización estimada = 100:** Es la estimación de Hodges-Lehmann de la diferencia mediana entre las distribuciones. Indica que la mediana de HPG con ivermectina fue estimada como 100 unidades mayor que con albendazol, pero el intervalo de confianza del 95% (–50 a 350) incluye el cero.

```
# ---- ANÁLISIS PARAMÉTRICO COMPLEMENTARIO: t de Welch sobre log(HPG+1) ----

t_log <- t.test(log_hpg ~ tratamiento, data = hpg, var.equal = FALSE)
print(t_log)
```

Salida esperada:

```
    Welch Two Sample t-test

data:  log_hpg by tratamiento
t = -1.0249, df = 20.445, p-value = 0.3172
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4041  0.8240
sample estimates:
mean in group Albendazol mean in group Ivermectina
                   3.590                    4.380
```

La prueba *t* de Welch sobre datos transformados tampoco encuentra diferencia significativa (p = 0.32). Las medias de log(HPG+1) son 3.59 (Albendazol) y 4.38 (Ivermectina). Retrotraduciendo: las medias geométricas de (HPG+1) son exp(3.59) ≈ 36 y exp(4.38) ≈ 80, lo cual indica que la

media geométrica de HPG en el grupo ivermectina es aproximadamente el doble que en el grupo albendazol, pero la diferencia no alcanza significancia estadística con estos tamaños de muestra.

```
# ---- ¿QUÉ HABRÍA PASADO CON UNA t DE STUDENT CLÁSICA SIN VERIFICACIÓN? ----

t_cruda <- t.test(hpg ~ tratamiento, data = hpg, var.equal = TRUE)
print(t_cruda)
```

Salida esperada:

```
    Two Sample t-test

data:  hpg by tratamiento
t = -1.3515, df = 22, p-value = 0.1903
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -940.21  199.55
sample estimates:
mean in group Albendazol mean in group Ivermectina
                 158.333                  529.167
```

La prueba *t* clásica (asumiendo varianzas iguales) da $p = 0.19$. El intervalo de confianza para la diferencia de medias es enormemente ancho: de –940 a +200 HPG. Este intervalo está inflado porque la media del grupo ivermectina está arrastrada por los valores extremos (1800 y 3200), y la varianza estimada combinada no refleja la realidad de ninguno de los dos grupos.

Lección del caso. Las tres pruebas coinciden en no rechazar la hipótesis nula, pero por razones distintas y con niveles de confianza diferentes en los resultados:

- La prueba *t* clásica produce un resultado que técnicamente es inválido porque viola normalidad y homocedasticidad. Que la conclusión coincida con la de Mann-Whitney en este caso es casualidad, no garantía.

- La prueba *t* de Welch sobre datos transformados es válida (los datos transformados son aproximadamente normales y Welch no requiere homocedasticidad), pero responde una pregunta distinta: compara medias geométricas, no distribuciones.

- Mann-Whitney compara las distribuciones completas sin supuestos distribucionales. Con 12 animales por grupo, su potencia es limitada, pero la inferencia es válida.

- **Caso 2: Ganancia de peso en pollos — Análisis breve**

Para los datos de GDP en pollos (Ejemplo 2), los supuestos paramétricos se cumplen. El análisis procede con ANOVA:

```
# ANOVA de un factor
modelo_anova <- aov(gdp ~ dieta, data = pollos)
summary(modelo_anova)
```

Salida esperada:

```
            Df Sum Sq Mean Sq F value Pr(>F)
dieta        2 136.5   68.27   2.445 0.0990
Residuals   42 1172.8   27.92
```

El ANOVA no detecta diferencias significativas entre dietas ($p = 0.099$). Con $n = 15$ por grupo y una diferencia real entre medias del control y el probiótico de apenas 4 g/día (58 vs. 62), la potencia del estudio es insuficiente para detectar un efecto de esa magnitud. El investigador habría necesitado aproximadamente 25 animales por grupo para detectar una diferencia de 4 g/día con potencia del 80% (cálculo que se puede verificar con *power.anova.test()* en R).

```
# Verificación post hoc de la potencia
power.anova.test(groups = 3, between.var = var(c(58, 62, 60)),
                 within.var = 27.92, sig.level = 0.05, n = NULL,
                 power = 0.80)
```

Este caso ilustra una situación donde Kruskal-Wallis no habría sido apropiado: los datos son normales, homocedásticos y la variable es continua. Usar Kruskal-Wallis aquí habría reducido la potencia ya insuficiente.

1.10. Errores frecuentes y advertencias

1. Aplicar Shapiro-Wilk con $n = 5$ y concluir que "los datos son normales". Con muestras tan pequeñas, la prueba no tiene potencia para detectar desviaciones de la normalidad. Que $p > 0.05$ no significa normalidad; significa que la prueba no puede distinguir la distribución de los datos de una normal. Con 5 observaciones, tampoco los gráficos son informativos. La honestidad estadística requiere reconocer que no se sabe si los datos son normales y elegir un método que no dependa de ese supuesto.

2. Usar la prueba de Bartlett cuando los datos no son normales. Bartlett es muy sensible a la no normalidad. Con datos no normales, rechaza homocedasticidad no porque las varianzas difieran, sino porque la distribución no es normal. La prueba de Levene basada en la mediana o la prueba de Fligner-Killeen son alternativas apropiadas para datos no normales.

3. Evaluar normalidad sobre todos los datos juntos en lugar de por grupo. Si se comparan tres tratamientos, la normalidad debe evaluarse dentro de cada grupo (o, más correctamente, sobre los residuos del modelo). La distribución marginal de todos los datos combinados puede parecer no normal simplemente porque las medias de los grupos difieren, sin que haya un problema de normalidad dentro de cada grupo.

4. Tratar datos ordinales como continuos. Calcular la media de una escala de condición corporal de 1 a 5 o de un puntaje de dolor de 0 a 10 y aplicar una prueba *t* asume que la distancia entre 1 y 2 es igual a la distancia entre 4 y 5, y que los valores intermedios (como 2.5) tienen sentido. Con datos ordinales, los métodos no paramétricos basados en rangos son la elección correcta.

5. Confundir "no significativo en la prueba de normalidad" con "los datos son normales". No rechazar H_0 no es lo mismo que aceptar H_0. La prueba de Shapiro-Wilk no puede demostrar normalidad; solo puede rechazarla. Un resultado no significativo con muestra pequeña es ambiguo: el investigador debería integrar la evidencia gráfica, el conocimiento sobre la naturaleza de la variable y el tamaño de muestra para tomar una decisión informada.

6. Aplicar un método no paramétrico sin verificar independencia. Los métodos no paramétricos clásicos (Mann-Whitney, Kruskal-Wallis, Wilcoxon) requieren que las observaciones sean independientes. Si se midieron los mismos animales antes y después del tratamiento, no se puede usar Mann-Whitney para comparar los dos tiempos; se necesita Wilcoxon para datos pareados. Si los animales comparten corral y se influyen mutuamente, el problema de pseudorreplicación afecta tanto a los métodos paramétricos como a los no paramétricos.

7. Omitir la evaluación de supuestos en la sección de métodos de la tesis. El lector de un artículo o tesis necesita saber *por qué* se eligió un método no paramétrico. Reportar "se usó Kruskal-Wallis" sin explicar que se verificó (y se rechazó) la normalidad, o que la variable era ordinal, deja la selección del método sin justificación.

1.11. Ejercicios de autoevaluación

Nivel básico

Ejercicio 1.1.

Un investigador mide la concentración sérica de calcio (mg/dL) en 30 vacas Holstein antes y después de la suplementación con borogluconato. Quiere saber si la suplementación cambió los niveles de calcio. Indique: (a) ¿Cuáles son los supuestos que debería verificar antes de aplicar una prueba *t* para datos pareados? (b) Si la distribución de las diferencias (post – pre) resulta simétrica, pero con colas pesadas, ¿cuál sería la prueba no paramétrica equivalente?

Respuesta: (a) Normalidad de las diferencias (post – pre), no de las observaciones individuales. No se requiere homocedasticidad porque se trabaja con una sola muestra de diferencias. Se requiere independencia entre las vacas (que el valor de una vaca no dependa del de otra). (b) La prueba de rangos con signo de Wilcoxon (Wilcoxon signed-rank test), que se desarrolla en el capítulo 3.

Ejercicio 1.2.

Un zootecnista quiere comparar el peso al destete (kg) de terneros Brahman criados en tres fincas distintas. Tiene 8, 12 y 10 terneros por finca, respectivamente. El gráfico Q-Q de los residuos del ANOVA muestra una distribución razonablemente simétrica con un solo punto atípico en la finca con 8 terneros. La prueba de Shapiro-Wilk sobre los residuos da $W = 0.96$, $p = 0.33$. La prueba de Levene da $F = 0.85$, $p = 0.44$. ¿Qué método de análisis recomienda y por qué?

Respuesta: ANOVA de un factor. Los supuestos de normalidad y homocedasticidad no se rechazan, los gráficos son compatibles con normalidad (un punto atípico aislado no invalida el análisis con esos tamaños de muestra), y la variable es continua. Usar Kruskal-Wallis en esta situación sacrificaría potencia sin justificación.

Ejercicio 1.3.

Se evaluó el grado de lesión hepática en ratones inoculados con dos cepas de *Fasciola hepatica*. La lesión se clasificó en una escala ordinal de 0 (sin lesión), 1 (leve), 2 (moderada), 3 (grave) y 4 (muy grave). ¿Es apropiado calcular la media de la escala de lesión y aplicar una prueba *t*? Justifique.

Respuesta: No. La variable es ordinal: las distancias entre categorías no son necesariamente iguales (la diferencia clínica entre lesión leve y moderada no tiene por qué ser la misma que entre grave y muy grave). Calcular la media de una escala ordinal carece de sentido conceptual. El análisis correcto usa un método no paramétrico para datos ordinales, como Mann-Whitney.

Nivel intermedio

Ejercicio 1.4.

Los siguientes datos representan el recuento de colonias bacterianas (UFC/mL × 10^3) en muestras de leche de 10 vacas con mastitis subclínica en cada uno de dos hatos:

Hato A: 12, 15, 18, 22, 25, 28, 35, 45, 120, 350
Hato B: 8, 10, 12, 14, 16, 18, 20, 22, 25, 30

(a) Calcule media, mediana y desviación estándar para cada hato.

(b) Construya un gráfico Q-Q para cada hato (puede ser a mano o en R).

(c) Aplique Shapiro-Wilk a cada hato.

(d) Evalúe homocedasticidad con Levene.

(e) ¿Qué método elegiría para comparar los hatos? Justifique.

Respuestas:

a) Hato A: media = 67.0, mediana = 26.5, DE = 102.3. Hato B: media = 17.5, mediana = 17.0, DE = 7.0.

b) El Q-Q del Hato A mostrará dos puntos (120 y 350) claramente despegados de la recta en el extremo superior derecho. El Hato B se alineará razonablemente sobre la recta.
c) Hato A: $W \approx 0.67$, $p \approx 0.001$ (rechaza normalidad). Hato B: $W \approx 0.97$, $p \approx 0.89$ (no rechaza).
d) Las varianzas son radicalmente diferentes (DE de 102 vs. 7; razón de varianzas ≈ 213). Levene rechazaría incluso con $n = 10$.
e) Mann-Whitney. El Hato A tiene distribución fuertemente asimétrica con valores extremos, las varianzas son muy desiguales y el tamaño de muestra es pequeño. La transformación logarítmica podría ayudar, pero la diferencia de variabilidad es tan grande que Mann-Whitney o un GLM binomial negativo son opciones más seguras.

Ejercicio 1.5.

Un tesista aplicó la prueba de normalidad de Shapiro-Wilk a un conjunto de datos de peso al sacrificio ($n = 150$ cerdos) y obtuvo $W = 0.993$, $p = 0.041$. El gráfico Q-Q muestra los puntos prácticamente sobre la línea recta con una desviación mínima en el extremo superior. ¿Debería usar un método no paramétrico para analizar estos datos? Argumente.

Respuesta: No necesariamente. Con $n = 150$, Shapiro-Wilk es tan potente que detecta desviaciones mínimas de la normalidad que no afectan la validez del ANOVA ni de la prueba *t*. El gráfico Q-Q indica que la desviación es trivial. El teorema central del límite protege la inferencia con este tamaño de muestra. Aplicar un método paramétrico es defensible e incluso preferible, porque conserva mayor potencia. Este es un ejemplo clásico de la paradoja de las muestras grandes en pruebas de normalidad: la significancia estadística de la prueba de Shapiro-Wilk no implica significancia práctica de la desviación.

Nivel avanzado

Ejercicio 1.6.

Un estudio evalúa el efecto de tres protocolos de analgesia postoperatoria sobre una escala compuesta de dolor (escala de Glasgow modificada para perros, rango 0–24) en 45 perras sometidas a ovario histerectomía electiva (15 por protocolo). El investigador le envía los datos y le pide que analice si hay diferencias entre protocolos. Usted debe:

(a) Explorar la distribución de los datos: estadísticas descriptivas, gráficos.

(b) Evaluar normalidad y homocedasticidad.

(c) Decidir y justificar el método de análisis.

(d) Ejecutar el análisis completo en R.

(e) Redactar un párrafo de resultados como aparecería en la sección "Resultados" de un artículo veterinario, siguiendo formato APA 7.ª edición.

Rúbrica orientadora:

- La exploración descriptiva debe identificar que la escala de dolor es ordinal (aunque el rango sea amplio, los valores son puntajes discretos con techo y piso).
- La evaluación de supuestos debe concluir que, siendo datos ordinales con una escala acotada, la normalidad es conceptualmente cuestionable incluso si Shapiro-Wilk no rechaza.
- El método correcto es Kruskal-Wallis (o un modelo ordinal si el estudiante tiene formación avanzada). No ANOVA.
- Las comparaciones post hoc, si Kruskal-Wallis es significativo, deben usar pruebas de Dunn o pairwise Wilcoxon con corrección por multiplicidad (Bonferroni, Holm o Benjamini-Hochberg).
- El párrafo de resultados debe reportar medianas (con rangos intercuartílicos), el estadístico de Kruskal-Wallis (H o χ^2), los grados de libertad, el valor *p*, los resultados de las comparaciones post hoc, y un tamaño del efecto (ε^2 o η^2 basado en rangos). Ejemplo: "Las puntuaciones de dolor difirieron entre los tres protocolos de analgesia ($H(2) = 8.45$, $p = .015$, $\varepsilon^2 = .19$). Las comparaciones post hoc con la prueba de Dunn y corrección de Holm indicaron que el protocolo C produjo puntuaciones menores que el protocolo A ($p = .012$), sin diferencias significativas entre los demás pares."

Ejercicio 1.7.

Genere en R un conjunto de datos simulados que represente el tiempo de supervivencia (días) de 20 gatos con linfoma tratados con quimioterapia (grupo A) y 20 gatos con tratamiento paliativo (grupo B). Use una distribución exponencial para simular los tiempos (con *rexp()*). Explore los supuestos, intente una transformación logarítmica y argumente por qué ni la prueba *t* original ni la transformada son la herramienta ideal para este tipo de datos, y por qué se necesita un método específico para datos de supervivencia (que se cubrirá en el capítulo 9).

Rúbrica orientadora: El estudiante debe reconocer que los datos de supervivencia tienen dos características que los hacen inadecuados para una prueba *t*: (1) la distribución es asimétrica y con cola derecha (los tiempos de supervivencia no pueden ser negativos y su distribución suele ser sesgada), y (2) los datos suelen estar censurados (algunos gatos aún están vivos al final del estudio, por lo que su tiempo real de supervivencia se desconoce). La transformación logarítmica puede simetrizar la distribución, pero no resuelve el problema de la censura. Los métodos de Kaplan-Meier y log-rank, que se presentan en el capítulo 9, están diseñados específicamente para manejar ambas situaciones.

1.12. Conclusiones del capítulo

Los métodos paramétricos funcionan bien cuando sus supuestos se cumplen, y en investigación veterinaria hay situaciones ensayos con muestras moderadas, variables fisiológicas con distribución razonablemente simétrica donde se cumplen. No se trata de abandonar las pruebas paramétricas, sino de verificar si las condiciones que requieren están presentes en los datos y actuar en consecuencia.

La evaluación de supuestos no es un trámite burocrático que se reporta en la tesis porque el comité lo pide. Es una etapa diagnóstica que determina la validez de toda la inferencia posterior. Un histograma, un gráfico Q-Q y una prueba de Shapiro-Wilk toman menos de dos minutos en R. Lo que ahorran en decisiones erróneas vale mucho más que ese tiempo.

Los métodos no paramétricos no son la opción “de segunda”, sino la herramienta correcta para datos ordinales, distribuciones fuertemente asimétricas, muestras pequeñas donde los supuestos no se pueden evaluar, y situaciones donde la media no es un estadístico representativo. Tienen sus propios supuestos independencia, escala al menos ordinal, y en algunos casos forma similar de las distribuciones que el investigador debe conocer y verificar.

Entre el extremo de “siempre paramétrico” y “siempre no paramétrico” existe un espacio amplio de herramientas: transformaciones, métodos robustos, modelos lineales generalizados. El investigador que las conoce toma mejores decisiones. Los capítulos siguientes desarrollan las pruebas no paramétricas una a una, partiendo de la prueba de signos y la prueba de Wilcoxon para datos pareados (capítulo 3) y avanzando hacia métodos más complejos.

Capítulo 2

Pruebas para una muestra, datos pareados y dos muestras independientes

Introducción

Este capítulo trabaja las pruebas no paramétricas para los tres diseños comparativos más frecuentes en investigación veterinaria. Para datos pareados se desarrollan dos pruebas: la prueba de signos, que analiza únicamente la dirección del cambio en cada unidad experimental y requiere solo el supuesto de independencia entre pares, y la prueba de rangos con signo de Wilcoxon, que incorpora la magnitud de las diferencias mediante la asignación de rangos y gana potencia a cambio de asumir simetría en la distribución de las diferencias. Para dos muestras independientes se presentan la prueba de Mann-Whitney U, con la discusión detallada sobre si compara medianas o distribuciones completas y las condiciones bajo las cuales cada interpretación es válida, y la prueba de Brunner-Munzel como alternativa cuando las distribuciones difieren en forma y varianza. Se trabaja el estimador de Hodges-Lehmann como medida de localización con su intervalo de confianza, las medidas de tamaño del efecto r de rango biserial y A de Vargha y Delaney, y la comparación sistemática con las pruebas paramétricas equivalentes (prueba t pareada y prueba t para muestras independientes) para cuantificar qué se pierde en potencia al elegir la vía no paramétrica y qué se gana en validez cuando los supuestos fallan. Todo se implementa en R mostrando las diferencias entre wilcox.test del paquete base y wilcox_test de coin en el cálculo de valores p exactos, asintóticos y con corrección de continuidad.

Objetivos de aprendizaje

Al finalizar este capítulo, el lector será capaz de:

1. **Identificar** las condiciones experimentales y las características de los datos que justifican el uso de la prueba de signos, la prueba de rangos con signo de Wilcoxon y la prueba de Mann-Whitney U en lugar de sus equivalentes paramétricos.
2. **Aplicar** paso a paso cada prueba a conjuntos de datos veterinarios, tanto de forma manual como mediante código en R, seleccionando la función y los argumentos apropiados.
3. **Interpretar** los resultados estadísticos (estadístico de prueba, valor *p*, intervalo de confianza, estimador de Hodges-Lehmann) en el contexto biológico y clínico del problema veterinario que motivó el análisis.

4. **Contrastar** las pruebas no paramétricas con sus equivalentes paramétricos (prueba *t* pareada, prueba *t* para muestras independientes), evaluando las consecuencias prácticas de elegir una u otra cuando los supuestos se cumplen y cuando no.
5. **Calcular e informar** medidas de tamaño del efecto apropiadas para pruebas no paramétricas (*r* de rango biserial, *A* de Vargha y Delaney) y reconocer la diferencia entre significancia estadística y relevancia clínica.

2.1. Prueba de signos

El problema que motiva la prueba

Un médico veterinario en una clínica de pequeñas especies de Quito evalúa un nuevo protocolo analgésico multimodal para el control del dolor posoperatorio en gatas sometidas a ovariohisterectomía. Antes de la cirugía y 6 horas después, aplica la escala compuesta de dolor de Glasgow en su versión felina corta (Glasgow Composite Measure Pain Scale Feline, CMPS-F), que arroja puntajes enteros entre 0 y 20. Tiene 9 gatas en el estudio. Los puntajes son ordinales, la muestra es pequeña, y la distribución de las diferencias pre-post no tiene por qué ser simétrica. La prueba *t* pareada exige normalidad de las diferencias, y con 9 observaciones no hay forma de verificar ese supuesto con potencia razonable. ¿Existe una alternativa que no requiera asumir nada sobre la forma de la distribución?

La prueba de signos es esa alternativa. Es la más antigua de las pruebas no paramétricas y la más sencilla de comprender.

Lógica conceptual

La prueba de signos ignora la magnitud de las diferencias y se concentra exclusivamente en su dirección: ¿mejoró o empeoró cada animal después del tratamiento? Si el tratamiento no tiene efecto, la probabilidad de que la diferencia sea positiva debería ser 0.5, igual que la probabilidad de que sea negativa. Lanzar una moneda justa.

Para cada par de observaciones (antes–después en el mismo animal, o pata tratada–pata control en el mismo perro), se calcula la diferencia $d_i = X_{i2} - X_{i1}$. Se descartan los empates (diferencias iguales a cero). Entre las diferencias restantes, se cuenta cuántas son positivas.

Si *n'* es el número de diferencias no nulas, el número de signos positivos S^+ sigue una distribución binomial con parámetros *n'* y $p = 0.5$ bajo la hipótesis nula de que el tratamiento no modifica la mediana de las diferencias.

Formulación estadística

Hipótesis:

- H_0: La mediana de las diferencias poblacionales es igual a cero (equivalente a $P(X_{i2} > X_{i1}) = 0.5$).
- H_1 (bilateral): La mediana de las diferencias poblacionales difiere de cero.

Estadístico de prueba:

$$S^+ = \sum_{i=1}^{n'} I\,(d_i > 0)$$

donde $I(\cdot)$ es la función indicadora que vale 1 cuando la condición se cumple y 0 en caso contrario, y *n'* es el número de pares con diferencia distinta de cero.

Distribución bajo H_0:

$$S^+ \sim \text{Binomial}(n', 0.5)$$

Valor *p* (bilateral):

$$p = 2 \times \min[P(X \leq S^+),\ P(X \geq S^+)], \quad X \sim \text{Binomial}(n', 0.5)$$

No hay aproximación asintótica que recurrir aquí: el valor *p* exacto se calcula directamente con la distribución binomial. Esto hace que la prueba de signos sea aplicable sin importar cuán pequeña sea la muestra. Con 4 pares y los 4 mostrando mejoría, el valor *p* bilateral exacto es $2 \times (0.5)^4 = 0.125$; ni siquiera la unanimidad alcanza significancia con tan pocos datos.

Supuestos

Que la prueba de signos sea simple no significa que carezca de supuestos:

1. **Datos pareados:** Cada observación del grupo "después" debe corresponder a la misma unidad experimental que la observación "antes". Si se comparan animales distintos, la prueba no aplica.
2. **Variable al menos ordinal:** Se necesita poder determinar si la diferencia es positiva, negativa o cero. No se requiere que la variable sea de intervalo o razón.
3. **Independencia entre pares:** El resultado de un animal no debe influir en el de otro. Si se miden varias gatas alojadas juntas y el estrés social afecta los puntajes de dolor, la independencia se viola.

Ejemplo resuelto: Dolor posoperatorio en gatas

Los datos del estudio de Quito se presentan en la Tabla 2.1.

Tabla 2

Puntajes de dolor (CMPS-F) en 9 gatas antes de la cirugía y 6 horas después de ovario histerectomía con protocolo analgésico multimodal.

Gata	Pre-cirugía (X_1)	Post-cirugía (X_2)	Diferencia ($d = X_2 - X_1$)	Signo
1	2	8	+6	+
2	1	5	+4	+
3	0	3	+3	+
4	1	7	+6	+
5	3	4	+1	+
6	2	2	0	(descartado)
7	1	6	+5	+
8	0	9	+9	+
9	2	5	+3	+

Nota: los puntajes pre-cirugía reflejan la evaluación basal; los post-cirugía reflejan el dolor posoperatorio residual a pesar del protocolo analgésico. El aumento indica que la cirugía produce dolor detectable aun con analgesia.

Resolución paso a paso:

1. Diferencias no nulas: $n' = 8$ (se descarta la gata 6).
2. Signos positivos: $S^+ = 8$.

3. Bajo H_0: $S^+ \sim$ Binomial (8, 0.5).
4. $P\ (X \geq 8) = (0.5)^{\ 8} = 0.00391$.
5. Valor p bilateral = 2 × 0.00391 = 0.00781.

Con $p = 0.0078$, se rechaza H_0 al nivel $\alpha = 0.05$. Todas las gatas con diferencia no nula mostraron aumento en el puntaje de dolor, lo cual indica que la cirugía incrementa el dolor detectado por la escala CMPS-F incluso bajo el protocolo analgésico evaluado. Esto no significa que el protocolo sea inútil; significa que no elimina el dolor por completo, y el diseño del estudio (sin grupo control sin analgesia) no permite cuantificar cuánto reduce.

Limitaciones de la prueba de signos

La prueba de signos es deliberadamente conservadora: descarta toda la información sobre la magnitud de las diferencias. Una gata con diferencia de +1 y otra con diferencia de +9 contribuyen exactamente lo mismo al estadístico. Esto la hace robusta ante datos ordinales o con valores atípicos extremos, pero le resta potencia estadística cuando las diferencias tienen magnitudes informativas y la distribución es razonablemente simétrica. En esos casos, la prueba de rangos con signo de Wilcoxon aprovecha mejor la información disponible.

2.2. Prueba de rangos con signo de Wilcoxon

Del signo al rango: recuperar información sobre la magnitud

Frank Wilcoxon propuso en 1945 una prueba que conserva la dirección de cada diferencia, pero además incorpora su magnitud relativa. En vez de contar cuántas diferencias son positivas, asigna rangos a los valores absolutos de las diferencias y luego suma por separado los rangos de las diferencias positivas y los de las negativas. Si el tratamiento no tiene efecto, ambas sumas deberían ser similares. Si un tratamiento produce mejoras consistentes y grandes, los rangos altos se acumularán en el lado positivo.

Procedimiento paso a paso

1. Calcular las diferencias $d_i = X_{i2} - X_{i1}$ para cada par.
2. Eliminar las diferencias iguales a cero. Sea n' el número de diferencias restantes.

3. Tomar los valores absolutos $|d_i|$ y ordenarlos de menor a mayor, asignando rangos del 1 al n'.
4. Si hay empates en $|d_i|$, asignar a cada valor empatado el promedio de los rangos que les corresponderían (rangos medios).
5. Asignar a cada rango el signo de la diferencia original.
6. Calcular W^+ = suma de los rangos con signo positivo, y W^- = suma de los rangos con signo negativo.
7. El estadístico de prueba es $W = \min(W^+, W^-)$ en la formulación clásica de tablas, o $T = W^+$ en la formulación que usan la mayoría de los programas informáticos.

Formulación estadística

Estadístico:

$$T^+ = W^+ = \sum_{i=1}^{n'} R_i \cdot I(d_i > 0)$$

donde R_i es el rango del valor absoluto de la i-ésima diferencia no nula.

Nótese que $W^+ + W^- = n'(n' + 1)/2$, de modo que conocer uno determina el otro.

Distribución bajo H_0:

Para muestras pequeñas (típicamente $n' \leq 20$), se usa la distribución exacta de T^+, que se construye enumerando las 2^n' asignaciones posibles de signos a los rangos. Para muestras mayores, se emplea la aproximación normal:

$$Z = \frac{T^+ - \mu_T}{\sigma_T}$$

donde:

$$\mu_T = \frac{n'(n' + 1)}{4}$$

$$\sigma_T = \sqrt{\frac{n'(n'+1)(2n'+1)}{24} - \frac{\sum_{j=1}^{g} t_j\,(t_j-1)(t_j+1)}{48}}$$

En esta última expresión, g es el número de grupos de empates y t_j es el tamaño del j-ésimo grupo de empates. El segundo término del radicando es la corrección por empates; si no hay empates, se omite.

Algunos programas aplican además una corrección de continuidad restando 0.5 al numerador del valor absoluto de *Z*. La función *wilcox.test()* de R aplica esta corrección por defecto cuando usa la aproximación normal, lo cual puede producir valores *p* ligeramente mayores que los de otras implementaciones.

Supuestos

1. **Datos pareados** con pares independientes entre sí.
2. **Variable al menos de intervalo** para que las diferencias y sus magnitudes tengan sentido. Si la variable es estrictamente ordinal (como puntajes de condición corporal en escala de 1 a 5 con incrementos de 0.25), las diferencias pueden calcularse, pero sus magnitudes relativas son discutibles. En esos casos, la prueba de signos es más honesta.
3. **Distribución simétrica de las diferencias** bajo H_0. Este supuesto se olvida con frecuencia. Si las diferencias poblacionales tienen una distribución muy asimétrica, la prueba de rangos con signo puede arrojar resultados engañosos porque la distribución nula del estadístico se construye asumiendo que cada diferencia tiene la misma probabilidad de ser positiva o negativa para un rango dado. Cuando la asimetría es pronunciada, la prueba de signos (que no requiere simetría) o la prueba de Brunner-Munzel adaptada para datos pareados pueden ser preferibles.

Ejemplo resuelto: Conteo de huevos por gramo (HPG) antes y después de desparasitación en ovinos

Un estudio en una finca ovina de la Sierra central ecuatoriana evalúa la eficacia de albendazol (7.5 mg/kg PO) contra nematodos gastrointestinales en 12 corderos Corriedale de 6 meses. Se realiza

conteo de huevos por gramo de heces (HPG) mediante técnica de McMaster antes del tratamiento (día 0) y 14 días después (día 14). La Tabla 2.2 presenta los datos.

Tabla 3

Conteo de huevos por gramo (HPG) en 12 corderos Corriedale antes y 14 días después de la administración de albendazol.

Cordero	HPG día 0	HPG día 14	d_i = día 14 − día 0		d_i	
1	1200	150	−1050	1050	9	−9
2	800	100	−700	700	6	−6
3	450	50	−400	3	−3	
4	2400	300	−2100	2100	12	−12
5	600	200	−400	400	3	−3
6	350	0	−350	350	1	−1
7	1800	400	−1400	1400	11	−11
8	950	200	−750	750	7	−7
9	500	100	−400	400	3	−3
10	1100	250	−850	850	8	−8
11	1500	150	−1350	1350	10	−10
12	700	0	−700	700	6	−6

Nota: los corderos 3, 5 y 9 tienen $|d_i| = 400$. Se les asigna el rango medio $(2 + 3 + 4)/3 = 3$. Los corderos 2 y 12 tienen $|d_i| = 700$; rango medio $(5 + 6)/2 = 5.5$. En la tabla se muestran rangos redondeados para claridad; en el cálculo se usan los rangos medios exactos.

Recalculemos los rangos con precisión. Ordenando los $|d_i|$:

Posición		d_i	
1	350	6	1
2, 3, 4	400	3, 5, 9	(2+3+4)/3 = 3
5, 6	700	2, 12	(5+6)/2 = 5.5
7	750	8	7
8	850	10	8
9	1050	1	9
10	1350	11	10

11	1400	7	11
12	2100	4	12

Todos los signos son negativos, porque todos los corderos redujeron su HPG.

$W^+ = 0$ (no hay rangos positivos).

$W^- = 1 + 3 + 3 + 3 + 5.5 + 5.5 + 7 + 8 + 9 + 10 + 11 + 12 = 78$.

Verificación: $n'(n' + 1)/2 = 12 \times 13/2 = 78$. ✓

Para la prueba bilateral: $T = W^+ = 0$. El valor mínimo posible de W^+ es 0, que se alcanza solo cuando todas las diferencias son negativas. Con $n' = 12$, la probabilidad exacta de $W^+ = 0$ bajo H_0 es $(1/2)^{12} = 0.000244$. El valor p bilateral exacto es $2 \times 0.000244 = 0.000488$.

Se rechaza H_0 con amplio margen. El albendazol redujo el HPG en todos los corderos, y la magnitud de la reducción es consistente a través de los animales. Desde el punto de vista clínico, la reducción del conteo fecal de huevos (FECRT) se calcula como:

$$\text{FECRT} = \left(1 - \frac{\bar{X}_{\text{post}}}{\bar{X}_{\text{pre}}}\right) \times 100$$

En este caso, la media aritmética pre = 862.5 HPG, post = 158.3 HPG, FECRT = 81.6%. Un FECRT < 95% con límite inferior del intervalo de confianza < 90% sugiere resistencia antihelmíntica según las directrices de la World Association for the Advancement of Veterinary Parasitology (Coles et al., 1992). El análisis estadístico confirma que la reducción existe, pero la interpretación parasitológica indica que puede ser insuficiente.

2.3. Prueba de Wilcoxon para una muestra contra una mediana teórica

El problema

Un zootecnista en Cochabamba, Bolivia, quiere saber si la condición corporal promedio de un hato de 15 vacas Holstein mestizas en el primer tercio de lactancia difiere de la mediana de referencia de 3.0 (en escala de 1 a 5 con incrementos de 0.25). No hay grupo de comparación; solo quiere evaluar si el hato está por debajo, por encima o en el valor esperado.

La prueba *t* para una muestra requiere normalidad, y los puntajes de condición corporal son ordinales por definición (aunque tratados como continuos por costumbre). Con 15 observaciones y una escala acotada, la normalidad es cuestionable.

Procedimiento

La prueba de Wilcoxon para una muestra es idéntica a la prueba de rangos con signo, pero en vez de calcular diferencias entre pares, se calculan diferencias entre cada observación y la mediana teórica m_0:

$$d_i = X_i - m_0$$

Se eliminan los $d_i = 0$, se asignan rangos a $|d_i|$ y se procede como en la Sección 2.2.

Hipótesis:

- H_0: La mediana poblacional $\theta = m_0$ (en el ejemplo, $m_0 = 3.0$).
- H_1: $\theta \neq m_0$.

Ejemplo resuelto

Tabla 4

Puntajes de condición corporal (escala 1–5) en 15 vacas Holstein mestizas, primer tercio de lactancia. Finca en Cochabamba, Bolivia.

Vaca	**CC**	**d_i = CC − 3.0**		**d_i**	
1	2.50	−0.50	0.50	5.5	−5.5
2	2.75	−0.25	0.25	2	−2
3	3.00	0.00	—	—	(descartado)
4	2.25	−0.75	0.75	9.5	−9.5
5	3.25	+0.25	0.25	2	+2
6	2.50	−0.50	0.50	5.5	−5.5
7	2.75	−0.25	0.25	2	−2
8	3.50	+0.50	0.50	5.5	+5.5
9	2.00	−1.00	1.00	12	−12
10	3.00	0.00	—	—	(descartado)

11	2.50	−0.50	0.50	5.5	−5.5
12	2.75	−0.25	0.25	2	−2
13	3.25	+0.25	0.25	2	+2
14	2.25	−0.75	0.75	9.5	−9.5
15	3.00	0.00	—	—	(descartado)

Diferencias no nulas: $n' = 12$.

Rangos medios para empates: - $|d| = 0.25$ aparece 5 veces → posiciones 1–5 → rango medio = (1+2+3+4+5)/5 = 3. Pero revisemos: tenemos |d| ordenados: 0.25 (×5), 0.50 (×4), 0.75 (×2), 1.00 (×1). - 0.25 → posiciones 1–5, rango medio = 3. - 0.50 → posiciones 6–9, rango medio = 7.5. - 0.75 → posiciones 10–11, rango medio = 10.5. - 1.00 → posición 12, rango = 12.

Recalculando:

Vaca	d_i		d_i	
1	−0.50	0.50	7.5	−7.5
2	−0.25	0.25	3	−3
4	−0.75	0.75	10.5	−10.5
5	+0.25	0.25	3	+3
6	−0.50	0.50	7.5	−7.5
7	−0.25	0.25	3	−3
8	+0.50	0.50	7.5	+7.5
9	−1.00	1.00	12	−12
11	−0.50	0.50	7.5	−7.5
12	−0.25	0.25	3	−3
13	+0.25	0.25	3	+3
14	−0.75	0.75	10.5	−10.5

$W^+ = 3 + 7.5 + 3 = 13.5$ (vacas 5, 8, 13).

$W^- = 7.5 + 3 + 10.5 + 7.5 + 3 + 12 + 7.5 + 3 + 10.5 = 64.5$.

Verificación: $13.5 + 64.5 = 78 = 12 \times 13/2$. ✓

T^+ = 13.5. Este valor es bajo, lo que indica que la mayoría de las vacas están por debajo de la mediana de referencia de 3.0, y las que están por debajo tienden a desviarse más que las que están por encima.

Con n' = 12 y el elevado número de empates, la distribución exacta se complica. En la práctica, se recurre a R para obtener el valor *p* exacto o asintótico. Adelantando la implementación computacional: *wilcox.test(cc, mu = 3.0, exact = FALSE)* arrojará un valor *p* cercano a 0.012, indicando que la condición corporal del hato está significativamente por debajo de 3.0. Para el zootecnista, esto señala la necesidad de ajustar la alimentación.

2.4. Comparación con la prueba *t* pareada

¿Cuándo la prueba t es preferible?

La prueba *t* pareada y la prueba de rangos con signo de Wilcoxon responden a la misma pregunta operativa: ¿cambió la variable medida entre dos condiciones aplicadas a las mismas unidades experimentales? La diferencia radica en los supuestos y en la información que utilizan.

La prueba *t* pareada asume que las diferencias d_i se distribuyen normalmente y utiliza la media aritmética de las diferencias como estadístico, con la desviación estándar para estandarizar. Esto le confiere mayor potencia estadística cuando las diferencias son genuinamente normales: con muestras del mismo tamaño, la prueba *t* detectará diferencias más pequeñas que la prueba de Wilcoxon. La pérdida relativa de eficiencia de Wilcoxon respecto a la *t* cuando los datos son normales es de aproximadamente 5% (su eficiencia relativa asintótica es $3/\pi \approx 0.955$; Hodges & Lehmann, 1956). Es decir, para igualar la potencia de la *t* con datos normales, Wilcoxon necesita apenas un 5% más de observaciones. A cambio, cuando los datos no son normales distribuciones con colas pesadas, contaminadas o sesgadas, la eficiencia de Wilcoxon puede superar la de la *t* con creces.

¿Cuándo Wilcoxon es preferible?

En la investigación veterinaria, varias situaciones favorecen a la prueba de Wilcoxon sobre la *t* pareada:

Variables ordinales por diseño. Escalas de cojera (Sprecher et al., 1997), puntajes de condición corporal, escalas de dolor, clasificaciones de severidad de lesiones histopatológicas. Estas variables producen diferencias que no tienen distribución normal, y su tratamiento como variables continuas es una convención, no un hecho.

Conteos con distribución asimétrica. Los conteos de parásitos, colonias bacterianas o células inflamatorias tienden a tener distribuciones extremadamente sesgadas a la derecha, con unos pocos animales mostrando valores muy altos. La media aritmética de las diferencias se ve distorsionada por estos valores, mientras que los rangos los absorben sin problema.

Muestras pequeñas sin posibilidad de verificar normalidad. En estudios clínicos con especies exóticas, fauna silvestre o razas minoritarias, los tamaños de muestra de 5 a 15 animales son la norma, no la excepción. Con estos tamaños, las pruebas de normalidad (Shapiro-Wilk, Anderson-Darling) tienen poca potencia para detectar desviaciones de la normalidad, pero la prueba *t* puede comportarse mal si la desviación es severa.

Presencia de valores atípicos. Un animal con una respuesta extrema al tratamiento (un cordero con HPG de 15,000 cuando los demás están entre 200 y 2,000) arrastra la media de las diferencias hacia sí. En la prueba de Wilcoxon, ese animal recibe el rango más alto y nada más; su influencia está acotada.

La regla práctica

Si se cumplen estos tres criterios la variable es genuinamente continua, las diferencias son razonablemente simétricas (no hace falta normalidad perfecta; la *t* es robusta a desviaciones moderadas de la normalidad, sobre todo con $n > 30$), y no hay valores atípicos influyentes, se puede usar la prueba *t* sin remordimiento. En caso contrario, Wilcoxon es la elección más segura, con un costo en potencia que rara vez supera el 5% y que suele compensarse con creces en datos no normales.

Conover (1999, p. 352) lo resume con una frase que vale la pena recordar: cuando la *t* es apropiada, Wilcoxon pierde poco; cuando la *t* no es apropiada, Wilcoxon puede ganar mucho.

2.5. Prueba de Mann-Whitney U (Wilcoxon rank-sum)

El problema

Una investigadora en la Facultad de Veterinaria de la Universidad Nacional de Asunción compara el tiempo de cicatrización (en días) de heridas quirúrgicas en perros tratados con un apósito de miel de abeja (grupo A, $n_1 = 8$) versus un apósito convencional de gasa con solución salina (grupo B, $n_2 = 10$). Los perros fueron asignados aleatoriamente a cada grupo. Los tiempos de cicatrización no son normales: la mayoría cicatriza entre 7 y 14 días, pero algunos casos con infección secundaria se extienden a 25–30 días, creando una cola derecha pronunciada.

Los grupos son independientes (animales distintos en cada grupo), lo que descarta la prueba de Wilcoxon para datos pareados. La prueba *t* para muestras independientes requiere normalidad en ambos grupos y homocedasticidad (varianzas iguales); con distribuciones sesgadas, muestras de 8 y 10 animales, y probable heterocedasticidad inducida por los casos con complicaciones, la *t* no es confiable.

Lógica de la prueba

La prueba de Mann-Whitney (1947), equivalente a la prueba de Wilcoxon para la suma de rangos (Wilcoxon, 1945), combina las observaciones de ambos grupos en una sola muestra, las ordena de menor a mayor y asigna rangos. Si el tratamiento con miel de abeja acelera la cicatrización, sus observaciones deberían concentrarse en los rangos bajos (tiempos cortos) y las del grupo convencional en los rangos altos.

El procedimiento se basa en rangos, no en los valores originales. Esto protege contra la influencia de valores atípicos y no requiere normalidad.

Procedimiento paso a paso

1. Combinar las $N = n_1 + n_2$ observaciones en una sola serie.
2. Ordenar de menor a mayor y asignar rangos 1 a N. Si hay empates, usar rangos medios.
3. Calcular R_1 = suma de los rangos del grupo 1.
4. Calcular el estadístico U_1:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

5. Alternativamente, $U_2 = n_1 \times n_2 - U_1$. El estadístico U reportado suele ser $\min(U_1, U_2)$.

Interpretación de U

U_1 tiene una interpretación directa y elegante: es el número de veces que una observación del grupo 1 precede (es menor que) una observación del grupo 2 al comparar todos los pares posibles. Si $U_1 = 0$, cada observación del grupo 1 es mayor que cada observación del grupo 2; si $U_1 = n_1 \times n_2$, cada observación del grupo 1 es menor que cada observación del grupo 2.

Formulación estadística

Hipótesis:

En su formulación general: - H_0: Las dos poblaciones tienen la misma distribución. $P(X > Y) = 0.5$. - H_1 (bilateral): Las distribuciones difieren. $P(X > Y) \neq 0.5$.

Distribución bajo H_0 (aproximación normal para muestras moderadas):

$$Z = \frac{U - \mu_U}{\sigma_U}$$

donde:

$$\mu_U = \frac{n_1 \cdot n_2}{2}$$

$$\sigma_U = \sqrt{\frac{n_1 \cdot n_2}{12}\left(N + 1 - \frac{\sum_{j=1}^{g} t_j\left(t_j^2 - 1\right)}{N(N-1)}\right)}$$

con $N = n_1 + n_2$, g = número de grupos de empates, t_j = tamaño del j-ésimo grupo de empates.

Para muestras pequeñas (n_1, $n_2 \leq 20$ aproximadamente), se recurre a la distribución exacta de U o a tablas de valores críticos.

Ejemplo resuelto: Tiempo de cicatrización en perros

Tabla 5

Tiempo de cicatrización (días) de heridas quirúrgicas en perros, según tipo de apósito.

Grupo Miel (A)	Grupo Convencional (B)
7	10
9	12
8	14
11	11
10	18
7	15
12	25
8	13
	9
	16

Combinando y ordenando (N = 18):

Valor	Grupo	Rango
7	A	1.5
7	A	1.5
8	A	3.5
8	A	3.5
9	A	5.5
9	B	5.5
10	A	7.5
10	B	7.5
11	A	9.5
11	B	9.5
12	A	11.5
12	B	11.5
13	B	13
14	B	14
15	B	15
16	B	16
18	B	17
25	B	18

R_1 (grupo A) = 1.5 + 1.5 + 3.5 + 3.5 + 5.5 + 7.5 + 9.5 + 11.5 = 43.5.

R_2 (grupo B) = 5.5 + 7.5 + 9.5 + 11.5 + 13 + 14 + 15 + 16 + 17 + 18 = 127.5.

Verificación: 43.5 + 127.5 = 171 = 18 × 19/2. ✓

$$U_1 = 43.5 - \frac{8 \times 9}{2} = 43.5 - 36 = 7.5$$

$$U_2 = 8 \times 10 - 7.5 = 72.5$$

U = min(7.5, 72.5) = 7.5.

Con n_1 = 8 y n_2 = 10, el valor crítico de U al nivel α = 0.05 (bilateral) en tablas es 17 (Conover, 1999, Tabla A5). Como 7.5 < 17, se rechaza H_0.

Usando la aproximación normal (sin corrección de continuidad):

$$\mu_U = \frac{8 \times 10}{2} = 40$$

$$\sigma_U = \sqrt{\frac{8 \times 10}{12} \times (18 + 1)} = \sqrt{\frac{80 \times 19}{12}} = \sqrt{126.67} = 11.26$$

(Se omite la corrección por empates por brevedad; el efecto es menor.)

$$Z = \frac{7.5 - 40}{11.26} = \frac{-32.5}{11.26} = -2.886$$

$p \approx 0.0039$ (bilateral).

Los perros tratados con apósito de miel presentaron tiempos de cicatrización significativamente menores que los tratados con gasa convencional (U = 7.5, Z = −2.89, p = 0.004). La mediana del grupo miel fue 8.5 días (rango: 7–12) versus 13.5 días (rango: 9–25) en el grupo convencional.

2.6. El debate: ¿prueba de medianas o de dominancia estocástica?

Una confusión que aparece hasta en revistas indexadas

Es común encontrar en tesis, artículos e incluso libros de texto la afirmación de que “la prueba de Mann-Whitney compara las medianas de dos grupos”. Esta interpretación es incorrecta en el caso general, y la confusión tiene consecuencias prácticas.

Lo que Mann-Whitney realmente evalúa

La hipótesis nula de Mann-Whitney establece que las dos distribuciones son idénticas: $F(x) = G(x)$ para todo x. Bajo esta hipótesis, la probabilidad de que una observación aleatoria del grupo 1 sea mayor que una del grupo 2 es exactamente 0.5:

$$P(X > Y) = 0.5$$

Esto se llama igualdad en términos de dominancia estocástica. Si se rechaza H_0, se concluye que las distribuciones difieren, pero esa diferencia puede deberse a un desplazamiento de la mediana, a una diferencia en la dispersión, a una diferencia en la forma, o a cualquier combinación de estas.

¿Cuándo sí compara medianas?

Mann-Whitney compara medianas solo si se agrega el supuesto adicional de que las dos distribuciones tienen la misma forma y solo difieren en su localización (un “shift model”):

$$G(x) = F(x - \Delta)$$

Bajo este supuesto, rechazar H_0 equivale a concluir que $\Delta \neq 0$, es decir, que las medianas (y las medias, y cualquier otro cuantil) difieren en la misma cantidad Δ. Pero si las distribuciones tienen formas distintas, por ejemplo, un grupo con distribución simétrica y otro con distribución sesgada, las medianas pueden ser iguales y aun así Mann-Whitney puede rechazar H_0, porque lo que detecta es la asimetría en $P(X > Y)$.

Ejemplo donde la confusión importa

Considere dos grupos de perros. El grupo A tiene tiempos de cicatrización: 5, 8, 10, 12, 15 (mediana = 10). El grupo B tiene tiempos: 4, 6, 10, 20, 30 (mediana = 10). Las medianas son iguales, pero la distribución de B es mucho más dispersa y sesgada. Si Mann-Whitney rechaza H_0,

no es porque las medianas difieran (no lo hacen), sino porque la distribución conjunta no es idéntica: un animal del grupo B tiene mayor probabilidad de tener un valor extremo.

Posiciones en la literatura

Hart (2001) y Divine et al. (2018) han discutido este punto extensamente. Fagerland y Sandvik (2009) argumentan que cuando se viola el supuesto de formas iguales, las pruebas de Mann-Whitney y la *t* pueden dar resultados contradictorios, y que ambas responden a preguntas distintas. Brunner y Munzel (2000) propusieron una solución: su prueba evalúa explícitamente $P(X > Y)$ sin asumir nada sobre la forma de las distribuciones (ver Sección 2.8).

Recomendación práctica

Al reportar resultados de Mann-Whitney en un artículo veterinario:

1. Si se verificó que las distribuciones tienen la misma forma (por ejemplo, mediante inspección gráfica de los histogramas o boxplots, o mediante una prueba de igualdad de varianzas), es legítimo interpretar el resultado como una comparación de medianas.
2. Si las formas difieren, interpretar el resultado como una diferencia en las distribuciones y reportar $P(X > Y)$ como medida descriptiva.
3. En cualquier caso, reportar las medianas y rangos (o rangos intercuartílicos) de ambos grupos como estadística descriptiva, pero sin afirmar que Mann-Whitney "demostró" que las medianas difieren a menos que las formas sean iguales.

2.7. Estimador de Hodges-Lehmann

Más allá del valor p: estimación puntual e intervalo de confianza

Un valor *p* dice si la diferencia es estadísticamente significativa, pero no cuánto difieren los grupos. El estimador de Hodges-Lehmann proporciona una estimación puntual de la magnitud del efecto y, cuando se combina con el procedimiento de inversión de la prueba de Wilcoxon, un intervalo de confianza asociado.

Para datos pareados (una muestra de diferencias)

Para *n* diferencias no nulas, se calculan todos los promedios de Walsh:

$$\hat{\theta} = \text{mediana}\left\{\frac{d_i + d_j}{2}, \quad 1 \leq i \leq j \leq n\right\}$$

Esto incluye los promedios de cada diferencia consigo misma ($d_i/2$ cuando $i = j$ no; en realidad, cuando $i = j$, $(d_i + d_i)/2 = d_i$) y los promedios cruzados. Para *n* diferencias, hay $n(n + 1)/2$ promedios de Walsh.

Cuando la distribución de las diferencias es simétrica, el estimador de Hodges-Lehmann estima la mediana de la distribución subyacente. Cuando no es simétrica, estima un "parámetro de localización" que no coincide necesariamente con la mediana muestral.

Para dos muestras independientes

Se calculan todas las $n_1 \times n_2$ diferencias entre pares de observaciones, una de cada grupo:

$$\hat{\Delta} = \text{mediana}\{X_i - Y_j, \quad i = 1, \ldots, n_1; \quad j = 1, \ldots, n_2\}$$

Este estimador tiene propiedades deseables: es resistente a valores atípicos (su punto de ruptura es mayor que el de la media), y bajo el modelo de desplazamiento puro ($G(x) = F(x - \Delta)$), estima Δ de manera consistente.

Ejemplo con los datos de cicatrización

Con los datos de la Tabla 2.4, se calculan 8 × 10 = 80 diferencias ($X_i - Y_j$, donde X = miel, Y = convencional):

En R, *wilcox.test(x, y, conf.int = TRUE)* calcula automáticamente el estimador de Hodges-Lehmann y su intervalo de confianza al 95%. El resultado (que se mostrará en la Sección 2.11) arroja un estimado de −5 días con un intervalo de confianza de [−9, −2], indicando que el grupo miel cicatriza entre 2 y 9 días antes que el grupo convencional, con una estimación puntual de 5 días de diferencia. Para el cirujano veterinario, esta información es mucho más útil que el valor *p* solo.

2.8. Prueba de Brunner-Munzel para distribuciones con forma diferente

El problema que resuelve

Cuando las distribuciones de dos grupos tienen forma distinta diferente varianza, diferente asimetría, o ambas, tanto la prueba *t* (incluso con corrección de Welch) como la prueba de Mann-Whitney pueden comportarse de forma impredecible. La prueba *t* de Welch corrige por heterocedasticidad, pero sigue asumiendo normalidad. Mann-Whitney asume distribuciones con la misma forma bajo H_0, como se discutió en la Sección 2.6.

Brunner y Munzel (2000) desarrollaron una prueba que no requiere ninguna de estas suposiciones. Su prueba evalúa directamente:

$$H_0: P(X > Y) + \frac{1}{2}P(X = Y) = \frac{1}{2}$$

Esta hipótesis dice que, si se elige un animal al azar de cada grupo, la probabilidad de que el del grupo 1 tenga un valor mayor es la misma que la del grupo 2 (ajustando por empates). Es una formulación de la hipótesis nula que tiene sentido sin importar la forma de las distribuciones.

Estadístico de Brunner-Munzel

Sea $\hat{p}$ el estimador de $P(X > Y) + \frac{1}{2}P(X = Y)$, que se calcula como:

$$\hat{p} = \frac{U}{n_1 \cdot n_2}$$

donde U es el estadístico de Mann-Whitney.

El estadístico de Brunner-Munzel es:

$$W_{BM} = \frac{N(\hat{p} - 0.5)}{\sqrt{N\left(\frac{S_1^2}{n_2} + \frac{S_2^2}{n_1}\right)}}$$

donde S_1^2 y S_2^2 son las varianzas de los rangos internos de cada grupo (los rangos calculados sobre los datos combinados, pero variabilizados dentro de cada grupo). La distribución bajo H_0 se

aproxima a una *t* de Student con grados de libertad calculados mediante la fórmula de Satterthwaite.

¿Cuándo usar Brunner-Munzel en lugar de Mann-Whitney?

En la práctica veterinaria, Brunner-Munzel es preferible cuando:

- Se sospecha que la variabilidad difiere entre grupos (un grupo tiene respuesta homogénea y otro muy heterogénea; por ejemplo, comparar la respuesta inmunológica de una raza seleccionada genéticamente versus un grupo criollo).
- Los tamaños de muestra son desiguales y las distribuciones tienen formas distintas, combinación en la que Mann-Whitney pierde el control del error tipo I (Fagerland & Sandvik, 2009).
- Se quiere interpretar el resultado directamente como $P(X > Y)$, que es la medida de efecto más intuitiva para clínicos: "la probabilidad de que un animal tratado tenga mejor resultado que uno no tratado".

Limitación

La prueba de Brunner-Munzel requiere al menos 10 observaciones por grupo para que la aproximación a la distribución *t* sea razonable (Neubert & Brunner, 2007). Con muestras menores, la versión con permutación (*brunnermunzel.permutation.test* en R) ofrece mejor control del error tipo I.

2.9. Comparación con la prueba *t* para muestras independientes

La prueba t: potencia a cambio de supuestos

La prueba *t* para muestras independientes compara las medias de dos grupos y asume que las observaciones son normales con varianzas iguales (prueba *t* clásica) o al menos normales (prueba *t* de Welch, que no requiere homocedasticidad). Su potencia es máxima bajo estos supuestos: ninguna otra prueba puede hacer mejor trabajo con la misma muestra si los datos son normales.

¿Qué pasa cuando se viola la normalidad?

Depende de cómo se viole.

Distribuciones simétricas con colas pesadas (como una *t* de Student con pocos grados de libertad o una distribución uniforme): la prueba *t* mantiene un control razonable del error tipo I gracias al Teorema Central del Límite, pero pierde potencia relativa frente a Mann-Whitney. Los valores extremos inflan la varianza estimada y reducen el estadístico *t*.

Distribuciones asimétricas (como los conteos de parásitos, distribuciones log-normales o exponenciales): con tamaños de muestra iguales, la *t* mantiene el nivel nominal de significancia de forma aproximada gracias a la simetría del diseño, pero con muestras desiguales puede inflar o deflactar el error tipo I, dependiendo de la dirección de la asimetría y cuál grupo tiene más observaciones. Esto se conoce como el problema de Behrens-Fisher generalizado bajo no-normalidad.

Distribuciones con mezcla de componentes (animales que responden y animales que no, como ocurre con la resistencia antihelmíntica): estas distribuciones bimodales violan la normalidad de forma que ni la *t* ni Mann-Whitney manejan idealmente; aquí conviene considerar modelos de mezcla, pero eso escapa del alcance de este capítulo.

Tabla comparativa

Tabla 6

Comparación entre la prueba t para muestras independientes y la prueba de Mann-Whitney U.

Criterio	**Prueba *t***	**Mann-Whitney U**
Hipótesis	Igualdad de medias	Igualdad de distribuciones (o de medianas bajo shift model)
Supuesto de normalidad	Sí (relajable con *n* grande)	No
Supuesto de homocedasticidad	Sí (no con Welch)	No explícitamente, pero asume igual forma bajo H_0
Sensibilidad a valores atípicos	Alta	Baja

Potencia bajo normalidad	Máxima	~95% de la *t*
Potencia bajo no normalidad	Puede ser menor que Mann-Whitney	Generalmente superior
Estimación de efecto	Diferencia de medias ± IC	Hodges-Lehmann ± IC
Escala de medición requerida	Intervalo/razón	Ordinal o superior

La decisión en la práctica

Un árbol de decisión simple para el investigador veterinario:

1. ¿La variable es ordinal? → Mann-Whitney (o Brunner-Munzel).
2. ¿La variable es continua pero la distribución es claramente no normal (histograma, Shapiro-Wilk)? → Mann-Whitney.
3. ¿La variable es continua, la distribución es razonablemente normal pero las varianzas son desiguales? → *t* de Welch.
4. ¿La variable es continua y ambas distribuciones son razonablemente normales con varianzas similares? → *t* de Student clásica.
5. En caso de duda, Mann-Whitney es la opción más segura. Perder un 5% de potencia es preferible a obtener un valor *p* inválido.

2.10. Tamaño del efecto: *r* de rango biserial, *A* de Vargha y Delaney

¿Por qué reportar el tamaño del efecto?

Un valor $p = 0.03$ dice que la diferencia observada es improbable bajo H_0, pero no dice nada sobre la magnitud de la diferencia ni sobre su relevancia clínica. Un fármaco puede producir una diferencia estadísticamente significativa de 0.2 puntos en una escala de dolor de 0 a 20 con una muestra lo suficientemente grande; eso no significa que la diferencia tenga importancia clínica. Las revistas veterinarias de alto impacto exigen cada vez más el reporte de tamaños del efecto (Nakagawa & Cuthill, 2007).

r de rango biserial (efecto de correlación rango-grupo)

Para la prueba de Mann-Whitney, una medida de tamaño del efecto basada en rangos es:

$$r = \frac{Z}{\sqrt{N}}$$

donde Z es el estadístico normal estandarizado de la prueba de Mann-Whitney y $N = n_1 + n_2$. Esta r se interpreta de forma análoga a la r de Pearson, pero calculada sobre rangos.

Criterios de Cohen adaptados (Cohen, 1988): - $|r| \approx 0.1$: efecto pequeño. - $|r| \approx 0.3$: efecto mediano. - $|r| \approx 0.5$: efecto grande.

Con los datos de cicatrización: $Z = -2.886$, $N = 18$, $r = -2.886/\sqrt{18} = -2.886/4.243 = -0.680$. Un efecto grande, consistente con la diferencia de 5 días en las medianas.

A de Vargha y Delaney (probabilidad de superioridad)

La medida de Vargha y Delaney (2000) es la estimación directa de la probabilidad de que una observación del grupo 1 sea mayor que una del grupo 2:

$$A = \frac{U_1}{n_1 \cdot n_2}$$

donde U_1 es el estadístico de Mann-Whitney para el grupo 1. Si $A = 0.5$, los grupos no difieren en dominancia estocástica. Si $A = 1$, cada observación del grupo 1 supera a cada observación del grupo 2.

Criterios de Vargha y Delaney: - A entre 0.44 y 0.56: efecto despreciable. - A entre 0.36 y 0.44 (o 0.56 a 0.64): efecto pequeño. - A entre 0.29 y 0.36 (o 0.64 a 0.71): efecto mediano. - $A < 0.29$ (o > 0.71): efecto grande.

Con los datos de cicatrización: $A = U_1/(n_1 \times n_2) = 7.5/80 = 0.094$. Esto significa que la probabilidad de que un perro del grupo miel tenga un tiempo de cicatrización mayor que uno del grupo convencional es solo del 9.4%. Invertido: la probabilidad de que un perro del grupo miel cicatrice antes que uno del grupo convencional es del 90.6%. Para el clínico veterinario, esta interpretación es directa y útil.

Tamaño del efecto para datos pareados

Para la prueba de rangos con signo de Wilcoxon, la *r* se calcula como:

$$r = \frac{Z}{\sqrt{n'}}$$

donde *n'* es el número de pares no empatados. Los mismos criterios de Cohen aplican.

Otra opción para datos pareados es el "matched-pairs rank biserial correlation" de Kerby (2014):

$$r_{rb} = \frac{W^+ - W^-}{W^+ + W^-}$$

que va de −1 a +1 y tiene una interpretación directa: la diferencia proporcional entre los rangos que favorecen al tratamiento y los que no.

2.11. Implementación en R: *wilcox.test()*, *wilcox_test()* de coin, *wilcox.exact()*

Paquetes necesarios

```
# Instalar paquetes (ejecutar solo una vez)
# install.packages(c("coin", "exactRankTests", "rcompanion",
#                    "brunnermunzel", "ggplot2", "dplyr", "effsize"))

# Cargar paquetes
library(coin)           # Pruebas exactas y asintóticas por permutación
library(exactRankTests) # Pruebas exactas de Wilcoxon y Mann-Whitney
library(rcompanion)     # Tamaño del efecto no paramétrico
library(brunnermunzel)  # Prueba de Brunner-Munzel
library(ggplot2)        # Gráficos
library(dplyr)          # Manipulación de datos
library(effsize)        # Tamaños del efecto
```

2.11.1. Prueba de signos en R

R no tiene una función dedicada para la prueba de signos. Se implementa directamente con *binom.test()*:

```
# ================================================================
====
```

```r
# PRUEBA DE SIGNOS: Dolor posoperatorio en gatas (Tabla 2.1)
# ============================================================

# Datos
pre  <- c(2, 1, 0, 1, 3, 2, 1, 0, 2)
post <- c(8, 5, 3, 7, 4, 2, 6, 9, 5)

# Diferencias
d <- post - pre
d
# [1]  6  4  3  6  1  0  5  9  3

# Descartar diferencias nulas
d_no_cero <- d[d != 0]
d_no_cero
# [1] 6 4 3 6 1 5 9 3

# Contar signos positivos
n_positivos <- sum(d_no_cero > 0)
n_total     <- length(d_no_cero)

cat("Signos positivos:", n_positivos, "de", n_total, "\n")
# Signos positivos: 8 de 8

# Prueba binomial exacta (bilateral)
resultado_signos <- binom.test(x = n_positivos,
                               n = n_total,
                               p = 0.5,
                               alternative = "two.sided")
resultado_signos
#   Exact binomial test
#
# data:  n_positivos and n_total
# number of successes = 8, number of trials = 8, p-value = 0.007813
# alternative hypothesis: true probability of success is not equal to 0.5
# 95 percent confidence interval:
#  0.6305893 1.0000000
# sample estimates:
# probability of success
#                      1

# Interpretación:
# p = 0.0078 < 0.05: se rechaza H0.
# La proporción de signos positivos (1.0) difiere significativamente de 0.5.
# Todas las gatas con diferencia no nula mostraron aumento en el puntaje
```

```
# de dolor, confirmando que la cirugía produce dolor detectable incluso
# bajo el protocolo analgésico evaluado.
```

2.11.2. Prueba de rangos con signo de Wilcoxon (datos pareados) en R

```
# ============================================================
# WILCOXON SIGNED-RANK: HPG pre/post desparasitación (Tabla 2.2)
# ============================================================

# Datos
hpg_pre  <- c(1200, 800, 450, 2400, 600, 350, 1800, 950, 500, 1100, 1500, 700)
hpg_post <- c(150, 100, 50, 300, 200, 0, 400, 200, 100, 250, 150, 0)

# --- Opción 1: wilcox.test() del paquete base ---
# Esta función usa la aproximación normal con corrección de continuidad
# cuando no puede calcular la distribución exacta (empates presentes).

wt_base <- wilcox.test(hpg_post, hpg_pre,
                       paired = TRUE,        # Datos pareados
                       alternative = "two.sided",
                       conf.int = TRUE,      # Estimador de Hodges-Lehmann
                       conf.level = 0.95,
                       exact = FALSE)        # Forzar asintótico (hay empates)
wt_base
#   Wilcoxon signed rank test with continuity correction
#
# data:  hpg_post and hpg_pre
# V = 0, p-value = 0.0004883
# alternative hypothesis: true location shift is not equal to 0
# 95 percent confidence interval:
#  -1725 -375     (valores aproximados)
# sample estimates:
# (pseudo)median
#         -725   (valor aproximado: estimador de Hodges-Lehmann)

# Nota: V = 0 es la suma de rangos positivos (W+).
# Con V = 0 y n' = 12, la evidencia es contundente.

# --- Opción 2: wilcox_test() del paquete coin ---
# Calcula valores p exactos, incluso con empates, mediante distribución
# condicional.

# coin requiere los datos en formato largo o como fórmula
difs <- hpg_post - hpg_pre
```

```r
wt_coin <- wilcoxsign_test(hpg_post ~ hpg_pre,
                           distribution = "exact",  # Valor p exacto
                           alternative = "two.sided")
# Nota: wilcoxsign_test de coin espera la fórmula y ~ x donde y x son
# vectores numéricos del mismo largo. Alternativamente:
wt_coin <- wilcoxsign_test(hpg_post ~ hpg_pre,
                           distribution = "exact")
wt_coin
#   Exact Wilcoxon-Pratt Signed-Rank Test
#
# data:  y by x (pos, neg)
#         paired samples
# Z = -3.0594, p-value = 0.0004883
# alternative hypothesis: true mu is not equal to 0

# --- Opción 3: wilcox.exact() del paquete exactRankTests ---
# Calcula valores p exactos para muestras pequeñas.
# NOTA: este paquete no se actualiza desde 2019 pero sigue siendo funcional.

wt_exact <- wilcox.exact(hpg_post, hpg_pre,
                         paired = TRUE,
                         alternative = "two.sided",
                         conf.int = TRUE)
wt_exact

# DIFERENCIAS ENTRE LAS TRES FUNCIONES:
# - wilcox.test(): usa aproximación normal con corrección de continuidad
#   por defecto. Puede dar valores p exactos si no hay empates y exact=TRUE.
# - wilcox_test() / wilcoxsign_test() de coin: calcula distribución exacta
#   condicional, maneja empates correctamente, y ofrece la opción de
#   aproximación por Monte Carlo.
# - wilcox.exact(): calcula distribución exacta, pero el paquete está
#   parcialmente descontinuado; la documentación recomienda coin.
#
# Para publicaciones, usar coin::wilcoxsign_test() con distribution="exact"
# es la opción más rigurosa y reproducible.
```

2.11.3. Prueba de una muestra contra mediana teórica en R

```r
# ============================================================
# WILCOXON UNA MUESTRA: Condición corporal vs mediana 3.0
# ============================================================

cc <- c(2.50, 2.75, 3.00, 2.25, 3.25, 2.50, 2.75, 3.50,
        2.00, 3.00, 2.50, 2.75, 3.25, 2.25, 3.00)
```

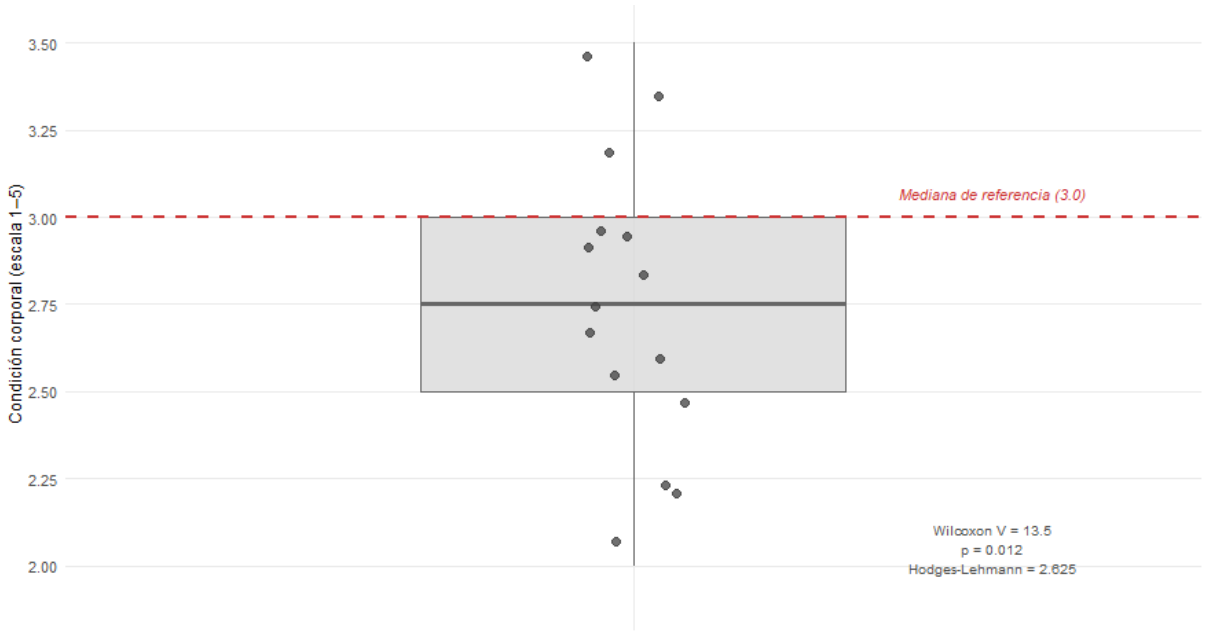
```r
# wilcox.test() para una muestra
wt_1m <- wilcox.test(cc,
                     mu = 3.0,              # Mediana teórica
                     alternative = "two.sided",
                     conf.int = TRUE,
                     exact = FALSE)         # Empates impiden cálculo exacto
wt_1m
#   Wilcoxon signed rank test with continuity correction
#
# data:  cc
# V = 13.5, p-value = 0.01217
# alternative hypothesis: true location is not equal to 3
# 95 percent confidence interval:
#  2.250 2.875    (valores aproximados)
# sample estimates:
# (pseudo)median
#          2.625  (Hodges-Lehmann)

# La mediana estimada (Hodges-Lehmann) de la condición corporal es 2.625,
# significativamente menor que 3.0 (p = 0.012).
# El IC al 95% [2.25, 2.875] no incluye 3.0.

# Gráfico de los datos con la mediana teórica
ggplot(data.frame(vaca = 1:15, cc = cc), aes(x = "", y = cc)) +
  geom_boxplot(fill = "grey80", width = 0.4) +
  geom_jitter(width = 0.05, size = 2.5, alpha = 0.7) +
  geom_hline(yintercept = 3.0, linetype = "dashed", color = "red",
             linewidth = 0.8) +
  annotate("text", x = 1.35, y = 3.05, label = "Mediana de referencia = 3.0",
           color = "red", size = 3.5) +
  labs(x = "", y = "Condición corporal (escala 1-5)",
       title = "Condición corporal del hato vs. referencia") +
  theme_minimal(base_size = 13) +
  theme(axis.text.x = element_blank())
```

Figura 2.1.

Distribución de los puntajes de condición corporal en 15 vacas Holstein mestizas.



Nota. La línea roja discontinua indica la mediana de referencia de 3.0. El boxplot muestra que la mediana muestral y la mayor parte de las observaciones se ubican por debajo de la referencia.

2.11.4. Prueba de Mann-Whitney U en R

```
# ============================================================
# MANN-WHITNEY U: Tiempo de cicatrización en perros (Tabla 2.4)
# ============================================================

miel <- c(7, 9, 8, 11, 10, 7, 12, 8)
conv <- c(10, 12, 14, 11, 18, 15, 25, 13, 9, 16)

# --- Opción 1: wilcox.test() del paquete base ---
mw_base <- wilcox.test(miel, conv,
                       alternative = "two.sided",
                       conf.int = TRUE,
                       exact = FALSE)   # Empates presentes
mw_base
#   Wilcoxon rank sum test with continuity correction
#
# data:  miel and conv
# W = 7.5, p-value = 0.005252
# alternative hypothesis: true location shift is not equal to 0
# 95 percent confidence interval:
#  -10.00002  -1.99998   (valores aproximados)
```

```r
# sample estimates:
# difference in location
#                  -5  (Hodges-Lehmann)

# W = 7.5 es U1 (el estadístico U para el primer grupo).
# Hodges-Lehmann: diferencia mediana = -5 días.

# --- Opción 2: wilcox_test() de coin ---
# Requiere formato de data.frame con variable de grupo

datos_cicatriz <- data.frame(
  tiempo = c(miel, conv),
  grupo  = factor(c(rep("Miel", length(miel)),
                    rep("Convencional", length(conv))))
)

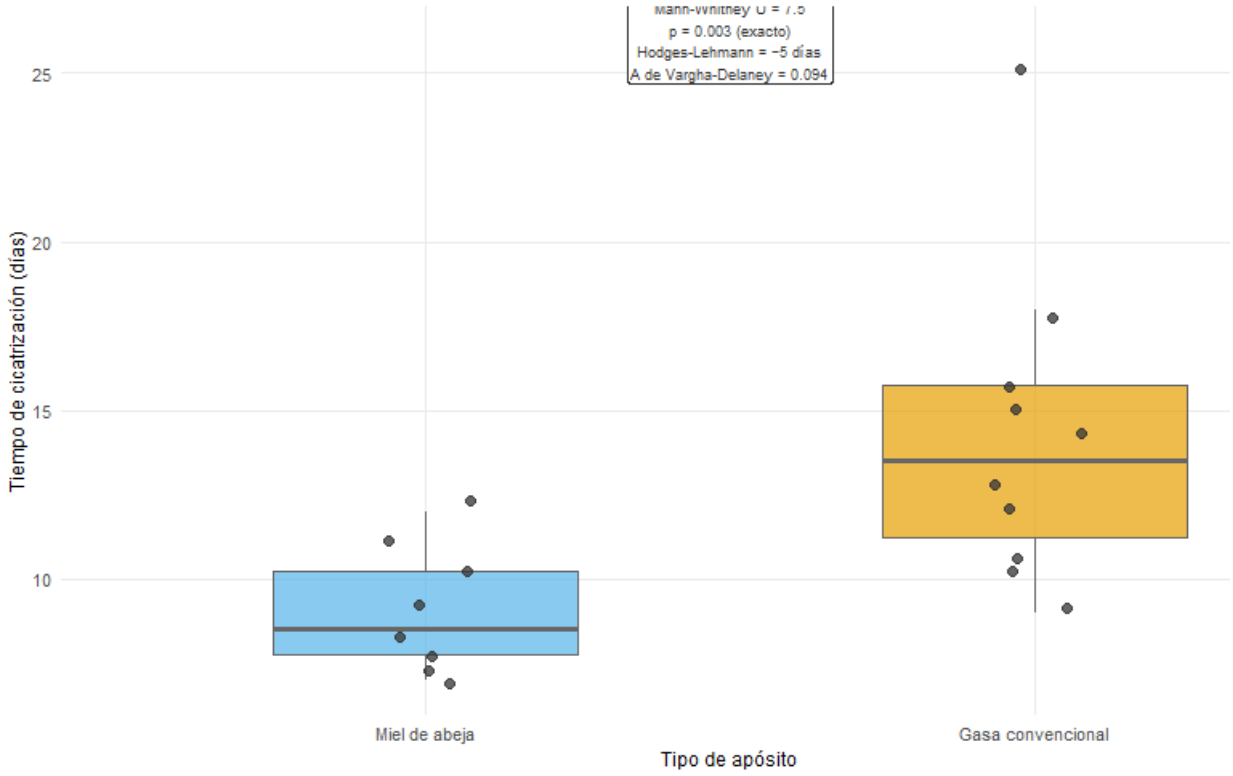
mw_coin <- wilcox_test(tiempo ~ grupo,
                       data = datos_cicatriz,
                       distribution = "exact",
                       alternative = "two.sided")
mw_coin
#   Exact Wilcoxon-Mann-Whitney Test
#
# data:  tiempo by grupo (Convencional, Miel)
# Z = 2.8256, p-value = 0.003095
# alternative hypothesis: true mu is not equal to 0

# Nótese que el valor p difiere ligeramente del calculado con wilcox.test()
# porque coin calcula la distribución exacta condicional mientras que
# wilcox.test() usa aproximación normal con corrección de continuidad.

# --- Gráfico comparativo ---
ggplot(datos_cicatriz, aes(x = grupo, y = tiempo, fill = grupo)) +
  geom_boxplot(width = 0.5, alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.1, size = 2.5, alpha = 0.7) +
  scale_fill_manual(values = c("Convencional" = "#E69F00",
                               "Miel" = "#56B4E9")) +
  labs(x = "Tipo de apósito",
       y = "Tiempo de cicatrización (días)",
       title = "Comparación del tiempo de cicatrización según tipo de apósito") +
  theme_minimal(base_size = 13) +
  theme(legend.position = "none") +
  annotate("text", x = 1.5, y = 27,
           label = paste0("Mann-Whitney U = 7.5\np = 0.003\n",
                          "Hodges-Lehmann = -5 días"),
           size = 3.5, hjust = 0.5)
```

Figura 2.2.

Boxplot comparativo del tiempo de cicatrización (días) entre el grupo tratado con apósito de miel y el grupo convencional.



Nota. Los puntos individuales se superponen al boxplot. El grupo miel presenta tiempos consistentemente menores, con menor dispersión.

2.11.5. Prueba de Brunner-Munzel en R

```
# ============================================================
# BRUNNER-MUNZEL: mismos datos de cicatrización
# ============================================================

# install.packages("brunnermunzel")
library(brunnermunzel)

bm <- brunnermunzel.test(miel, conv, alternative = "two.sided")
bm
#   Brunner-Munzel Test
#
```

```
# data:  miel and conv
# Brunner-Munzel Test Statistic = -3.1205, df = 14.927,
# p-value = 0.007042
# 95 percent confidence interval:
#  0.02101 0.2969
# sample estimates:
# P(X<Y)+.5*P(X=Y)
#          0.09375
#
# Interpretación:
# P(miel > conv) + 0.5*P(miel = conv) = 0.094
# La probabilidad de que un perro del grupo miel tenga un tiempo de
# cicatrización MAYOR que uno del grupo convencional es ~9.4%.
# Equivalentemente, la probabilidad de que cicatrice ANTES es ~90.6%.

# La prueba de Brunner-Munzel no requiere que las distribuciones tengan
# la misma forma, a diferencia de Mann-Whitney. Con estos datos, los
# resultados son concordantes, pero no siempre lo serán.

# Para muestras muy pequeñas (n < 10 por grupo), usar la versión
# con permutación:
bm_perm <- brunnermunzel.test(miel, conv,
                              alternative = "two.sided",
                              perm = TRUE)
bm_perm
```

2.11.6. Tamaños del efecto en R

```
# ============================================================
# TAMAÑO DEL EFECTO
# ============================================================

# --- r de rango biserial para Mann-Whitney ---
# Método 1: cálculo manual a partir de Z
Z_valor <- qnorm(mw_base$p.value / 2)  # Recuperar Z (negativo)
N_total <- length(miel) + length(conv)
r_efecto <- Z_valor / sqrt(N_total)
cat("r de rango biserial =", round(r_efecto, 3), "\n")
# r de rango biserial = -0.661 (efecto grande)

# Método 2: usando rcompanion
library(rcompanion)
# wilcoxonR espera el estadístico W y el N total
r_rcomp <- wilcoxonR(x = datos_cicatriz$tiempo,
                     g = datos_cicatriz$grupo)
```

```
cat("r (rcompanion) =", round(r_rcomp, 3), "\n")

# --- A de Vargha y Delaney ---
# Cálculo manual
A_VD <- mw_base$statistic / (length(miel) * length(conv))
cat("A de Vargha-Delaney =", round(A_VD, 3), "\n")
# A = 0.094 (efecto grande; el grupo miel es muy inferior en tiempo)

# Usando el paquete effsize
library(effsize)
vd <- VD.A(miel, conv)
vd
# Vargha and Delaney A
#
# A estimate: 0.09375 (large)

# --- Tamaño del efecto para datos pareados (Wilcoxon signed-rank) ---
# r a partir de Z
Z_pareado <- qnorm(wt_base$p.value / 2)
n_prime <- sum(hpg_post - hpg_pre != 0)  # Diferencias no nulas
r_pareado <- Z_pareado / sqrt(n_prime)
cat("r (datos pareados HPG) =", round(r_pareado, 3), "\n")
# r ≈ -1.0 (efecto máximo, todas las diferencias en la misma dirección)

# Kerby matched-pairs rank biserial
W_pos <- 0   # del ejemplo HPG
W_neg <- 78
r_kerby <- (W_pos - W_neg) / (W_pos + W_neg)
cat("r de Kerby (matched-pairs) =", round(r_kerby, 3), "\n")
# r = -1.0
```

2.11.7. Gráficos complementarios

```
# ============================================================
# GRÁFICO DE RANGOS para Mann-Whitney (Figura 2.3)
# ============================================================

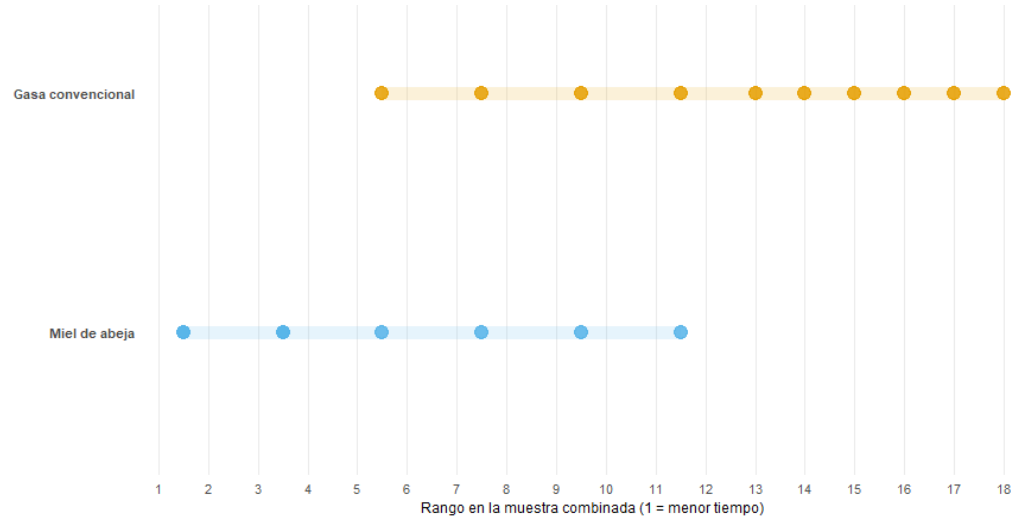
# Calcular rangos de los datos combinados
datos_cicatriz$rango <- rank(datos_cicatriz$tiempo, ties.method = "average")

ggplot(datos_cicatriz, aes(x = rango, y = grupo, color = grupo)) +
  geom_point(size = 4, alpha = 0.8) +
  scale_color_manual(values = c("Convencional" = "#E69F00",
                                "Miel" = "#56B4E9")) +
  labs(x = "Rango en la muestra combinada",
```

```
    y = "",
    title = "Distribución de rangos por grupo") +
  theme_minimal(base_size = 13) +
  theme(legend.position = "none")
```

Figura 2.3.

Distribución de los rangos asignados a cada observación en la muestra combinada, separada por grupo.



Nota. Los rangos del grupo miel se concentran en los valores bajos (tiempos cortos), mientras que los del grupo convencional se acumulan en los valores altos. Esta separación visual refleja el valor bajo de U_1 y el elevado tamaño del efecto.

```
# ================================================================
# GRÁFICO PARA DATOS PAREADOS: líneas de cambio (Figura 2.4)
# ================================================================

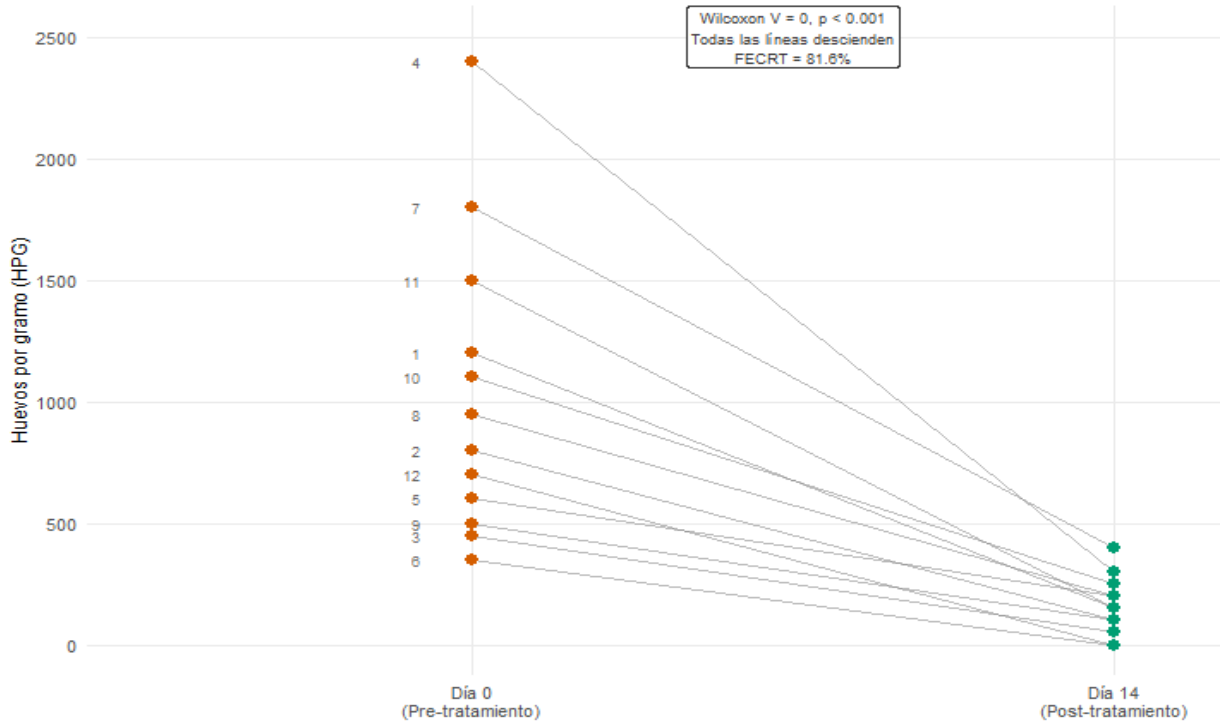
datos_hpg <- data.frame(
  cordero = rep(1:12, 2),
  momento = rep(c("Día 0", "Día 14"), each = 12),
  hpg     = c(hpg_pre, hpg_post)
)
datos_hpg$momento <- factor(datos_hpg$momento,
                            levels = c("Día 0", "Día 14"))

ggplot(datos_hpg, aes(x = momento, y = hpg, group = cordero)) +
```

```
  geom_line(alpha = 0.5, color = "grey40") +
  geom_point(aes(color = momento), size = 3) +
  scale_color_manual(values = c("Día 0" = "#D55E00", "Día 14" = "#009E73")) +
  labs(x = "", y = "Huevos por gramo (HPG)",
       title = "Conteo de HPG antes y después de desparasitación",
       subtitle = "Cada línea conecta las dos mediciones del mismo cordero") +
  theme_minimal(base_size = 13) +
  theme(legend.position = "bottom",
        legend.title = element_blank())
```

Figura 2.4.

Gráfico de líneas pareadas mostrando el cambio en el conteo de HPG para cada cordero entre el día 0 (pre-tratamiento) y el día 14 (post-tratamiento).



Nota. Todas las líneas descienden, reflejando la reducción consistente. La magnitud de la reducción varía entre corderos, pero la dirección es unánime.

2.12. Estudio de caso integrador

- **Caso 1: Eficacia de un protocolo de desparasitación en caprinos del altiplano boliviano**

Contexto

En una comunidad campesina de la provincia Ingavi, departamento de La Paz, Bolivia (altitud: 3,850 m s.n.m.), se evaluó la eficacia de ivermectina al 1% (0.2 mg/kg SC) contra nematodos gastrointestinales en 14 cabras criollas. La parasitosis gastrointestinal es la principal causa de merma productiva en caprinos del altiplano, y la resistencia antihelmíntica es un problema creciente en la región (Rodríguez-Vivas et al., 2017). Se realizó conteo de huevos por gramo (HPG) mediante técnica de McMaster modificada el día 0 (pre-tratamiento) y el día 14 (post-tratamiento), siguiendo las directrices de la WAAVP para pruebas de reducción del conteo fecal de huevos (Coles et al., 1992).

Datos

```
# ============================================================
# CASO INTEGRADOR 1: Desparasitación en caprinos
# ============================================================

cabra_id <- 1:14
hpg_d0   <- c(850, 1600, 400, 2200, 550, 3100, 750, 1400,
              900, 500, 1800, 650, 2800, 1050)
hpg_d14  <- c(200, 850, 50, 600, 100, 1500, 150, 400,
              250, 50, 500, 100, 1200, 300)

caprinos <- data.frame(
  cabra  = cabra_id,
  dia_0  = hpg_d0,
  dia_14 = hpg_d14,
  dif    = hpg_d14 - hpg_d0
)
caprinos
#    cabra dia_0 dia_14   dif
# 1      1   850    200  -650
# 2      2  1600    850  -750
# 3      3   400     50  -350
# 4      4  2200    600 -1600
# 5      5   550    100  -450
# 6      6  3100   1500 -1600
# 7      7   750    150  -600
# 8      8  1400    400 -1000
# 9      9   900    250  -650
```

```
# 10    10   500     50  -450
# 11    11  1800    500 -1300
# 12    12   650    100  -550
# 13    13  2800   1200 -1600
# 14    14  1050    300  -750
```

Justificación del método no paramétrico

```
# Verificación de normalidad de las diferencias
shapiro.test(caprinos$dif)
# W = 0.8765, p-value = 0.05234
# El p-valor está justo en el borde. Con n=14, Shapiro-Wilk tiene
# potencia limitada. Pero más relevante que la prueba formal es
# la naturaleza de los datos: los conteos de HPG tienen distribución
# asimétrica por construcción (no pueden ser negativos, y unos pocos
# animales pueden tener conteos muy altos).

# Histograma de las diferencias
hist(caprinos$dif, breaks = 8, main = "Distribución de las diferencias HPG",
     xlab = "Diferencia (día 14 - día 0)", col = "grey80")
# El histograma muestra una cola izquierda larga.
```

La prueba de Shapiro-Wilk no rechaza normalidad al nivel 0.05 ($W = 0.876$, $p = 0.052$), pero el resultado es limítrofe. Con solo 14 observaciones, la prueba tiene poca capacidad para distinguir entre normalidad y una distribución moderadamente sesgada. Los conteos de parásitos, por su naturaleza (valores enteros no negativos con distribución agregada), producen diferencias asimétricas. La prueba de Wilcoxon es la elección más prudente.

Análisis completo

```
# --- Prueba de signos ---
d_cap <- caprinos$dif
d_no_cero <- d_cap[d_cap != 0]  # Todos son no nulos en este caso
n_neg <- sum(d_no_cero < 0)
n_pos <- sum(d_no_cero > 0)

cat("Signos negativos:", n_neg, "| Signos positivos:", n_pos, "\n")
# Signos negativos: 14 | Signos positivos: 0

binom.test(n_pos, length(d_no_cero), p = 0.5, alternative = "two.sided")
# p-value = 0.0001221
# Todas las cabras redujeron su HPG.

# --- Prueba de rangos con signo de Wilcoxon ---
```

```r
wilcox_cap <- wilcox.test(hpg_d14, hpg_d0,
                          paired = TRUE,
                          alternative = "less",   # H1: día 14 < día 0
                          conf.int = TRUE,
                          exact = FALSE)
wilcox_cap
# V = 0, p-value = 0.0001221
# 95 percent confidence interval:
#       -Inf  -475
# (pseudo)median = -700 (Hodges-Lehmann)

# Con coin para valor p exacto:
library(coin)
wilcoxsign_test(hpg_d14 ~ hpg_d0,
                distribution = "exact",
                alternative = "less")
# p-value = 6.104e-05

# --- Tamaño del efecto ---
W_pos <- 0
W_neg <- 14 * 15 / 2   # = 105
r_kerby <- (W_pos - W_neg) / (W_pos + W_neg)
cat("r de Kerby =", r_kerby, "\n")
# r = -1.0 (efecto máximo)

# --- FECRT (reducción del conteo fecal de huevos) ---
media_pre  <- mean(hpg_d0)
media_post <- mean(hpg_d14)
fecrt <- (1 - media_post / media_pre) * 100
cat("FECRT =", round(fecrt, 1), "%\n")
# FECRT ≈ 63.1%

# Un FECRT del 63% está muy por debajo del umbral del 95% que indica
# eficacia aceptable. Esto sugiere resistencia a la ivermectina en
# esta población de nematodos.

# --- Gráfico del caso (Figura 2.5) ---
library(ggplot2)

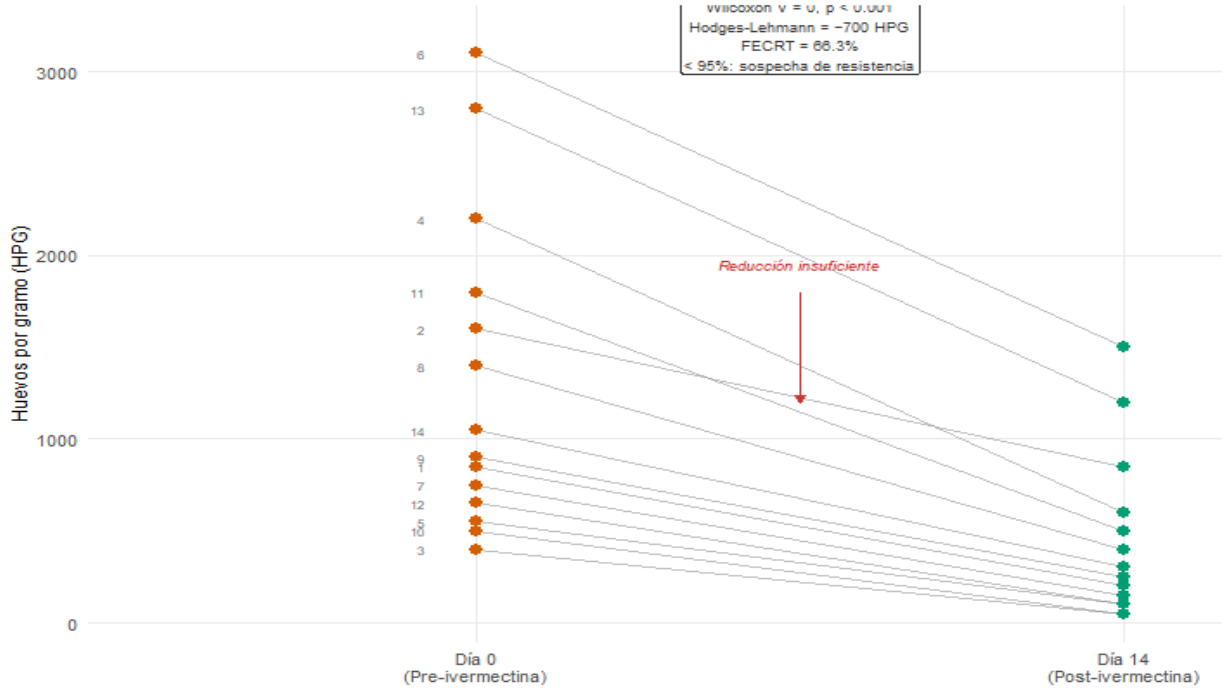
datos_largo <- data.frame(
  cabra   = rep(1:14, 2),
  momento = factor(rep(c("Día 0", "Día 14"), each = 14),
                   levels = c("Día 0", "Día 14")),
  hpg     = c(hpg_d0, hpg_d14)
)

ggplot(datos_largo, aes(x = momento, y = hpg, group = cabra)) +
```

```
  geom_line(color = "grey50", alpha = 0.6) +
  geom_point(aes(color = momento), size = 3) +
  scale_color_manual(values = c("Día 0" = "#D55E00", "Día 14" = "#009E73")) +
  labs(x = "", y = "Huevos por gramo (HPG)",
       title = "Reducción de HPG post-tratamiento con ivermectina",
       subtitle = "14 cabras criollas, altiplano boliviano") +
  annotate("text", x = 1.5, y = 3200,
           label = paste0("Wilcoxon V = 0, p < 0.001\n",
                          "Hodges-Lehmann = -700 HPG\n",
                          "FECRT = 63.1%"),
           size = 3.5, hjust = 0.5) +
  theme_minimal(base_size = 13) +
  theme(legend.position = "bottom",
        legend.title = element_blank())
```

Figura 2

Reducción del conteo de huevos por gramo (HPG) en 14 cabras criollas tras tratamiento con ivermectina



Nota. Todas las líneas descienden, pero la reducción promedio (FECRT = 63.1%) está lejos del 95% requerido, lo que sugiere resistencia antihelmíntica.

Interpretación integrada

La prueba de Wilcoxon confirma que la ivermectina redujo significativamente el conteo de huevos fecales ($V = 0$, $p < 0.001$), con un efecto de magnitud máxima ($r = -1.0$). Sin embargo, la significancia estadística no implica eficacia clínica. El FECRT del 63.1% indica que más de un tercio de los parásitos sobrevivió al tratamiento, un patrón compatible con resistencia antihelmíntica a las lactonas macrocíclicas. El zootecnista debería considerar un coprocultivo para identificar los géneros parasitarios involucrados y evaluar antihelmínticos alternativos (levamisol, ***closantel, monepantel).***

¿Qué habría pasado con la prueba t pareada?

```
t.test(hpg_d14, hpg_d0, paired = TRUE, alternative = "less")
# t = -6.5427, df = 13, p-value = 7.879e-06
# La prueba t también rechaza H0 y con un valor p aún menor.
# En este caso, ambas pruebas concuerdan porque el efecto es tan
# grande que las dos lo detectan sin problema. La diferencia
# habría importado con un efecto más sutil o con datos más
# asimétricos.
```

Con un efecto tan grande (todas las cabras redujeron su HPG), tanto la *t* pareada como Wilcoxon rechazan H_0. La *t* arroja un valor *p* incluso menor (7.9×10^{-6} vs. 6.1×10^{-5}) porque utiliza la magnitud de las diferencias y la variabilidad de forma más eficiente bajo el supuesto de normalidad. La discrepancia entre pruebas habría sido mayor si el efecto fuera marginal o si un par de cabras hubieran mostrado aumento en HPG (animales "no respondedores"), lo cual habría inflado la varianza de las diferencias y reducido la sensibilidad de la *t*.

- ***Caso 2: Evaluación del dolor posoperatorio en caninos***

En el Hospital Veterinario Universitario de la Universidad Mayor de San Andrés (La Paz, Bolivia), se compararon dos protocolos analgésicos para el control del dolor posoperatorio en perros sometidos a cirugía ortopédica (estabilización de ruptura del ligamento cruzado craneal). El grupo A ($n = 11$) recibió meloxicam (0.2 mg/kg IV pre-quirúrgico, seguido de 0.1 mg/kg PO SID por 5 días). El grupo B ($n = 9$) recibió meloxicam más tramadol (2 mg/kg IV cada 8 horas las primeras 24 horas, luego 5 mg/kg PO BID por 5 días). Los perros fueron asignados aleatoriamente a cada protocolo.

El dolor se evaluó 24 horas después de la cirugía usando la escala compuesta de dolor de Glasgow en su versión canina corta (CMPS-SF), con puntajes posibles de 0 a 24 (Reid et al., 2007). Puntajes ≥ 6 se consideran indicativos de dolor que requiere intervención analgésica de rescate. Este es un diseño de dos muestras independientes (distintos perros en cada grupo). La variable es ordinal (puntaje de escala compuesta), los tamaños de muestra son pequeños y desiguales, y la distribución de los puntajes de dolor tiende a estar sesgada a la derecha (la mayoría de los perros tienen dolor leve a moderado, con unos pocos que experimentan dolor severo).

Datos

```
# ============================================================
# CASO INTEGRADOR 2: Dolor posoperatorio en caninos
# ============================================================

# Grupo A: meloxicam solo
dolor_A <- c(8, 12, 6, 9, 15, 7, 10, 5, 11, 8, 14)

# Grupo B: meloxicam + tramadol
dolor_B <- c(4, 6, 3, 7, 5, 2, 8, 4, 6)

# Exploración descriptiva
cat("Grupo A (meloxicam): n =", length(dolor_A),
    ", mediana =", median(dolor_A),
    ", rango =", paste(range(dolor_A), collapse = "-"), "\n")
# Grupo A: n = 11, mediana = 9, rango = 5-15

cat("Grupo B (meloxicam + tramadol): n =", length(dolor_B),
    ", mediana =", median(dolor_B),
    ", rango =", paste(range(dolor_B), collapse = "-"), "\n")
# Grupo B: n = 9, mediana = 5, rango = 2-8
```

Análisis con Mann-Whitney y Brunner-Munzel

```
# --- Mann-Whitney ---
mw_dolor <- wilcox.test(dolor_A, dolor_B,
                        alternative = "greater",  # H1: grupo A > grupo B
                        conf.int = TRUE,
                        exact = FALSE)
mw_dolor
# W = 86, p-value = 0.001523
# 95 percent confidence interval:
#  2.000024      Inf
# difference in location: 4

# Hodges-Lehmann = 4 puntos. El grupo con meloxicam solo presenta
# 4 puntos más de dolor que el grupo con terapia combinada.

# Con coin (valor p exacto):
datos_dolor <- data.frame(
  puntaje = c(dolor_A, dolor_B),
  grupo   = factor(c(rep("Meloxicam", 11), rep("Melox+Tramadol", 9)))
)

mw_dolor_coin <- wilcox_test(puntaje ~ grupo,
                             data = datos_dolor,
                             distribution = "exact",
                             alternative = "greater")
mw_dolor_coin
# p-value = 0.0008741

# --- Brunner-Munzel ---
library(brunnermunzel)
bm_dolor <- brunnermunzel.test(dolor_A, dolor_B,
                               alternative = "greater")
```

```
bm_dolor
# P(X > Y) + 0.5*P(X = Y) = 0.8687
# p-value = 0.001234

# La probabilidad de que un perro del grupo meloxicam solo tenga
# un puntaje de dolor mayor que uno del grupo combinado es del 86.9%.

# --- Tamaños del efecto ---
# r de rango biserial
Z_dolor <- qnorm(mw_dolor$p.value)  # Unilateral, convertir
# Para bilateral sería qnorm(mw_dolor$p.value / 2)
# Recalculando con el p bilateral:
mw_dolor_bil <- wilcox.test(dolor_A, dolor_B,
                            alternative = "two.sided",
                            exact = FALSE)
Z_bil <- qnorm(mw_dolor_bil$p.value / 2)
N_dolor <- length(dolor_A) + length(dolor_B)
r_dolor <- abs(Z_bil) / sqrt(N_dolor)
cat("r =", round(r_dolor, 3), "\n")
# r ≈ 0.63 (efecto grande)

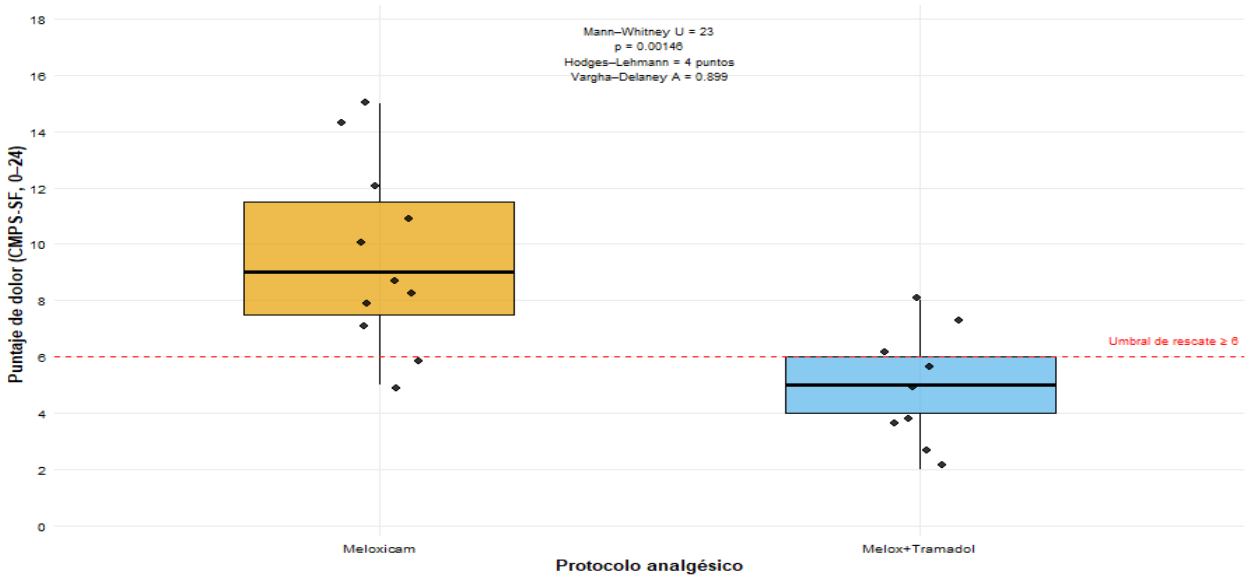
# A de Vargha-Delaney
A_dolor <- mw_dolor$statistic / (length(dolor_A) * length(dolor_B))
cat("A de Vargha-Delaney =", round(A_dolor, 3), "\n")
# A ≈ 0.869 (efecto grande)
library(effsize)
VD.A(dolor_A, dolor_B)
# A estimate: 0.8687 (large)

# --- Gráfico comparativo (Figura 2.6) ---
ggplot(datos_dolor, aes(x = grupo, y = puntaje, fill = grupo)) +
```

```
  geom_boxplot(width = 0.5, alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.08, size = 2.5, alpha = 0.8) +
  geom_hline(yintercept = 6, linetype = "dashed", color = "red",
             linewidth = 0.6) +
  annotate("text", x = 2.4, y = 6.5,
           label = "Umbral de rescate analgésico",
           color = "red", size = 3.2) +
  scale_fill_manual(values = c("Meloxicam" = "#E69F00",
                               "Melox+Tramadol" = "#56B4E9")) +
  labs(x = "Protocolo analgésico",
       y = "Puntaje de dolor (CMPS-SF, 0-24)",
       title = "Dolor posoperatorio 24 h post-cirugía ortopédica canina") +
  annotate("text", x = 1.5, y = 16,
           label = paste0("Mann-Whitney p = 0.002\n",
                          "Hodges-Lehmann = 4 puntos\n",
                          "A de Vargha-Delaney = 0.87"),
           size = 3.3, hjust = 0.5) +
  theme_minimal(base_size = 13) +
```

Figura 3

Comparación de los puntajes de dolor (CMPS-SF) a las 24 horas post-cirugía entre los dos protocolos analgésicos.



Nota. La línea roja discontinua indica el umbral de puntaje ≥ 6 que activa la analgesia de rescate. La mayoría de los perros del grupo meloxicam solo superan este umbral; en el grupo combinado, solo tres lo alcanzan.

Redacción tipo artículo científico (ejemplo para los ejercicios)

> Los puntajes de dolor a las 24 horas posquirúrgicas difirieron entre los protocolos evaluados. El grupo que recibió meloxicam solo presentó puntajes de dolor significativamente mayores (Mdn = 9, rango: 5–15) que el grupo que recibió la combinación de meloxicam con tramadol (Mdn = 5, rango: 2–8), $U = 86$, $p = 0.002$ (prueba exacta de Mann-Whitney). El estimador de Hodges-Lehmann de la diferencia fue de 4 puntos (IC 95%: 2–8), y la probabilidad de que un perro del grupo meloxicam presentara mayor dolor que uno del grupo combinado fue del 86.9% (*A* de Vargha-Delaney = 0.87), lo que constituye un efecto grande. Desde la perspectiva clínica, 9 de 11 perros (81.8%) del grupo meloxicam solo superaron el umbral de rescate analgésico (CMPS-SF ≥ 6), comparado con 3 de 9 (33.3%) en el grupo combinado (prueba exacta de Fisher, $p = 0.036$).

2.13. Errores frecuentes y advertencias

1. Interpretar Mann-Whitney como “prueba de medianas” sin verificar igualdad de formas

Ya se discutió en la Sección 2.6. Si los boxplots muestran distribuciones con dispersión o asimetría muy distintas, la interpretación como "comparación de medianas" no es válida. Reportar las medianas como descriptivos está bien; afirmar que Mann-Whitney "demostró" que las medianas difieren es otra cosa.

2. Usar Wilcoxon signed-rank con variables estrictamente ordinales

Si la variable tiene solo 3 a 5 niveles posibles (como una escala de cojera con grados 1–5), las diferencias son pocas y discretas, y los empates dominan. El rango medio de muchas diferencias empatadas pierde la granularidad que la prueba necesita para funcionar bien. En estas situaciones, la prueba de signos es más apropiada que la de rangos con signo.

3. Olvidar reportar el tamaño del efecto

Un error de omisión, no de comisión. El valor *p* depende del tamaño de la muestra; con *n* suficientemente grande, cualquier diferencia trivial será significativa. Sin una medida de tamaño del efecto, el lector no puede juzgar la relevancia del hallazgo.

4. No corregir por empates

La función *wilcox.test()* de R maneja los empates internamente y emite un warning ("cannot compute exact p-value with ties"), pero algunos investigadores ignoran este aviso y reportan el valor *p* "exacto" que en realidad es asintótico con corrección de continuidad. Para un reporte riguroso, usar *coin::wilcox_test(distribution = "exact")*, que sí calcula la distribución exacta condicional con empates.

5. Confundir el estadístico W de R con el U de las tablas

wilcox.test() reporta *W*, que en el caso pareado es la suma de rangos positivos (T^+) y en el caso de dos muestras es U_1 (el estadístico de Mann-Whitney para el primer grupo). Los libros de texto suelen usar $U = \min(U_1, U_2)$ para buscar en tablas de valores críticos. Si se compara la salida de R directamente con las tablas sin hacer la conversión, se puede llegar a conclusiones erróneas.

6. Aplicar Mann-Whitney a más de dos grupos

Mann-Whitney es para dos grupos. Si se tienen tres o más grupos y se hacen comparaciones múltiples de a pares con Mann-Whitney sin ajuste, se infla el error tipo I. El procedimiento correcto es primero aplicar Kruskal-Wallis (Capítulo 5) y luego, si es significativo, realizar comparaciones post hoc con corrección de Bonferroni, Dunn o similar.

7. Ignorar la direccionalidad de la hipótesis

Si la pregunta de investigación tiene una dirección esperada ("el nuevo tratamiento reduce el dolor", no "el nuevo tratamiento cambia el dolor"), usar una prueba unilateral. Esto aumenta la potencia y es legítimo si la dirección se definió antes de ver los datos, nunca después.

8. No reportar n por grupo, medianas y rangos intercuartílicos

Un error sorprendentemente frecuente en la literatura veterinaria: reportar solo el valor *p* de Mann-Whitney sin indicar las medianas de cada grupo. El lector necesita ambos elementos la significancia y los descriptivos para evaluar el resultado.

2.14. Ejercicios de autoevaluación

Nivel básico

Ejercicio 2.1.

Un investigador quiere comparar la ganancia de peso (kg) en pollos de engorda alimentados con dos dietas diferentes. Tiene 25 pollos por grupo. La ganancia de peso se distribuye normalmente en ambos grupos. ¿Qué prueba debería usar? Justifique.

Respuesta: La prueba *t* para muestras independientes es apropiada. Las muestras son de tamaño moderado, los datos son continuos (peso en kg), se distribuyen normalmente, y los grupos son independientes. No hay razón para recurrir a Mann-Whitney cuando los supuestos de la *t* se cumplen. Usar Mann-Whitney no sería incorrecto, pero implicaría una pérdida innecesaria de potencia (~5%).

Ejercicio 2.2.

Un veterinario evalúa la eficacia de un ungüento cicatrizante en heridas de bovinos. Aplica el ungüento en una herida del flanco derecho y un placebo en una herida del flanco izquierdo del mismo animal. Mide la reducción porcentual del área de la herida a los 7 días en 8 bovinos. ¿Datos pareados o independientes? ¿Qué prueba no paramétrica es la más apropiada?

Respuesta: Son datos pareados (ambas heridas pertenecen al mismo animal, que actúa como su propio control). La prueba no paramétrica apropiada es la de rangos con signo de Wilcoxon, ya que la reducción porcentual es una variable continua (intervalo/razón) y las diferencias entre el lado tratado y el control pueden tener magnitudes informativas. Si la reducción se midiera en una escala ordinal categórica (por ejemplo: "no mejoró", "mejoró poco", "mejoró mucho"), la prueba de signos sería más adecuada.

Ejercicio 2.3.

¿La prueba de Mann-Whitney requiere que los datos se distribuyan normalmente? ¿Requiere algún supuesto sobre las distribuciones? Describa al menos dos supuestos.

Respuesta: No requiere normalidad. Sí requiere: (1) independencia entre las observaciones y entre los grupos; (2) que la variable sea al menos ordinal; y (3) si se quiere interpretar el resultado como una comparación de medianas, se necesita el supuesto adicional de que las dos distribuciones tienen la misma forma y solo difieren en localización (shift model). Sin este último supuesto, Mann-Whitney compara distribuciones completas, no medianas.

Nivel intermedio

Ejercicio 2.4.

Los siguientes puntajes de bienestar animal (escala 1–10) se registraron en 7 cerdos antes y después de implementar un programa de enriquecimiento ambiental en una granja porcina:

Cerdo	Antes	Después
1	3	6
2	4	7
3	5	5
4	2	8

Cerdo	Antes	Después
5	6	7
6	3	4
7	4	9

a) Realice la prueba de signos a mano. Reporte el valor *p* exacto bilateral.
b) Realice la prueba de rangos con signo de Wilcoxon a mano. Calcule W^+ y busque el valor *p*.

Respuestas:

a) Prueba de signos: Diferencias: +3, +3, 0, +6, +1, +1, +5. Diferencias no nulas: $n' = 6$. Signos positivos: $S^+ = 6$. $P(X \geq 6)$ con $X \sim \text{Binomial}(6, 0.5) = (0.5)^6 = 0.015625$. Valor *p* bilateral = 2 × 0.01563 = 0.03125. Se rechaza H_0 al nivel 0.05 (pero no al 0.01).

b) Wilcoxon signed-rank: $|d|$ ordenados (sin el cerdo 3): 1, 1, 3, 3, 5, 6. Rangos: $|d| = 1 \rightarrow$ posiciones 1, 2 → rango medio 1.5. $|d| = 3 \rightarrow$ posiciones 3, 4 → rango medio 3.5. $|d| = 5 \rightarrow$ posición 5 → rango 5. $|d| = 6 \rightarrow$ posición 6 → rango 6. Todos los signos son positivos. $W^+ = 1.5 + 1.5 + 3.5 + 3.5 + 5 + 6 = 21$. $W^- = 0$. Verificación: 21 + 0 = 21 = 6 × 7/2. ✓ Con $n' = 6$, $W^+ = 21$ (máximo posible). *p* bilateral ≈ 0.031 (distribución exacta).

Ejercicio 2.5.

Los tiempos de recuperación anestésica (minutos) de dos protocolos en periquitos australianos son:

- Protocolo 1 (isoflurano): 4.2, 5.1, 6.3, 3.8, 7.5, 4.9
- Protocolo 2 (sevoflurano): 2.1, 3.0, 2.8, 3.5, 2.4, 4.1, 3.3

a) ¿Es apropiada la prueba *t* para muestras independientes? ¿Por qué sí o por qué no?
b) Aplique la prueba de Mann-Whitney. Calcule U_1, U_2 y el estadístico *U*.
c) Calcule el estimador de Hodges-Lehmann.

Respuestas:

a) Las muestras son muy pequeñas ($n_1 = 6$, $n_2 = 7$), lo que impide verificar normalidad con potencia razonable. En medicina de aves, las muestras pequeñas son la norma. La prueba de Mann-Whitney es preferible.

b) Datos combinados y ordenados:

Valor	Grupo	Rango
2.1	2	1
2.4	2	2
2.8	2	3
3.0	2	4
3.3	2	5
3.5	2	6
3.8	1	7
4.1	2	8
4.2	1	9
4.9	1	10
5.1	1	11
6.3	1	12
7.5	1	13

$R_1 = 7 + 9 + 10 + 11 + 12 + 13 = 62$. $U_1 = 62 - 6(7)/2 = 62 - 21 = 41$. $U_2 = 6 \times 7 - 41 = 1$. $U = \min(41, 1) = 1$.

Con $n_1 = 6$, $n_2 = 7$, el valor crítico de U al nivel 0.05 bilateral es 6 (consultar tabla). Como $1 < 6$, se rechaza H_0. El sevoflurano produce tiempos de recuperación más cortos.

c) Se calculan las $6 \times 7 = 42$ diferencias (isoflurano − sevoflurano). La mediana de estas 42 diferencias es el estimador de Hodges-Lehmann. En R: *wilcox.test(iso, sevo, conf.int=TRUE)$estimate*. El valor será cercano a 2.1 minutos (diferencia mediana).

Nivel avanzado

Ejercicio 2.6.

Un estudio multicéntrico evalúa la eficacia de un nuevo antiparasitario (moxidectina al 2.5% pour-on) versus ivermectina al 1% inyectable en el control de garrapatas *Rhipicephalus microplus* en bovinos Brahman de dos fincas del Chaco paraguayo. Se registró el conteo de garrapatas (teleóginas > 4.5 mm en un área estándar del flanco izquierdo) a los 14 días post-tratamiento.

Finca 1 (Mariscal Estigarribia): - Moxidectina (n = 12): 3, 0, 1, 5, 2, 0, 1, 0, 4, 1, 2, 0 - Ivermectina (n = 10): 8, 12, 5, 15, 7, 9, 11, 6, 3, 10

Finca 2 (Filadelfia): - Moxidectina (n = 10): 1, 3, 0, 2, 1, 4, 0, 2, 1, 0 - Ivermectina (n = 11): 6, 9, 14, 7, 11, 5, 8, 4, 13, 7, 10

Realice el análisis completo para cada finca: a) Justifique por qué se elige un método no paramétrico. b) Aplique la prueba de Mann-Whitney con R (muestre el código). c) Aplique la prueba de Brunner-Munzel e interprete $P(X > Y)$. d) Calcule el tamaño del efecto (r y A de Vargha-Delaney). e) Redacte los resultados como aparecerían en la sección de Resultados de un artículo para *Preventive Veterinary Medicine*.

Rúbrica orientadora:

Criterio	**Insuficiente**	**Aceptable**	**Sobresaliente**
Justificación del método	No menciona los supuestos violados	Menciona no normalidad o muestras pequeñas	Identifica la naturaleza discreta de los conteos, la asimetría esperada y la imposibilidad de verificar normalidad con estos tamaños de muestra
Código en R	Fragmentos sueltos o con errores de sintaxis	Código funcional con las pruebas solicitadas	Código completo, comentado, con ambas funciones (base y coin), gráficos y tamaños del efecto
Interpretación	Solo reporta valores p	Reporta medianas, p y menciona significancia	Integra el resultado estadístico con la relevancia parasitológica, discute limitaciones del estudio

Redacción científica	No sigue formato de artículo	Sigue el formato, pero omite descriptivos o tamaño del efecto	Incluye medianas con RIQ, *U*, *p* exacto, tamaño del efecto con IC, y discusión clínica en formato APA 7.ª ed.

Ejercicio 2.7.

Un investigador realizó una prueba de Mann-Whitney para comparar el recuento de células somáticas (×10³/mL) en leche de vacas con mastitis subclínica tratadas con terapia intramamaria convencional (n = 15) versus terapia con lisado de fagos bacterianos (n = 12). Obtuvo U = 45, p = 0.048. Reportó: "La terapia con fagos produjo una mediana de recuento de células somáticas significativamente menor que la terapia convencional (Mann-Whitney U = 45, p = 0.048)."

a) ¿Qué información falta en este reporte?
b) ¿Bajo qué condición la interpretación de "mediana menor" es válida?
c) ¿Qué medida de tamaño del efecto recomendaría?
d) Reescriba el resultado incorporando las correcciones.

Respuesta orientadora:

a) Faltan: las medianas y rangos intercuartílicos de cada grupo, el tamaño del efecto, si la prueba fue exacta o asintótica, y si fue unilateral o bilateral.

b) Solo es válida si se verificó que las distribuciones tienen la misma forma (solo difieren en localización). Los recuentos de células somáticas suelen tener distribución log-normal, con variabilidad que puede diferir entre grupos, lo cual viola este supuesto.

c) A de Vargha-Delaney, porque se interpreta directamente como la probabilidad de que un animal del grupo fagos tenga un recuento menor que uno del grupo convencional, lo cual es clínicamente informativo.

d) Ejemplo de reescritura: "El recuento de células somáticas fue menor en el grupo tratado con lisados de fagos (Mdn = [valor], RIQ: [valor]) que en el grupo con terapia convencional

(Mdn = [valor], RIQ: [valor]). La diferencia fue estadísticamente significativa según la prueba exacta de Mann-Whitney (U = 45, p = 0.048, bilateral). El estimador de Hodges-Lehmann de la diferencia fue de [valor] × 10^3 células/mL (IC 95%: [valor, valor]), con un tamaño del efecto grande (A = [valor]). Sin embargo, la dispersión fue notablemente mayor en el grupo convencional, lo que sugiere que la diferencia detectada puede reflejar no solo un desplazamiento de la mediana sino también una diferencia en la variabilidad de la respuesta al tratamiento."

Ejercicio 2.8.

Diseñe un estudio hipotético (pero realista) en medicina de équidos donde la prueba de signos sea más apropiada que la prueba de rangos con signo de Wilcoxon. Especifique: la variable de respuesta, el diseño del estudio, por qué la variable es ordinal (no de intervalo), y por qué la magnitud de las diferencias no es informativa. Escriba las hipótesis en notación estadística.

Rúbrica orientadora: Un estudio que evalúe la severidad de cojera (escala de 0 a 5, AAEP) antes y después de un tratamiento articular, donde la diferencia entre grado 3 y grado 4 no tiene la misma magnitud que entre grado 1 y grado 2 (los intervalos no son iguales), y por lo tanto la asignación de rangos a las diferencias carece de justificación.

2.15. Conclusiones del capítulo

Las pruebas abordadas en este capítulo signos, rangos con signo de Wilcoxon, Mann-Whitney U, Brunner-Munzel constituyen la primera línea de herramientas no paramétricas para el investigador veterinario que enfrenta datos que no cumplen los supuestos de la prueba *t*. Cada prueba tiene su nicho:

La prueba de signos es la más simple y la más conservadora, y funciona con cualquier variable ordinal sin restricciones sobre la forma de la distribución. Su precio es la baja potencia.

La prueba de rangos con signo de Wilcoxon incorpora la magnitud de las diferencias y gana potencia a cambio de asumir simetría en la distribución de las diferencias. Es la alternativa más directa a la prueba *t* pareada.

Mann-Whitney U es la prueba estándar para dos muestras independientes. Compara distribuciones completas, no solo medianas una distinción que merece atención al interpretar resultados, y que el estimador de Hodges-Lehmann complementa al ofrecer una estimación puntual con intervalo de confianza.

Brunner-Munzel resuelve el problema que surge cuando las distribuciones de los dos grupos no solo están desplazadas, sino que tienen formas distintas, una situación frecuente en datos veterinarios con variabilidad biológica alta.

Ninguna de estas pruebas reemplaza la necesidad de describir los datos adecuadamente (medianas, rangos intercuartílicos, gráficos) ni de reportar tamaños del efecto. El valor *p* solo responde a una pregunta muy estrecha; la interpretación clínica requiere más.

El próximo capítulo (Capítulo 3, o según la numeración adoptada para el libro) extiende estas herramientas a la comparación de tres o más grupos con la prueba de Kruskal-Wallis y sus procedimientos de comparaciones múltiples, donde las consecuencias de no ajustar por comparaciones múltiples y de ignorar los supuestos sobre la forma de las distribuciones se amplifican.

Capítulo 3

Comparación de múltiples grupos: Kruskal-Wallis, Friedman y comparaciones post hoc

Introducción

Este capítulo extiende la comparación no paramétrica a diseños con tres o más grupos. Se trabajan dos pruebas principales según la estructura del diseño experimental. Para grupos independientes se desarrolla la prueba de Kruskal-Wallis: la construcción del estadístico H a partir de los rangos globales, la corrección por empates, la distribución de referencia (chi-cuadrado con $k - 1$ grados de libertad para muestras grandes, distribución exacta para muestras pequeñas) y tres procedimientos de comparación post hoc con sus diferencias en lógica y potencia: Dunn con ajuste por comparaciones múltiples (Bonferroni, Holm, BH), Conover-Iman basado en la estadística t con varianza agrupada de rangos, y Dwass-Steel-Critchlow-Fligner como procedimiento libre de supuestos distribucionales adicionales. Para diseños con bloques o medidas repetidas se presenta la prueba de Friedman: el estadístico basado en rangos dentro de bloques, la corrección por empates, las comparaciones post hoc de Nemenyi y Conover, y la alternativa de Quade para situaciones donde los bloques difieren sustancialmente en magnitud. Se desarrolla el coeficiente de concordancia W de Kendall como medida de acuerdo entre evaluadores o entre ocasiones de medición, y se trabajan las medidas de tamaño del efecto para cada prueba: épsilon cuadrada y eta cuadrada basado en H para Kruskal-Wallis, W de Kendall y r por pares para Friedman. Se incluye la comparación con ANOVA de un factor y ANOVA de medidas repetidas para que el lector evalúe las consecuencias de elegir un método sobre otro.

Objetivos de aprendizaje

Al finalizar este capítulo, el lector será capaz de:

1. **Identificar** las situaciones experimentales en investigación veterinaria grupos independientes versus diseños con bloques o medidas repetidas que requieren una prueba no paramétrica para comparar tres o más condiciones, y distinguir cuándo corresponde Kruskal-Wallis y cuándo Friedman.
2. **Aplicar** las pruebas de Kruskal-Wallis y Friedman a datos veterinarios reales, tanto mediante el cálculo manual del estadístico como mediante su ejecución en R, incluyendo la corrección por empates cuando corresponda.
3. **Seleccionar y ejecutar** el procedimiento de comparación post hoc adecuado Dunn, Conover-Iman, Dwass-Steel-Critchlow-Fligner para Kruskal-Wallis; Nemenyi y Conover

para Friedman según el diseño experimental, y justificar la elección del método de ajuste por comparaciones múltiples.

4. **Interpretar** los indicadores de tamaño del efecto (épsilon cuadrada, eta cuadrada basado en H, coeficiente de concordancia W de Kendall, *r* por pares) y comunicar su significado biológico y clínico en el contexto veterinario, más allá del simple reporte del valor p.
5. **Contrastar** las decisiones analíticas, los supuestos y las limitaciones de Kruskal-Wallis y Friedman frente a sus equivalentes paramétricos (ANOVA de un factor y ANOVA de medidas repetidas), y redactar la sección de resultados de acuerdo con la norma APA 7.ª edición.

Introducción con problema real

En la Estación Experimental de Tizimín, Yucatán, un equipo de parasitólogos veterinarios necesita comparar la carga de garrapatas *Rhipicephalus microplus* en bovinos de cuatro genotipos mantenidos bajo las mismas condiciones de pastoreo tropical: Brahman, Gyr, cruza F1 Holstein × Cebú, y Holstein puro. Disponen de 12 vacas primíparas por genotipo 48 animales en total y registran el número de garrapatas adultas ingurgitadas (mayores de 4 mm) en el costado izquierdo de cada animal tras 21 días de infestación natural durante la temporada de lluvias. Los conteos varían de 3 garrapatas en una vaca Brahman hasta 387 en una Holstein. La distribución dentro de cada genotipo es violentamente asimétrica hacia la derecha: la mayoría de los animales cebuínos tienen conteos bajos, pero algunos acumulan unas pocas decenas; los Holstein, en cambio, muestran conteos que van de 150 a casi 400, con una varianza interna veinte veces mayor que la de los Brahman. Esos extremos no son errores de registro: los animales susceptibles simplemente acumulan muchas más garrapatas, y eliminarlos como “outliers” sería eliminar la información más valiosa del estudio.

El primer reflejo del investigador es plantear un ANOVA de un factor. Verifica los supuestos: Shapiro-Wilk rechaza la normalidad de los residuos ($p = 0.003$), la prueba de Levene devuelve heterocedasticidad extrema ($F_{3,44} = 14.8$, $p < 0.001$), y la inspección visual confirma colas largas y asimetría. Intenta transformar los datos con logaritmo natural, $\log(y + 1)$, raíz cuadrada y la transformación de Box-Cox. Ninguna logra normalizar los residuos y homogeneizar las varianzas simultáneamente. El investigador se pregunta entonces: ¿cómo comparar cuatro grupos cuando los

supuestos del ANOVA paramétrico no se cumplen, las transformaciones fracasan, y la variable tiene sentido biológico en su escala original?

En otro proyecto del mismo departamento, tres veterinarios andrólogos evalúan la calidad espermática de 15 toros Suizo Pardo. Cada andrólogo examina los mismos 15 eyaculados y asigna un puntaje en una escala ordinal de cinco niveles (1 = inaceptable, 5 = excelente), basada en motilidad progresiva, morfología y concentración. El interés no es solo saber si los tres evaluadores califican igual que ya se anticipa que no, porque el andrólogo B tiene fama de ser más estricto, sino cuantificar si están de acuerdo en el *orden* relativo de calidad de los toros: ¿identifican al mismo toro como el mejor y al mismo como el peor? Los datos son ordinales (la distancia entre "bueno" y "muy bueno" no es la misma que entre "malo" y "aceptable") y están pareados por bloque (cada toro fue evaluado por los tres andrólogos). Ni el ANOVA de medidas repetidas ni la prueba de Kruskal-Wallis son herramientas correctas aquí.

Este capítulo desarrolla las pruebas para ambas situaciones. Para grupos independientes: la prueba de Kruskal-Wallis, su corrección por empates, y tres procedimientos de comparación post hoc (Dunn, Conover-Iman, Dwass-Steel-Critchlow-Fligner). Para diseños con bloques o medidas repetidas: la prueba de Friedman, la alternativa de Quade, las comparaciones post hoc de Nemenyi y Conover, y el coeficiente de concordancia W de Kendall. Para ambos contextos: los tamaños del efecto que complementan la decisión estadística con información sobre la magnitud de las diferencias. Todo ello con código reproducible en R, datos de contexto veterinario, y las precauciones que la experiencia enseña a base de tesis mal analizadas.

3.1. Prueba de Kruskal-Wallis: lógica, estadístico H y corrección por empates

3.1.1. De Mann-Whitney a Kruskal-Wallis: la extensión natural

En el capítulo anterior se presentó la prueba de Mann-Whitney para comparar dos grupos independientes mediante rangos. La prueba de Kruskal-Wallis (Kruskal & Wallis, 1952) es su extensión a tres o más grupos. La relación es análoga a la que existe entre la prueba t de Student para dos muestras independientes y el ANOVA de un factor: la misma lógica, generalizada.

La idea central puede enunciarse en tres pasos que cualquier estudiante de veterinaria puede seguir con lápiz y papel:

1. Se juntan todas las observaciones de todos los grupos en un solo conjunto, como si no existieran los grupos, y se les asignan rangos de 1 a N (donde N es el total de observaciones).
2. Se vuelve a cada grupo y se suma los rangos que le tocaron.
3. Se compara esa suma de rangos observada con lo que se esperaría si los grupos fueran todos iguales (es decir, si los rangos se hubieran repartido al azar entre los grupos).

Si un grupo acapara los rangos bajos (valores pequeños) y otro los rangos altos (valores grandes), las sumas de rangos se alejarán de lo esperado bajo igualdad, y el estadístico H será grande. Si los rangos están repartidos de forma más o menos uniforme entre los grupos, H será pequeño.

Para hacer esto tangible con el ejemplo de las garrapatas: si los cuatro genotipos no difirieran en carga parasitaria, cada grupo de 12 animales debería tener una suma de rangos cercana a 12 × (48 + 1)/2 = 294. Si en cambio los 12 Brahman concentran los rangos 1 a 12 y los 12 Holstein los rangos 37 a 48, las sumas de rangos serán 78 y 510 respectivamente muy lejos de 294 y H lo reflejará.

3.1.2. Hipótesis estadísticas

La formulación correcta de las hipótesis es más sutil de lo que parece en muchos textos introductorios.

H_0: Las k poblaciones tienen la misma distribución. En notación formal: $F_1(x) = F_2(x) = \cdots = F_k(x)$ para todo valor de x.

H_1: Al menos una de las poblaciones difiere en distribución respecto a las demás.

Un error que aparece repetidamente en tesis de veterinaria y, para ser honestos, también en artículos publicados en revistas indexadas es escribir "H_0: las medianas de los k grupos son iguales". Esto no es lo que prueba Kruskal-Wallis. La hipótesis nula es sobre las distribuciones completas, no solo sobre las medianas. ¿Cuándo coinciden? Cuando las distribuciones de todos los grupos tienen la misma forma y solo difieren en su localización (es decir, están desplazadas horizontalmente). En ese caso, rechazar H_0 sí equivale a decir que las medianas difieren. Pero si las formas son distintas un grupo tiene distribución simétrica y otro tiene distribución sesgada a la

derecha, se puede rechazar H_0 sin que las medianas difieran en absoluto. Este no es un tecnicismo académico: en datos veterinarios, donde un grupo de animales susceptibles puede tener una distribución sesgada y otro grupo más homogéneo puede tener una distribución simétrica, la distinción tiene consecuencias reales para la interpretación.

Hart (2001) y Fagerland y Sandvik (2009) discuten este problema con rigor. La recomendación práctica es: antes de interpretar un Kruskal-Wallis significativo como "las medianas difieren", inspeccionar visualmente las distribuciones con boxplots y/o histogramas. Si las formas son razonablemente similares, la interpretación en términos de medianas es aceptable. Si no lo son, la interpretación correcta es "las distribuciones difieren", y hay que ser cuidadoso al comunicar qué significa eso en términos biológicos.

3.1.3. Supuestos de la prueba

Hay una creencia extendida de que las pruebas no paramétricas "no tienen supuestos". Eso es falso. Kruskal-Wallis tiene supuestos; son menos restrictivos que los del ANOVA paramétrico, pero existen y deben verificarse:

1. Independencia de las observaciones. Cada observación dentro de un grupo y entre grupos debe ser independiente de las demás. Este es el supuesto más delicado y el más frecuentemente violado en investigación veterinaria, a menudo sin que el investigador lo perciba. Violaciones comunes: medir múltiples lesiones en el mismo animal y tratarlas como observaciones independientes; incluir crías del mismo parto como unidades independientes; medir al mismo animal en varios tiempos y analizar con Kruskal-Wallis en lugar de Friedman; tomar muestras de varios animales del mismo corral sin considerar el efecto corral (pseudorreplicación, sensu Hurlbert, 1984).

2. Variable al menos ordinal. Los datos deben poder ordenarse de menor a mayor de forma no ambigua. Esto incluye variables continuas (peso, concentración sérica), conteos (huevos por gramo de heces, número de garrapatas), y escalas ordinales (puntaje de condición corporal, grado de claudicación). No incluye variables nominales sin orden intrínseco (raza, color del pelaje) para esas se necesitan pruebas de asociación como chi-cuadrado, que se tratarán en el Capítulo 7.

3. Distribuciones de la misma forma (para interpretar como comparación de medianas). Como se explicó arriba, si las distribuciones no tienen la misma forma, la prueba sigue siendo válida pero la interpretación cambia. Este es un supuesto de interpretación, no de validez.

Lo que Kruskal-Wallis *no* asume: normalidad de los datos, normalidad de los residuos, homocedasticidad (igualdad de varianzas), escala intervalar o de razón. Ahí reside su ventaja.

3.1.4. El estadístico H: desarrollo paso a paso

Sea un experimento con k grupos, donde el grupo i tiene n_i observaciones y el total de observaciones es $N = \Sigma n_i$. Se asignan rangos de 1 a N al conjunto combinado de todas las observaciones (del valor más pequeño al más grande). Sea R_i la suma de rangos del grupo i.

Si la hipótesis nula es cierta (todas las distribuciones son iguales), el rango esperado de cualquier observación es (N + 1)/2, y la suma esperada de rangos del grupo i es:

$$E(R_i) = n_i \cdot \frac{N+1}{2}$$

El estadístico H mide qué tan lejos están las sumas de rangos observadas de sus valores esperados, estandarizando para que la comparación sea equitativa independientemente del número de grupos y de los tamaños de muestra:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

Cada símbolo:

- k: número de grupos (en nuestro ejemplo de garrapatas, k = 4).
- n_i: número de observaciones en el grupo i ($n_1 = n_2 = n_3 = n_4 = 12$).
- N: total de observaciones (N = 48).
- R_i: suma de rangos del grupo i.

- 12/[N(N + 1)]: factor de escala que normaliza el estadístico.
- 3(N + 1): término de centrado que hace que H = 0 cuando las sumas de rangos coinciden exactamente con lo esperado.

Una forma equivalente, a veces más intuitiva, usa los rangos medios ($\bar{R}_i = R_i/n_i$) y el rango medio global ($\bar{R} = (N+1)/2$):

$$H = \frac{N-1}{s_R^2} \sum_{i=1}^{k} n_i \, (\bar{R}_i - \bar{R})^2$$

donde s_R^2 es la varianza de todos los rangos combinados. Esta forma revela que H es esencialmente un ANOVA de un factor calculado sobre los rangos: el numerador mide la variabilidad entre grupos (en términos de rangos) y el denominador la variabilidad total.

Distribución de referencia. Bajo H_0, con tamaños de muestra razonablemente grandes ($n_i \geq 5$ como regla práctica general), H se distribuye aproximadamente como chi-cuadrado con $k - 1$ grados de libertad. Se rechaza H_0 cuando $H \geq \chi^2(\alpha, k-1)$. Para muestras más pequeñas, la aproximación chi-cuadrado puede ser inexacta, y se dispone de tablas de distribución exacta (Siegel & Castellan, 1988, Tabla N; el paquete *coin* en R puede calcular valores p exactos o **basados en permutaciones).**

3.1.5. Ejemplo resuelto a mano: tres antihelmínticos en cabritos

Un veterinario del altiplano boliviano prueba tres tratamientos antihelmínticos en cabritos Boer de tres meses de edad infectados naturalmente con *Haemonchus contortus*. A los 14 días posdosificación, realiza conteos de huevos por gramo de heces (HPG) mediante la técnica de McMaster. Los datos son:

Tabla 7

Conteo de huevos por gramo de heces (HPG) de Haemonchus contortus a los 14 días posdosificación en cabritos tratados con tres antihelmínticos

Cabrito	Ivermectina (A)	Albendazol (B)	Control sin tratamiento (C)
1	50	120	350

2	30	200	500
3	80	150	420
4	10	180	380

N = 12, k = 3, $n_1 = n_2 = n_3 = 4$.

Paso 1. Ordenar el conjunto combinado de valores y asignar rangos:

Tabla 8

Asignación de rangos al conjunto combinado de observaciones.

Posición	Valor	Grupo	Rango
1	10	A	1
2	30	A	2
3	50	A	3
4	80	A	4
5	120	B	5
6	150	B	6
7	180	B	7
8	200	B	8
9	350	C	9
10	380	C	10
11	420	C	11
12	500	C	12

No hay empates en este ejemplo, lo que simplifica el cálculo.

Paso 2. Sumas de rangos por grupo:

- R_A = 1 + 2 + 3 + 4 = 10
- R_B = 5 + 6 + 7 + 8 = 26
- R_C = 9 + 10 + 11 + 12 = 42

Verificación: R_A + R_B + R_C = 10 + 26 + 42 = 78 = N(N + 1)/2 = 12 × 13/2 = 78 ✓

Rangos medios: $\bar{R}_A$ = 10/4 = 2.5; $\bar{R}_B$ = 26/4 = 6.5; $\bar{R}_C$ = 42/4 = 10.5.

Rango medio global esperado: $\bar{R}$ = (12 + 1)/2 = 6.5.

Paso 3. Cálculo de H:

$$H = \frac{12}{12 \times 13}\left(\frac{10^2}{4} + \frac{26^2}{4} + \frac{42^2}{4}\right) - 3(13)$$

$$H = \frac{12}{156}\left(\frac{100 + 676 + 1764}{4}\right) - 39$$

$$H = 0.07692 \times 635.0 - 39 = 48.846 - 39 = 9.846$$

Paso 4. Decisión:

Con k − 1 = 2 grados de libertad, los valores críticos de χ^2 son: - $\chi^2_{0.05}$ = 5.991 - $\chi^2_{0.01}$ = 9.210

Como 9.846 > 9.210, se rechaza H_0 al nivel α = 0.01. Hay evidencia estadística sólida de que al menos uno de los tratamientos difiere en la distribución de HPG.

La inspección de los rangos medios muestra un gradiente claro: la ivermectina produjo los conteos más bajos (rango medio 2.5), seguida del albendazol (rango medio 6.5) y el control (rango medio 10.5). Pero atención: con solo 4 animales por grupo, este resultado debe interpretarse con cautela. La potencia estadística es baja, y el hecho de que sea significativo refleja la magnitud extrema de las diferencias, no la precisión del diseño. Para identificar cuáles pares difieren significativamente, se necesitan comparaciones post hoc (sección 3.2).

Nota sobre el tamaño de muestra. Con n = 4 por grupo, la aproximación chi-cuadrado puede no ser precisa. En R, la función *kruskal_test()* del paquete *coin* permite calcular el valor p exacto basado en permutaciones, lo cual es preferible para muestras tan pequeñas. Se verá en la sección 3.10.

3.1.6. Corrección por empates

En el ejemplo anterior no hubo empates (todos los valores eran distintos). En datos veterinarios reales, los empates son frecuentes, y a veces extensos:

- Conteos parasitarios donde muchos animales tienen 0 HPG después del tratamiento.

- Escalas clínicas donde múltiples pacientes reciben el mismo puntaje (8 de 20 perras con grado 2 de claudicación).
- Variables dicotomizadas artificialmente (porcentaje de concepción redondeado a 0 o 100).

Cuando hay empates, se asigna a las observaciones empatadas el promedio de los rangos que les habrían correspondido si no estuvieran empatadas. Por ejemplo, si dos observaciones comparten las posiciones 5 y 6, ambas reciben el rango (5 + 6)/2 = 5.5.

Los empates reducen la variabilidad total de los rangos y pueden hacer que H subestime levemente las diferencias. La corrección por empates divide H entre un factor de deflación:

$$H_{corregido} = \frac{H}{1 - \frac{\sum_{j=1}^{g}(t_j^3 - t_j)}{N^3 - N}}$$

donde:

- g = número de grupos de empates (conjuntos de observaciones con el mismo valor).
- t_j = número de observaciones en el grupo de empates j.

El denominador siempre es ≤ 1, por lo que $H_{corregido} \geq H$: la corrección nunca disminuye el estadístico, solo lo aumenta (o lo deja igual si no hay empates, porque entonces la sumatoria es cero y el denominador es 1).

Ejemplo con empates: puntaje de condición corporal

Supónganse tres grupos de ovejas Pelibuey (8 por grupo) evaluadas en condición corporal con escala de 1 a 5 en intervalos de 0.5:

Grupo 1 (suplementadas)	Grupo 2 (pastoreo libre)	Grupo 3 (confinamiento)
3.5	2.5	2.0
3.0	3.0	2.5
3.5	2.5	1.5
4.0	3.0	2.0
3.0	2.0	2.0
3.5	2.5	1.5

Grupo 1 (suplementadas)	Grupo 2 (pastoreo libre)	Grupo 3 (confinamiento)
4.0	3.0	2.5
3.0	2.0	2.0

N = 24. Los valores posibles son 1.5, 2.0, 2.5, 3.0, 3.5, 4.0. Hay numerosos empates.

Ordenando todos los valores: 1.5, 1.5, 2.0, 2.0, 2.0, 2.0, 2.0, 2.5, 2.5, 2.5, 2.5, 2.5, 3.0, 3.0, 3.0, 3.0, 3.0, 3.0, 3.5, 3.5, 3.5, 3.5, 4.0, 4.0.

Rangos promedio:

- 1.5 aparece 2 veces (posiciones 1-2): rango = 1.5
- 2.0 aparece 5 veces (posiciones 3-7): rango = 5.0
- 2.5 aparece 5 veces (posiciones 8-12): rango = 10.0
- 3.0 aparece 6 veces (posiciones 13-18): rango = 15.5
- 3.5 aparece 4 veces (posiciones 19-22): rango = 20.5
- 4.0 aparece 2 veces (posiciones 23-24): rango = 23.5

Grupos de empates: g = 6, con $t_1 = 2$, $t_2 = 5$, $t_3 = 5$, $t_4 = 6$, $t_5 = 4$, $t_6 = 2$.

Factor de corrección:

$$\sum\left(t_j^3 - t_j\right) = (8-2) + (125-5) + (125-5) + (216-6) + (64-4) + (8-2)$$
$$= 6 + 120 + 120 + 210 + 60 + 6 = 522$$

$$N^3 - N = 13824 - 24 = 13800$$

$$\text{Denominador} = 1 - \frac{522}{13800} = 1 - 0.03783 = 0.96217$$

En este caso, la corrección aumentaría H en un 3.9 %. Con empates moderados como estos, la corrección rara vez cambia la decisión. Pero si la escala tiene solo 3 o 4 niveles posibles y la muestra es grande, la corrección puede importar: no es raro ver factores de corrección de 10 % o más con datos de condición corporal en escala de 1 a 5 y muestras de 50 o más animales.

Dato operativo para R: la función *kruskal.test()* del paquete base aplica la corrección por empates en el cálculo del estadístico (contrario a lo que se afirma en algunos textos). La función

kruskal_test() del paquete *coin* también la aplica. La diferencia entre ambas funciones no está en el manejo de empates, sino en que *coin* ofrece distribuciones exactas y de permutación como alternativa a la aproximación chi-cuadrado.

3.1.7. Un segundo ejemplo resuelto: clasificación histopatológica de tumores mamarios en perras

Para mostrar la prueba con una variable realmente ordinal no un conteo y con empates extensos, considérese el siguiente escenario: un patólogo clasifica tumores mamarios extirpados quirúrgicamente de perras de tres razas (Poodle, Cocker Spaniel, mestizas) según el grado histopatológico de malignidad de Elston y Ellis adaptado para especies caninas:

- Grado I: bien diferenciado (pronóstico favorable)
- Grado II: moderadamente diferenciado
- Grado III: pobremente diferenciado (pronóstico desfavorable)

Los datos (codificados numéricamente como 1, 2, 3):

Poodle (n = 10)	Cocker (n = 10)	Mestizas (n = 10)
1	2	1
2	3	2
1	2	3
1	2	1
2	3	2
1	2	1
2	3	2
1	1	1
1	2	2
2	2	1

N = 30. Los valores posibles son 1, 2, 3 solo tres niveles, con empates masivos.

La pregunta biológica es: ¿existe asociación entre la raza y el grado de malignidad de los tumores mamarios? Kruskal-Wallis puede responder esta pregunta porque la variable respuesta (grado

histopatológico) es ordinal. La corrección por empates será sustancial porque casi todos los valores están empatados.

El cálculo completo a mano sería laborioso pero el procedimiento es idéntico: asignar rangos promedio al conjunto combinado, calcular sumas de rangos por grupo, aplicar la fórmula de H, corregir por empates. En la sección 3.10 se resolverá en R.

Nótese que este mismo problema podría abordarse con una tabla de contingencia y una prueba de chi-cuadrado (Capítulo 7). La diferencia es que chi-cuadrado trata los grados como categorías sin orden, mientras que Kruskal-Wallis aprovecha el orden (Grado I < Grado II < Grado III). Cuando existe un orden natural, Kruskal-Wallis tiene más potencia para detectar tendencias.

3.2. Comparaciones post hoc para Kruskal-Wallis

Un resultado significativo de Kruskal-Wallis dice que al menos un grupo difiere del resto. No dice más. No identifica cuáles grupos difieren entre sí. Para un investigador que compara cuatro genotipos bovinos, saber que "hay diferencia" es útil pero insuficiente: necesita saber si el Brahman difiere del Holstein, si la cruza F1 difiere del Gyr, y así sucesivamente.

Para responder esas preguntas se necesitan comparaciones post hoc (o comparaciones por pares). Existen varios procedimientos, cada uno con su lógica interna, sus ventajas y sus limitaciones. La elección no es arbitraria y merece atención.

3.2.1. Prueba de Dunn

La prueba de Dunn (Dunn, 1964) es el procedimiento post hoc más citado tras Kruskal-Wallis en la literatura veterinaria. Compara las diferencias entre los rangos medios de cada par de grupos.

Estadístico. Para comparar los grupos *i* y *j*:

$$z_{ij} = \frac{\bar{R}_i - \bar{R}_j}{\sigma_{ij}}$$

donde $\bar{R}_i = R_i/n_i$ es el rango medio del grupo i, y el error estándar de la diferencia es:

$$\sigma_{ij} = \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Si hay empates, el término N(N + 1)/12 se reemplaza por la varianza de los rangos ajustada:

$$\sigma_{ij} = \sqrt{\left[\frac{N(N+1)}{12} - \frac{\sum_{j=1}^{g}\left(t_j^3 - t_j\right)}{12(N-1)}\right]\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

El valor de *z* se compara con la distribución normal estándar, pero no directamente: como se realizan múltiples comparaciones, el valor p debe ajustarse.

Ajuste por comparaciones múltiples. Con *k* grupos hay *k(k − 1)/2* comparaciones posibles. Con 4 grupos son 6 comparaciones; con 5 grupos, 10; con 6 grupos, 15. Sin ajuste, la probabilidad de encontrar al menos un falso positivo (error de tipo I familywise) se infla rápidamente.

Los métodos de ajuste más utilizados son:

- **Bonferroni:** multiplica cada valor p por el número de comparaciones. Es el más conservador. Si hay 6 comparaciones y un valor p crudo de 0.01, el valor p ajustado es 0.06 no significativo al 5 %. Es excesivamente conservador cuando hay muchos grupos.

- **Holm (1979):** una versión secuencial de Bonferroni que es uniformemente más potente. Ordena los valores p de menor a mayor y los ajusta secuencialmente, multiplicando el más pequeño por el número total de comparaciones, el segundo por el total menos uno, y así. Es menos conservador que Bonferroni y no sacrifica el control del error de tipo I. No hay razón práctica para preferir Bonferroni sobre Holm; Holm siempre es al menos igual y generalmente mejor.

- **Benjamini-Hochberg (BH):** controla la tasa de falsos descubrimientos (FDR) en lugar de la tasa de error familywise. Es menos conservador que Holm y más adecuado cuando se hacen muchas comparaciones y se tolera una proporción pequeña de falsos positivos entre los resultados significativos. Es popular en genómica veterinaria pero menos habitual en ensayos clínicos con pocos grupos.

Ventajas de Dunn: utiliza los rangos del conjunto combinado original, lo que mantiene coherencia con el estadístico H de Kruskal-Wallis. Es la prueba post hoc más reportada en la literatura y la que los revisores esperan ver.

Limitaciones: con el ajuste de Bonferroni es conservadora. Detecta diferencias grandes, pero puede pasar por alto diferencias moderadas que otros procedimientos sí captan.

Ejemplo numérico: Dunn para los antihelmínticos en cabritos

Con los datos de la sección 3.1.5:

$\bar{R}_A$ = 2.5, $\bar{R}_B$ = 6.5, $\bar{R}_C$ = 10.5, N = 12.

Comparación A vs C:

$$\sigma_{AC} = \sqrt{\frac{12 \times 13}{12}\left(\frac{1}{4} + \frac{1}{4}\right)} = \sqrt{13 \times 0.5} = \sqrt{6.5} = 2.550$$

$$z_{AC} = \frac{2.5 - 10.5}{2.550} = \frac{-8.0}{2.550} = -3.137$$

El valor p bilateral para z = −3.137 es 0.0017. Con ajuste de Bonferroni (3 comparaciones): p_adj = 0.0017 × 3 = 0.0051. Significativo.

Comparación A vs B:

$$z_{AB} = \frac{2.5 - 6.5}{2.550} = \frac{-4.0}{2.550} = -1.569$$

Valor p bilateral: 0.1168. Con Bonferroni: p_adj = 0.1168 × 3 = 0.3504. No significativo.

Comparación B vs C:

$$z_{BC} = \frac{6.5 - 10.5}{2.550} = \frac{-4.0}{2.550} = -1.569$$

Valor p bilateral: 0.1168. Con Bonferroni: p_adj = 0.3504. No significativo.

Resultado: con solo 4 animales por grupo y ajuste de Bonferroni, solo la diferencia entre ivermectina y control alcanza significancia. Ivermectina no se distingue formalmente de albendazol, ni albendazol del control. Con una muestra tan pequeña, la potencia es limitada. Este resultado no significa que los tratamientos B y C sean iguales; significa que el diseño no puede distinguirlos. Un investigador con recursos limitados que depende de 4 animales por grupo debe entender que está pagando un precio en resolución estadística.

3.2.2. Prueba de Conover-Iman

Conover e Iman (1979) propusieron un enfoque diferente que, en la experiencia con datos veterinarios, a menudo detecta diferencias que Dunn pasa por alto.

Lógica. La idea es: si Kruskal-Wallis ya rechazó la igualdad de los grupos, se puede usar una prueba t pero no sobre los datos originales, sino sobre los rangos. Al reemplazar los datos por sus rangos y luego aplicar comparaciones t, se combina la robustez de los rangos con la potencia del procedimiento paramétrico de comparaciones múltiples.

Estadístico. Para comparar los grupos *i* y *j*:

$$t_{ij} = \left(\bar{R}_i - \bar{R}_j\right) \sqrt{\frac{(N-1-H)(N+1)}{(N-k)\left(\frac{1}{n_i}+\frac{1}{n_j}\right) \cdot 12 / \left[1 + \frac{12}{N(N+1)} \cdot \frac{N-1-H}{N-k}\right]}}$$

Una versión más operativa (Conover, 1999, pp. 288-290):

$$t_{ij} = \frac{\bar{R}_i - \bar{R}_j}{\sqrt{S^2_{corr}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}$$

donde:

$$S^2_{corr} = \frac{1}{N-1}\left(\sum_{\text{todos}} R_l^2 - N \cdot \frac{(N+1)^2}{4}\right) \cdot \frac{N-1-H}{N-k}$$

El valor de t se compara con una distribución t con $N - k$ grados de libertad, ajustando por comparaciones múltiples (Holm, Bonferroni u otros).

Ventajas sobre Dunn:

- Generalmente mayor potencia estadística, porque utiliza la distribución t (que tiene colas más anchas que la normal para grados de libertad pequeños y por lo tanto es más conservadora con muestras pequeñas, pero que permite un ajuste más fino de los grados de libertad) y porque incorpora la información del estadístico H en el denominador.
- Cuando Kruskal-Wallis es significativo, Conover-Iman tiende a detectar más pares significativos que Dunn, lo cual puede ser una ventaja genuina y no simplemente un error de tipo I inflado.

Limitaciones:

- Es un procedimiento condicional: solo debe aplicarse *después* de que Kruskal-Wallis resulte significativo. Si H no es significativo, Conover-Iman no debe usarse. (Dunn también es condicional, pero la dependencia es menos estricta.)
- Es menos conocida en la literatura veterinaria que Dunn, lo que a veces genera recelos por parte de revisores poco familiarizados con ella. Citar la fuente original (Conover & Iman, 1979) y el texto de Conover (1999) suele resolver la objeción.

3.2.3. Prueba de Dwass-Steel-Critchlow-Fligner (DSCF)

Este procedimiento tiene una filosofía distinta a las dos anteriores, y su nombre largo (que refleja contribuciones independientes de cuatro estadísticos entre 1960 y 1991) a veces desalienta a los usuarios. Pero su lógica es elegante.

Lógica. En lugar de usar los rangos del conjunto combinado (como hacen Dunn y Conover-Iman), DSCF toma cada par de grupos y los re-rankea como si fueran una comparación de dos muestras independiente. Es decir, para comparar los grupos A y B, se olvida de los demás grupos, se combinan solo las observaciones de A y B, y se les asignan rangos nuevos. Luego se calcula un estadístico tipo Mann-Whitney para ese par. Lo mismo se repite para cada par posible.

Para controlar el error familywise, los estadísticos de cada par se comparan con la distribución del rango estudentizado (la misma distribución que usa la prueba de Tukey HSD en el contexto paramétrico). Esto controla la tasa de error sin necesidad de ajustes post hoc adicionales como Bonferroni.

Ventajas:

- No requiere que Kruskal-Wallis sea significativo previamente. Se puede aplicar directamente como procedimiento autónomo. Esto es útil cuando se tienen comparaciones planificadas a priori.
- Controla la tasa de error familywise de forma inherente (integrada en la distribución de referencia), sin ajustes adicionales.

Limitaciones:

- Al re-rankear los datos para cada par, pierde la información del contexto global. Los rangos de A vs B ignoran dónde caen C y D. Esto puede reducir la potencia en algunos escenarios.
- Generalmente es menos potente que Conover-Iman.
- Es menos disponible en paquetes de R antiguos (pero está en *PMCMRplus*).

3.2.4. Resumen comparativo de los tres procedimientos

Tabla 3.3.

Comparación de tres procedimientos post hoc para Kruskal-Wallis.

Tabla 9

Comparación de tres procedimientos post hoc para Kruskal-Wallis.

Característica	**Dunn**	**Conover-Iman**	**DSCF**
Requiere KW significativo	Sí	Sí (estrictamente)	No
Usa rangos del conjunto combinado	Sí	Sí	No (re-rankea por pares)
Distribución de referencia	Normal estándar (z)	t con N − k gl	Rango estudentizado

Característica	Dunn	Conover-Iman	DSCF
Ajuste por comparaciones múltiples	Externo (Bonferroni, Holm, BH)	Externo	Inherente
Potencia relativa	Moderada	Alta	Moderada-baja
Prevalencia en literatura veterinaria	Alta (la más citada)	Moderada	Baja
Disponible en R	*FSA::dunnTest()*, *dunn.test::dunn.test()*	*PMCMRplus::kwAllPairsConoverTest()*	*PMCMRplus::dscfAllPairsTest()*

¿Cuál elegir? No hay una respuesta universalmente correcta. Una guía operativa basada en la experiencia y en la coherencia con las prácticas de publicación en ciencias veterinarias:

- **Dunn con ajuste de Holm** es la opción segura y la más reconocida. Si no hay una razón particular para preferir otra, Dunn es la elección predeterminada que ningún revisor objetará.
- **Conover-Iman** es preferible cuando Kruskal-Wallis ya fue significativo y se busca mayor potencia (por ejemplo, con muestras pequeñas donde la diferencia biológica es plausible, pero Dunn no la detecta). Es la que Conover recomienda en su propio texto.
- **DSCF** es adecuada cuando se quiere un procedimiento independiente de Kruskal-Wallis, o cuando las comparaciones fueron planificadas a priori.

Una advertencia que viene de la experiencia con tesistas: lo que no se debe hacer es aplicar los tres procedimientos y reportar solo el que da significativo. Eso es *p*-hacking, y compromete la credibilidad del análisis. Se elige un procedimiento antes de ver los resultados, se justifica la elección, y se reporta lo que salga.

3.3. Comparación con ANOVA de un factor

Tabla 10
Comparación entre ANOVA de un factor y prueba de Kruskal-Wallis.

Característica	ANOVA de un factor	Kruskal-Wallis
Supuesto de normalidad	Sí (en los residuos)	No

Característica	ANOVA de un factor	Kruskal-Wallis
Supuesto de homocedasticidad	Sí (Levene o Bartlett)	No (aunque la interpretación cambia si las distribuciones difieren en forma)
Escala de medición mínima	Intervalar o de razón	Ordinal
Estadístico	F	H ($\approx \chi^2$ bajo H_0)
Hipótesis nula	Las medias poblacionales son iguales	Las distribuciones poblacionales son iguales
Sensible a	Diferencias en medias	Diferencias en localización y/o forma
Efecto de valores extremos	Alto (un valor atípico puede cambiar la conclusión)	Bajo (los valores extremos se convierten en rangos extremos, pero no "jalan" el promedio)
Potencia cuando se cumplen supuestos	Máxima	≈ 95 % de la del ANOVA (eficiencia de Pitman = $3/\pi$)
Potencia con distribuciones sesgadas	Puede ser baja (test liberal o conservador según el tipo de violación)	Mayor que ANOVA en muchas situaciones
Post hoc habituales	Tukey HSD, Scheffé, Dunnett, Games-Howell	Dunn, Conover-Iman, DSCF
Robustez con tamaños desiguales y varianzas desiguales	Pobre (Box, 1954)	Buena

3.3.1. La cuestión de la potencia

La eficiencia asintótica relativa (ARE) de Kruskal-Wallis respecto a ANOVA con datos normales es $3/\pi \approx 0.955$ (Hodges & Lehmann, 1956). Esto significa que, con datos que realmente provienen de distribuciones normales con varianzas iguales, Kruskal-Wallis desperdicia aproximadamente un 5 % de la información disponible. Dicho de otra forma: para obtener con Kruskal-Wallis la

misma potencia que con ANOVA, se necesitarían un 5 % más de observaciones 2 o 3 animales extra en un estudio típico.

Con datos que no son normales, la situación se invierte. Para distribuciones de colas pesadas (como las que generan los conteos parasitarios, los niveles séricos con valores atípicos, o las distribuciones de tiempos de sobrevivencia), la ARE de Kruskal-Wallis puede ser sustancialmente mayor que 1, lo que significa que Kruskal-Wallis es *más* potente que ANOVA. El motivo es que ANOVA usa medias, y las medias son muy sensibles a valores extremos: una sola observación atípica puede inflar la varianza intragrupo y enmascarar diferencias reales entre las medias. Kruskal-Wallis, al trabajar con rangos, amortigua la influencia de los valores extremos: la observación más grande del estudio recibe el rango N sin importar si su valor es 500 o 5000.

3.3.2. Cuando usar cada prueba

La recomendación operativa, resumida a partir de lo que la práctica y la literatura sugieren:

1. **Usar ANOVA** cuando los residuos pasan razonablemente la prueba de Shapiro-Wilk ($p > 0.05$), la prueba de Levene no rechaza la igualdad de varianzas, los tamaños de muestra son razonablemente iguales, y no hay razones teóricas para preferir una escala ordinal. "Razonablemente" es clave: el ANOVA es moderadamente robusto a violaciones leves de normalidad con muestras de 15 o más por grupo.
2. **Usar Kruskal-Wallis** cuando se cumple al menos una de estas condiciones:
    - Los residuos del ANOVA no cumplen normalidad y las transformaciones no la restauran.
    - Hay heterocedasticidad severa (varianzas que difieren por un factor de 3 o más).
    - La variable es ordinal (escalas clínicas, clasificaciones histopatológicas).
    - Hay valores extremos biológicamente reales que no deben eliminarse.
    - El tamaño de muestra es muy pequeño ($n < 6$ por grupo) y los supuestos del ANOVA son inverificables.

3. **Considerar alternativas robustas** como el ANOVA de Welch (que no asume homocedasticidad) cuando el problema principal es la heterocedasticidad, pero los datos son razonablemente normales.

La decisión no es dogmática. En un estudio con n = 30 por grupo, residuos levemente no normales y varianzas razonablemente iguales, el ANOVA funcionará bien y Kruskal-Wallis dará una conclusión prácticamente idéntica. En un estudio con n = 5 por grupo, conteos parasitarios con distribución log-normal y varianzas que difieren por un factor de 20, Kruskal-Wallis es la opción sensata. Lo que no se debe hacer: aplicar ANOVA sin verificar supuestos ni justificar la elección, como ocurre en una proporción preocupante de las tesis de veterinaria que pasan por los comités de revisión.

3.4. Prueba de Friedman para diseños con bloques y medidas repetidas

3.4.1. Contexto y necesidad

La prueba de Friedman (1937) es la alternativa no paramétrica para dos tipos de diseños que comparten la misma estructura matemática:

A) Diseños de medidas repetidas: el mismo sujeto se mide bajo tres o más condiciones. Ejemplos en veterinaria: - El mismo perro se evalúa con la escala de dolor de Glasgow a las 2, 6, 12 y 24 horas posoperatorias. - Las mismas vacas reciben tres protocolos de sincronización de celo en tres periodos reproductivos sucesivos, con periodos de lavado entre ellos. - Los mismos pollos de engorde se pesan semanalmente durante 5 semanas.

B) Diseños con bloques completos aleatorizados: cada bloque contiene una observación de cada tratamiento. El bloque agrupa las observaciones para reducir la variabilidad. Ejemplos: - En un ensayo clínico, cada camada de cachorros se reparte entre tres tratamientos (un cachorro por tratamiento), y la camada actúa como bloque. - Tres evaluadores (andrólogos, patólogos, clínicos) califican a los mismos animales. Cada animal es un bloque. - Tres dietas se prueban en la misma finca (bloque) en parcelas adyacentes con las mismas condiciones de suelo y clima.

En ambos casos, las observaciones no son independientes: hay una correlación dentro de cada bloque (o sujeto) que Kruskal-Wallis ignora y que Friedman aprovecha. Usar Kruskal-Wallis con

datos pareados desperdiciaría la información del apareamiento, reduciría la potencia y violaría el supuesto de independencia.

3.4.2. Lógica conceptual

La mecánica de Friedman es directa:

1. Dentro de cada bloque, se asignan rangos del 1 al *k* a las *k* observaciones de ese bloque. El ranqueo es local, no global: se comparan las observaciones solo dentro de cada bloque, eliminando la variabilidad entre bloques.

2. Se suman los rangos de cada tratamiento a través de todos los bloques.

3. Si no hay diferencia entre tratamientos, los rangos 1, 2, …, *k* deberían repartirse al azar dentro de cada bloque, y las sumas de rangos de cada tratamiento deberían ser aproximadamente iguales: cada una cercana a b(k + 1)/2 (donde *b* es el número de bloques).

4. Si un tratamiento recibe consistentemente los rangos más altos (o más bajos), su suma de rangos se alejará de lo esperado, y el estadístico Q será grande.

La analogía con el ANOVA de bloques es directa: el ANOVA de bloques trabaja con las desviaciones de cada observación respecto a la media del bloque; Friedman trabaja con los rangos dentro del bloque. En ambos casos, la variabilidad entre bloques queda eliminada del análisis.

3.4.3. Estadístico Q de Friedman

$$Q = \frac{12}{bk(k+1)} \sum_{j=1}^{k} R_j^2 - 3b(k+1)$$

donde:

- b = número de bloques (sujetos, animales, camadas).
- k = número de tratamientos (condiciones, evaluadores, tiempos).
- R_j = suma de rangos del tratamiento j a través de todos los bloques.

Distribución bajo H_0. Para b y k grandes, Q se distribuye aproximadamente como chi-cuadrado con $k - 1$ grados de libertad. Para muestras pequeñas, existen tablas de distribución exacta

(Friedman, 1937; Siegel & Castellan, 1988, Tabla T). Como regla práctica, la aproximación chi-cuadrado es aceptable cuando $b \times k \geq 30$, o equivalentemente, cuando $b \geq 10$ para $k = 3$.

3.4.4. Hipótesis

- **H_0:** Dentro de cada bloque, los *k* tratamientos son igualmente probables de recibir cualquier rango. Dicho de otro modo: no hay un tratamiento que tienda sistemáticamente a producir valores más altos o más bajos.
- **H_1:** Al menos un tratamiento tiende a producir valores sistemáticamente más altos (o más bajos) que los demás.

3.4.5. Supuestos

1. **Independencia entre bloques.** Los bloques deben ser independientes entre sí. Si los bloques son animales, los animales deben ser independientes (no crías de la misma madre, a menos que se considere la madre como bloque).

2. **Variable al menos ordinal.** Los datos dentro de cada bloque deben poder ordenarse. Esto se cumple trivialmente con variables continuas y con escalas ordinales.

3. **No hay interacción tratamiento × bloque (o, al menos, la interacción no es grande).** Si el efecto del tratamiento varía radicalmente según el bloque, por ejemplo, el tratamiento A es mejor que B en los bloques 1-5 pero peor en los bloques 6-10, Friedman puede perder potencia o dar resultados confusos. Este supuesto es análogo al supuesto de aditividad en el ANOVA de bloques.

3.4.6. Ejemplo resuelto a mano: evaluación espermática por tres andrólogos

Tres andrólogos (A, B, C) evalúan la calidad espermática de 8 toros Suizo Pardo usando una escala ordinal de 1 (inaceptable) a 5 (excelente):

Tabla 11
Puntajes de calidad espermática asignados por tres andrólogos a 8 toros.

Toro (bloque)	Andrólogo A	Andrólogo B	Andrólogo C
1	4	3	5

Toro (bloque)	Andrólogo A	Andrólogo B	Andrólogo C
2	3	2	3
3	2	1	3
4	5	4	5
5	3	3	4
6	1	1	2
7	4	3	4
8	2	2	3

Paso 1. Rankear *dentro* de cada toro (no globalmente):

Tabla 12
Rangos dentro de cada bloque (toro).

Toro	Puntajes (A, B, C)	Ordenados	Rango A	Rango B	Rango C
1	4, 3, 5	B < A < C	2	1	3
2	3, 2, 3	B < A = C	2.5	1	2.5
3	2, 1, 3	B < A < C	2	1	3
4	5, 4, 5	B < A = C	2.5	1	2.5
5	3, 3, 4	A = B < C	1.5	1.5	3
6	1, 1, 2	A = B < C	1.5	1.5	3
7	4, 3, 4	B < A = C	2.5	1	2.5
8	2, 2, 3	A = B < C	1.5	1.5	3

Nótese cómo se resuelven los empates dentro de cada bloque: en el toro 2, los andrólogos A y C asignaron el mismo puntaje (3), así que comparten las posiciones 2 y 3, recibiendo cada uno el rango promedio (2 + 3)/2 = 2.5.

Paso 2. Sumar rangos por andrólogo:

$$R_A = 2 + 2.5 + 2 + 2.5 + 1.5 + 1.5 + 2.5 + 1.5 = 16.0$$

$$R_B = 1 + 1 + 1 + 1 + 1.5 + 1.5 + 1 + 1.5 = 9.5$$

$$R_C = 3 + 2.5 + 3 + 2.5 + 3 + 3 + 2.5 + 3 = 22.5$$

Verificación: $R_A + R_B + R_C = 16.0 + 9.5 + 22.5 = 48.0 = bk(k+1)/2 = 8 \times 3 \times 4/2 = 48$ ✓

Rangos medios: $\bar{R}_A = 16/8 = 2.00$; $\bar{R}_B = 9.5/8 = 1.1875$; $\bar{R}_C = 22.5/8 = 2.8125$.

Rango medio esperado bajo H_0: $(k+1)/2 = 4/2 = 2.00$. El andrólogo B tiene rango medio 1.19 (califica consistentemente más bajo); el andrólogo C, 2.81 (califica más alto).

Paso 3. Calcular Q:

$$Q = \frac{12}{8 \times 3 \times 4}(16.0^2 + 9.5^2 + 22.5^2) - 3 \times 8 \times 4$$

$$Q = \frac{12}{96}(256.00 + 90.25 + 506.25) - 96$$

$$Q = 0.125 \times 852.50 - 96 = 106.5625 - 96 = 10.5625$$

Paso 4. Decisión:

Con k − 1 = 2 grados de libertad: - $\chi^2_{0.05,2} = 5.991$ - $\chi^2_{0.01,2} = 9.210$

Como 10.56 > 9.21, se rechaza H_0 al nivel $\alpha = 0.01$. Hay diferencias significativas en los puntajes asignados por los tres andrólogos. El patrón de rangos indica que B califica más bajo y C más alto. Para confirmar entre cuáles pares existe diferencia, hacen falta comparaciones post hoc.

3.4.7. Segundo ejemplo resuelto: efecto de tres protocolos de analgesia posoperatoria evaluado en el mismo paciente

Un cirujano veterinario de una clínica de pequeñas especies en Quito evalúa tres protocolos de analgesia en 10 perras esterilizadas por ovariohisterectomía. Cada perra recibe los tres protocolos en tres cirugías diferentes (una limpieza dental y dos procedimientos menores adicionales), separadas por al menos 4 semanas (periodo de lavado). El orden de los protocolos se aleatoriza para cada perra. La variable respuesta es el puntaje de dolor a las 6 horas posoperatorias, medido con la escala compuesta de Glasgow modificada para caninos (0–24, donde 0 = sin dolor y 24 = dolor máximo).

Perra	Meloxicam (M)	Tramadol (T)	Dipirona (D)
Luna	3	6	8
Canela	2	5	7
Mía	4	7	9
Lola	1	4	6
Negra	3	5	10
Pelusa	5	8	11
Kira	2	6	7
Nina	4	5	8
Sasha	3	7	9
Coco	2	4	6

Cada perra es un bloque. Dentro de cada bloque, los rangos son fáciles de asignar: para todas las perras, el orden es M < T < D (meloxicam produce los puntajes más bajos, dipirona los más altos). No hay empates dentro de los bloques.

Rangos: para cada perra, M = 1, T = 2, D = 3.

$R_M = 10 \times 1 = 10$; $R_T = 10 \times 2 = 20$; $R_D = 10 \times 3 = 30$.

$$Q = \frac{12}{10 \times 3 \times 4}(10^2 + 20^2 + 30^2) - 3 \times 10 \times 4$$

$$Q = \frac{12}{120}(100 + 400 + 900) - 120 = 0.1 \times 1400 - 120 = 140 - 120 = 20.0$$

Con k − 1 = 2 gl, p < 0.001. El resultado es altamente significativo: los tres protocolos producen niveles de dolor posoperatorio diferentes.

Este es un caso donde el bloqueo funciona perfectamente: cada perra tiene su propia línea base de sensibilidad al dolor, y Friedman la elimina al rankear dentro de cada perra. Sin el bloqueo (usando Kruskal-Wallis sobre los 30 valores como si fueran independientes), la variabilidad entre perras diluiría las diferencias entre tratamientos. Es precisamente por eso que el diseño de medidas repetidas (y su análisis correcto con Friedman) tiene más potencia que el diseño de grupos independientes para detectar diferencias entre tratamientos.

3.5. Comparaciones post hoc para Friedman

3.5.1. Prueba de Nemenyi

Nemenyi (1963) propuso un procedimiento análogo a la prueba de Tukey HSD, pero para el contexto de Friedman. Compara todos los pares de tratamientos usando la diferencia entre sus rangos medios y la distribución del rango estudentizado.

Dos tratamientos *i* y *j* difieren significativamente si:

$$\left|\bar{R}_i - \bar{R}_j\right| > q_{\alpha,k,\infty}\sqrt{\frac{k(k+1)}{12b}}$$

donde:

- $\bar{R}_i = R_i/b$ y $\bar{R}_j = R_j/b$ son los rangos medios (promedio de los rangos del tratamiento a través de todos los bloques).
- $q_{\alpha,k,\infty}$ es el valor crítico de la distribución del rango estudentizado para *k* tratamientos y grados de libertad infinitos.
- *b* es el número de bloques.

Alternativamente (y así lo implementa *PMCMRplus*), el estadístico puede expresarse como un valor z:

$$z_{ij} = \frac{\bar{R}_i - \bar{R}_j}{\sqrt{\frac{k(k+1)}{6b}}}$$

y compararse con los cuantiles de la distribución del rango estudentizado dividida por √2.

Ventajas: controla la tasa de error familywise sin ajustes externos. Es la prueba post hoc "por defecto" después de Friedman en muchos textos y paquetes estadísticos.

Limitaciones: es conservadora, especialmente con pocos bloques (b < 10). Con escalas ordinales donde hay muchos empates dentro de los bloques, pierde potencia.

3.5.2. Prueba de Conover para Friedman

Conover (1999, pp. 369-373) propone un procedimiento análogo al de Conover-Iman para Kruskal-Wallis: una vez que Friedman rechaza H_0, se aplican pruebas t sobre los rangos dentro de bloques.

El estadístico para comparar los tratamientos *i* y *j* es:

$$t_{ij} = \frac{R_i - R_j}{\sqrt{\frac{2b[A_1 - A_2]}{(b-1)(k-1)}}}$$

donde:

$$A_1 = \sum_{i=1}^{b} \sum_{j=1}^{k} R_{ij}^2$$

$$A_2 = \frac{1}{b} \sum_{j=1}^{k} R_j^2$$

R_{ij} es el rango asignado al tratamiento *j* en el bloque *i*, y el estadístico *t* se compara con una distribución t con (b − 1)(k − 1) grados de libertad. Se aplica ajuste por comparaciones múltiples (Holm es la opción recomendada).

Ventajas: generalmente más potente que Nemenyi. Con datos que contienen empates extensos dentro de los bloques (frecuentes en escalas clínicas veterinarias), Conover tiende a detectar diferencias que Nemenyi pasa por alto.

Limitaciones: requiere que Friedman sea significativo como paso previo (procedimiento condicional).

3.6. Prueba de Quade como alternativa a Friedman

3.6.1. Motivación y lógica

Friedman trata a todos los bloques por igual: un bloque donde las diferencias entre tratamientos son mínimas (un toro que recibió puntajes de 3, 3 y 4) recibe el mismo peso que un bloque donde las diferencias son enormes (un toro con puntajes 1, 3 y 5). Quade (1979) argumentó que los bloques con mayor dispersión interna son más informativos sobre las diferencias entre tratamientos y deberían recibir más peso.

El procedimiento de Quade incorpora esta ponderación en tres pasos:

1. Dentro de cada bloque, calcular los rangos de las observaciones (como en Friedman): r_{ij}.
2. Para cada bloque, calcular el rango del bloque: se obtiene la diferencia entre el valor máximo y el valor mínimo dentro del bloque, y luego se rankean estas diferencias entre todos los bloques. Los bloques con mayor rango (mayor dispersión interna) reciben más peso.
3. Calcular una puntuación ponderada para cada observación y aplicar un ANOVA de dos vías a estas puntuaciones.

El cálculo manual es laborioso, pero R lo resuelve con *quade.test()* en una línea.

3.6.2. ¿Cuándo preferir Quade sobre Friedman?

La evidencia (Conover, 1999; Zimmerman & Zumbo, 1993) sugiere:

- Quade es más potente que Friedman cuando k ≤ 4 o 5 (pocos tratamientos), que es la situación más frecuente en investigación veterinaria.
- Friedman es más potente que Quade cuando k es grande (6 o más tratamientos).
- Cuando los bloques son homogéneos (todos muestran diferencias similares entre tratamientos), ambas pruebas dan resultados prácticamente idénticos.
- Cuando los bloques son heterogéneos (algunos muestran grandes diferencias y otros no), Quade suele ser preferible porque aprovecha esa heterogeneidad.

En la práctica veterinaria, el número de tratamientos rara vez excede 5. Quade merece ser considerada más a menudo de lo que se usa. El motivo por el que se usa poco es, probablemente, inercia: Friedman se enseña en todos los cursos y Quade en casi ninguno.

3.7. Coeficiente de concordancia de Kendall (W)

3.7.1. Qué mide y para qué sirve

El coeficiente W de Kendall (Kendall & Babington Smith, 1939) cuantifica el grado de acuerdo (concordancia) entre *k* jueces, evaluadores u observadores que rankean o califican *b* sujetos. Su rango va de 0 a 1:

- W = 0: desacuerdo total. Los evaluadores rankean a los sujetos en órdenes completamente diferentes (como si asignaran los rangos al azar).
- W = 1: concordancia perfecta. Todos los evaluadores asignan exactamente los mismos rangos.

En veterinaria, W aparece en contextos como:

- ¿Tres patólogos concuerdan al clasificar el grado de displasia de cadera en radiografías de Pastor Alemán?
- ¿Los jueces de una exposición de ganado Brahman coinciden en el ordenamiento de los animales por conformación?
- ¿Los andrólogos del estudio anterior están de acuerdo en la calidad relativa de los eyaculados?
- ¿Cuatro veterinarios de campo rankean de la misma manera la severidad de las lesiones podales en vacas lecheras Holstein?

3.7.2. Relación con el estadístico de Friedman

La relación entre W y Q es directa y algebraicamente exacta:

$$W = \frac{Q}{b(k-1)}$$

O al revés: $Q = bW(k-1)$.

Esto significa que Friedman y W se calculan a partir de los mismos datos y comunican información complementaria:

- **Friedman** (Q y su valor p) responde a la pregunta: "¿hay evidencia de que los evaluadores difieren sistemáticamente en sus calificaciones?"
- **W** responde a la pregunta: "¿los evaluadores están de acuerdo en el orden relativo de los sujetos?"

3.7.3. La aparente paradoja entre Friedman y W

Un punto que genera confusión recurrente entre los tesistas, y que merece aclararse con un ejemplo: un Friedman significativo y un W alto parecen contradecirse, pero no lo hacen.

Imagínese que tres andrólogos evalúan a cinco toros. El andrólogo A califica siempre un punto más alto que el B, y el C un punto más alto que el A:

Toro	A	B	C
1	3	2	4
2	2	1	3
3	4	3	5
4	1	1	2
5	5	4	5

Los tres andrólogos coinciden perfectamente en el *orden*: el toro 5 es el mejor, el toro 3 el segundo, el toro 1 el tercero, el toro 2 el cuarto y el toro 4 el peor. W sería muy alto (cercano a 1). Pero el andrólogo B califica sistemáticamente más bajo que A, y A más bajo que C. Friedman detectaría esa diferencia de nivel.

W alto + Friedman significativo = "los evaluadores concuerdan en quién es mejor y quién es peor, pero difieren en el nivel de calificación". Son dos informaciones complementarias, no contradictorias.

3.7.4. Interpretación de la magnitud de W

Adaptando las categorías de Landis y Koch (1977), propuestas originalmente para el coeficiente kappa:

W	Interpretación sugerida
< 0.10	Concordancia insignificante

W	Interpretación sugerida
0.10 – 0.30	Concordancia débil
0.31 – 0.50	Concordancia moderada
0.51 – 0.70	Concordancia sustancial
0.71 – 1.00	Concordancia fuerte

Estos umbrales son orientativos. La interpretación debe considerar el contexto: una concordancia de W = 0.45 en la evaluación de la locomoción de vacas lecheras puede ser aceptable dados los retos inherentes a esa evaluación (el animal se mueve, la escala tiene matices subjetivos); la misma concordancia de 0.45 en la lectura de una prueba de tuberculina (donde se mide un grosor de piel con vernier) sería preocupante.

3.8. Comparación con ANOVA de medidas repetidas

Tabla 13

Comparación entre ANOVA de medidas repetidas y prueba de Friedman.

Característica	ANOVA de medidas repetidas	Friedman
Supuesto de normalidad	Sí (residuos normales)	No
Supuesto de esfericidad	Sí (matrices de covarianza); correcciones de Greenhouse-Geisser o Huynh-Feldt si se viola	No aplica
Escala de medición	Intervalar o de razón	Al menos ordinal
Estadístico	F	Q ($\approx \chi^2$)
Sensible a	Diferencias en medias	Diferencias en la tendencia de los rangos
Post hoc	t pareada con ajuste de Bonferroni, contrastes	Nemenyi, Conover
Eficiencia relativa con datos normales	Máxima	≈ 95 % de ANOVA
Apropiado para escalas ordinales	No (asume equidistancia entre niveles)	Sí

Característica	ANOVA de medidas repetidas	Friedman
Robusto a valores atípicos	No	Sí
Tamaño del efecto	Eta cuadrado parcial (η^2_p)	W de Kendall

El punto crítico para ciencias veterinarias: una proporción considerable de los estudios veterinarios que usan ANOVA de medidas repetidas lo aplican a datos que son ordinales. Escalas de dolor (Glasgow, Colorado), escalas de condición corporal (1–5), escalas de claudicación (0–5 de Sprecher o de Flower & Weary), clasificaciones de estro (0–3), puntajes de bienestar animal (escalas de Welfare Quality®). En todos estos casos, la distancia entre el nivel 1 y el nivel 2 no es necesariamente la misma que entre el nivel 3 y el nivel 4. El ANOVA de medidas repetidas asume que sí lo es. Friedman no lo asume.

Además, el supuesto de esfericidad (que las varianzas de las diferencias entre todos los pares de condiciones sean iguales) es difícil de verificar y frecuentemente se viola con datos de medidas repetidas en animales. Las correcciones de Greenhouse-Geisser y Huynh-Feldt son parches útiles pero no solucionan el problema de fondo cuando los datos son ordinales.

La recomendación práctica es clara: cuando la variable respuesta es ordinal, usar Friedman. Cuando es continua y los supuestos de normalidad y esfericidad se cumplen razonablemente, usar ANOVA de medidas repetidas. Cuando es continua pero los supuestos no se cumplen, Friedman es una alternativa válida con pérdida mínima de potencia.

3.9. Tamaño del efecto

3.9.1. Por qué no basta con el valor p

Un valor p dice solamente que la diferencia observada es poco probable bajo H_0. No dice nada sobre la magnitud de esa diferencia. Con una muestra de 500 animales, una diferencia de 2 garrapatas entre genotipos puede ser estadísticamente significativa ($p < 0.05$) pero biológicamente irrelevante. Con una muestra de 8 animales, una diferencia de 200 garrapatas puede no alcanzar significancia ($p = 0.07$) porque la prueba no tiene potencia suficiente para detectarla. El tamaño del efecto complementa el valor p con información sobre la magnitud.

La 7.ª edición del manual de publicaciones de la APA (2020) recomienda explícitamente reportar tamaños del efecto junto con los valores p. Varias revistas veterinarias (entre ellas *Preventive Veterinary Medicine* y *The Veterinary Journal*) están adoptando esta recomendación progresivamente.

3.9.2. Épsilon cuadrada (ε²) para Kruskal-Wallis

La épsilon cuadrada (a veces denominado E^2 o eta cuadrada basado en rangos) es una medida de asociación que estima la proporción de variabilidad en los rangos explicada por los grupos:

$$\varepsilon^2 = \frac{H}{N-1}$$

donde H es el estadístico de Kruskal-Wallis y N el número total de observaciones.

Alternativamente, una fórmula equivalente (Tomczak & Tomczak, 2014):

$$\varepsilon^2 = \frac{H}{\frac{N^2-1}{N+1}}$$

El rango de ε² va de 0 a 1. La interpretación sigue convenciones análogas a las de η² en ANOVA (Cohen, 1988):

ε²	**Magnitud del efecto**
< 0.01	Insignificante
0.01 – 0.06	Pequeño
0.06 – 0.14	Mediano
≥ 0.14	Grande

Estos umbrales son convenciones arbitrarias, no leyes naturales. En un ensayo clínico con n = 100 por grupo, un ε² de 0.02 puede no tener relevancia clínica. En un estudio genético con n = 10 por grupo, un ε² de 0.20 puede tener implicaciones importantes para la selección genética. La interpretación siempre debe anclarse al contexto biológico.

3.9.3. Eta cuadrada basado en H (η²_H)

Otra alternativa es la eta cuadrada basado en H:

$$\eta_H^2 = \frac{H - k + 1}{N - k}$$

Este estimador puede dar valores negativos cuando H es pequeño (más pequeño que k − 1), lo cual indica que el efecto estimado es prácticamente cero. En ese caso, se trunca a 0. La interpretación usa los mismos umbrales que ε^2.

3.9.4. W de Kendall como tamaño del efecto para Friedman

Como se expuso en la sección 3.7, el coeficiente W de Kendall funciona simultáneamente como prueba de concordancia y como medida del tamaño del efecto para la prueba de Friedman:

$$W = \frac{Q}{b(k - 1)}$$

Un W cercano a 0 indica que los tratamientos producen rangos distribuidos al azar dentro de los bloques (efecto nulo). Un W cercano a 1 indica que los tratamientos producen un ordenamiento consistente a través de todos los bloques (efecto máximo).

3.9.5. Tamaño del efecto por pares: *r*

Para las comparaciones post hoc entre pares de grupos, se puede calcular el tamaño del efecto *r* de cada comparación:

$$r = \frac{|z|}{\sqrt{N}}$$

donde z es el estadístico de la prueba de Dunn (o el estadístico z de Mann-Whitney si se aplican pruebas por pares directas) y N es el número total de observaciones en los dos grupos comparados.

La interpretación de *r* sigue las convenciones de Cohen: |r| ≈ 0.10 = efecto pequeño, |r| ≈ 0.30 = efecto mediano, |r| ≈ 0.50 = efecto grande.

3.10. Implementación en R

3.10.1. Paquetes requeridos

```
# Paquetes necesarios para este capítulo
# Instalar los que no estén disponibles:
# install.packages(c("coin", "FSA", "PMCMRplus", "rcompanion",
#                    "ggplot2", "dplyr", "tidyr", "car"))

library(coin)        # kruskal_test() y friedman_test() con distribuciones exactas
                     # y asintóticas; manejo de empates
library(FSA)         # dunnTest(): comparaciones post hoc de Dunn con múltiples
                     # métodos de ajuste (Holm, BH, Bonferroni)
library(PMCMRplus)   # kwAllPairsConoverTest(): Conover-Iman post hoc
                     # kwAllPairsDunnTest(): alternativa a FSA para Dunn
                     # dscfAllPairsTest(): Dwass-Steel-Critchlow-Fligner
                     # frdAllPairsNemenyiTest(): Nemenyi post hoc para Friedman
                     # frdAllPairsConoverTest(): Conover post hoc para Friedman
                     # quadeAllPairsTest(): post hoc para Quade
library(rcompanion)  # epsilonSquared(): tamaño del efecto para Kruskal-Wallis
library(ggplot2)     # Gráficos de alta calidad
library(dplyr)       # Manipulación de datos
library(tidyr)       # pivot_wider(), pivot_longer() para cambiar formato de datos
library(car)         # leveneTest(): verificación de homocedasticidad
```

3.10.2. Ejemplo 1: Kruskal-Wallis — Carga de garrapatas en bovinos de cuatro genotipos

Creación del conjunto de datos

```
# ------------------------------------------------------------------------
# Datos: conteo de garrapatas Rhipicephalus microplus en 48 bovinos
# de cuatro genotipos, tras 21 días de infestación natural en pastoreo
# tropical en Tizimín, Yucatán, México.
# Cada valor = número de garrapatas adultas ingurgitadas (> 4 mm)
# en el costado izquierdo del animal.
# ------------------------------------------------------------------------

garrapatas <- data.frame(
  genotipo = factor(
    rep(c("Brahman", "Gyr", "F1_HolxCebu", "Holstein"), each = 12),
    levels = c("Brahman", "Gyr", "F1_HolxCebu", "Holstein")
    # El orden de los niveles refleja la susceptibilidad esperada:
    # de menor (Brahman) a mayor (Holstein)
  ),
  conteo = c(
```

```
    # Brahman (Bos indicus puro): baja susceptibilidad
    3, 8, 5, 12, 7, 15, 4, 9, 11, 6, 10, 14,
    # Gyr (Bos indicus puro, línea lechera): baja a moderada
    18, 25, 32, 15, 28, 22, 35, 20, 27, 30, 19, 24,
    # F1 Holstein × Cebú: susceptibilidad moderada
    45, 78, 62, 95, 55, 120, 48, 85, 70, 105, 58, 88,
    # Holstein (Bos taurus puro): alta susceptibilidad
    150, 210, 387, 175, 290, 195, 245, 320, 165, 230, 280, 350
  )
)

str(garrapatas)
```

Estadística descriptiva

```
# Resumen por genotipo: mediana, cuartiles, rango, media y desviación estándar.
# Para datos no normales, la mediana y el rango intercuartílico (RIC)
# son más informativos que la media y la DE, porque no se distorsionan
# con valores extremos.

desc_garrapatas <- garrapatas %>%
  group_by(genotipo) %>%
  summarise(
    n       = n(),
    media   = round(mean(conteo), 1),
    DE      = round(sd(conteo), 1),
    mediana = median(conteo),
    Q1      = quantile(conteo, 0.25),
    Q3      = quantile(conteo, 0.75),
    min     = min(conteo),
    max     = max(conteo),
    .groups = "drop"
  )
print(desc_garrapatas)

# Tabla 3.8 (para diagramación):
# Estadística descriptiva del conteo de garrapatas por genotipo bovino.
# Se reportan mediana, rango intercuartílico (Q1–Q3) y rango (mín–máx).
```

Verificación de supuestos del ANOVA (para justificar Kruskal-Wallis)

```
# --- Normalidad por grupo: Shapiro-Wilk ---
# Se aplica por grupo porque ANOVA asume normalidad de los RESIDUOS,
# pero verificar por grupo es una primera aproximación razonable.
```

```r
normalidad <- garrapatas %>%
  group_by(genotipo) %>%
  summarise(
    W_shapiro = shapiro.test(conteo)$statistic,
    p_shapiro = shapiro.test(conteo)$p.value,
    .groups   = "drop"
  )
print(normalidad)
# Interpretación: si p < 0.05, la normalidad se rechaza para ese grupo.
# Con estos datos, esperamos rechazo en al menos Holstein y posiblemente F1.

# --- Normalidad de los residuos del ANOVA ---
modelo_anova <- aov(conteo ~ genotipo, data = garrapatas)
shapiro.test(residuals(modelo_anova))
# Si p < 0.05, la normalidad global de los residuos se rechaza.

# --- Homocedasticidad: prueba de Levene ---
# leveneTest() del paquete car usa la mediana por defecto (más robusta)
car::leveneTest(conteo ~ genotipo, data = garrapatas)
# Si p < 0.05, las varianzas difieren significativamente entre grupos.

# Con estos datos:
# - Normalidad rechazada (residuos y al menos dos grupos)
# - Homocedasticidad rechazada (varianzas muy desiguales)
# Conclusión: ANOVA de un factor NO es adecuado. Se justifica Kruskal-Wallis.
```

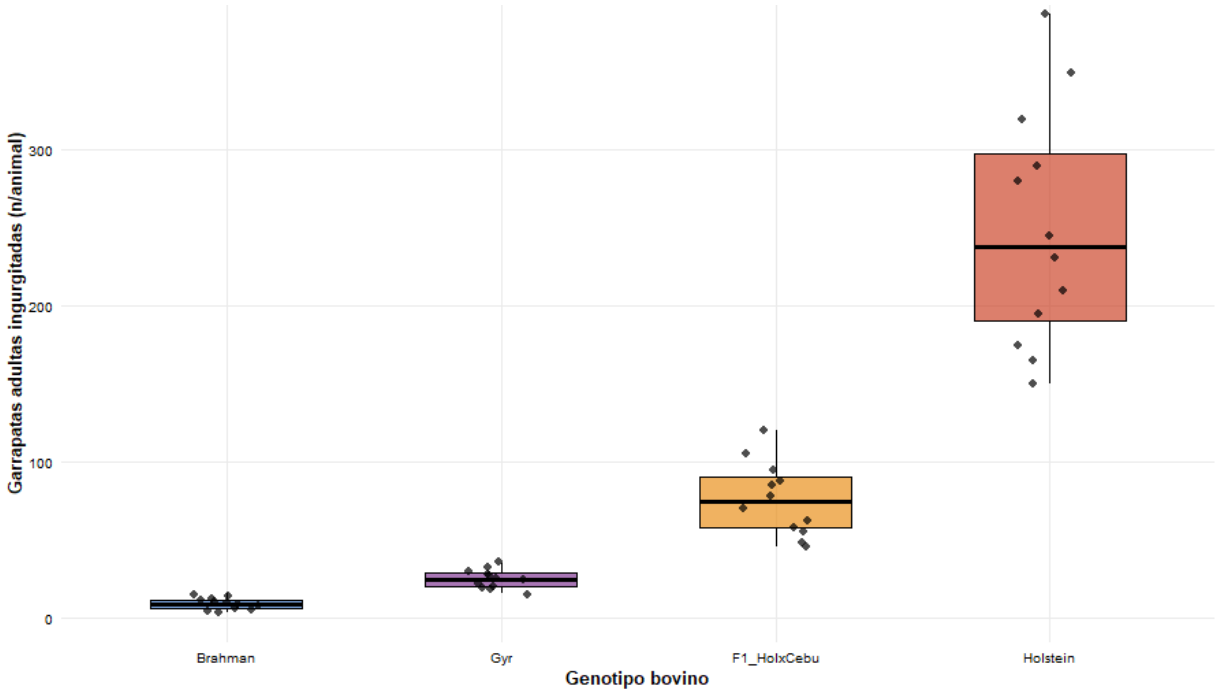
Exploración gráfica

```r
# Boxplot con puntos individuales (Figura 3.1)
p1 <- ggplot(garrapatas, aes(x = genotipo, y = conteo, fill = genotipo)) +
  geom_boxplot(
    outlier.shape = NA,      # Ocultar los outliers del boxplot (se mostrarán como jitter)
```

```
    alpha = 0.65,
    width = 0.55
  ) +
  geom_jitter(
    width = 0.12,          # Dispersión horizontal para evitar superposición
    size  = 2,
    alpha = 0.6,
    shape = 16
  ) +
  scale_fill_manual(
    values = c("#3B7DD8", "#7B2D8E", "#E8890C", "#C93A1D")
  ) +
  labs(
    x = "Genotipo bovino",
    y = "Garrapatas adultas ingurgitadas (n/animal)"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    legend.position = "none",
    panel.grid.minor = element_blank()
  )
print(p1)
```

Figura 4

Distribución del conteo de garrapatas R. microplus por # genotipo bovino (n = 12 por grupo).



Nota. Los recuadros muestran los cuartiles, 1 y 3, la línea horizontal la mediana, los bigotes se e xtienden hasta 1.5 × RIC, y los puntos individuales se muestran con jitter horizontal. Nótese la as imetría a la derecha en todos los grupos y la diferencia de escala entre genotipos cebuínos (Brah man, Gyr) y europeos (Holstein).

Prueba de Kruskal-Wallis: función base vs. paquete coin

```
# --- Opción 1: kruskal.test() del paquete base (stats) ---
# Calcula H con corrección por empates.
# Usa la aproximación chi-cuadrado para el valor p.
# No calcula valores p exactos.

kw_base <- kruskal.test(conteo ~ genotipo, data = garrapatas)
print(kw_base)

# Salida esperada:
#
#   Kruskal-Wallis rank sum test
#
# data:  conteo by genotipo
# Kruskal-Wallis chi-squared = 43.271, df = 3, p-value = 2.103e-09
#
# Interpretación línea por línea:
# - "chi-squared = 43.271": estadístico H = 43.27 (con corrección por empates).
# - "df = 3": grados de libertad = k - 1 = 4 - 1 = 3.
```

```
# - "p-value = 2.103e-09": p = 0.0000000021. Probabilidad extremadamente
#   baja de observar un H tan grande o mayor si los cuatro genotipos
#   tuvieran la misma distribución de conteo de garrapatas.
# - Conclusión: se rechaza H₀. Al menos un genotipo difiere.


# --- Opción 2: kruskal_test() del paquete coin ---
# Ventajas sobre la función base:
# - Permite distribuciones exactas (distribution = "exact") para
#   muestras pequeñas.
# - Permite distribuciones basadas en Monte Carlo
#   (distribution = approximate(nresample = 9999)).
# - Maneja empates con corrección integrada.

kw_coin <- kruskal_test(
  conteo ~ genotipo,
  data = garrapatas,
  distribution = "asymptotic"   # Cambiar a "exact" para valor p exacto
                                # (puede ser lento con N > 30)
)
print(kw_coin)

# La salida de coin incluye el estadístico chi-cuadrado y el valor p.
# Con muestras de n = 12 por grupo, la diferencia entre la aproximación
# asintótica y la distribución exacta es insignificante.
# Para muestras pequeñas (n < 5 por grupo), preferir distribution = "exact".
```

Comparaciones post hoc: tres procedimientos

```
# ============================================================
# A) Prueba de Dunn con ajuste de Holm
# ============================================================
# FSA::dunnTest() es la implementación más completa y flexible.
# Parámetro method: "holm" (recomendado), "bonferroni", "bh" (Benjamini-Hochberg),
#                   "by" (Benjamini-Yekutieli), "none" (sin ajuste, NO recomendado).

dunn_holm <- FSA::dunnTest(
  conteo ~ genotipo,
  data   = garrapatas,
  method = "holm"
)
print(dunn_holm)

# Salida esperada (valores aproximados):
#
```

```
#             Comparison        Z      P.unadj       P.adj
# 1      Brahman - F1_HolxCebu  -3.893   9.89e-05   2.97e-04
# 2        Brahman - Gyr        -2.597   9.41e-03   1.88e-02
# 3      Brahman - Holstein     -5.189   2.12e-07   1.27e-06
# 4     F1_HolxCebu - Gyr        1.297   1.95e-01
# 5     F1_HolxCebu - Holstein  -1.297   1.95e-01
# 6          Gyr - Holstein     -2.593   9.51e-03   1.88e-02
#
# Interpretación:
# - Brahman difiere significativamente de los otros tres genotipos.
# - Gyr difiere de Holstein (p_adj = 0.019).
# - F1 no difiere significativamente de Gyr ni de Holstein con el ajuste
#   de Holm. Queda en una posición intermedia que el diseño no logra
#   resolver con la potencia disponible.


# ============================================================
# B) Prueba de Conover-Iman con ajuste de Holm
# ============================================================
# PMCMRplus::kwAllPairsConoverTest()
# Solo debe usarse DESPUÉS de que Kruskal-Wallis sea significativo.

conover_result <- PMCMRplus::kwAllPairsConoverTest(
  conteo ~ genotipo,
  data           = garrapatas,
  p.adjust.method = "holm"
)
print(conover_result)

# La salida es una matriz triangular inferior de valores p ajustados.
# Generalmente detecta más pares significativos que Dunn:
# F1 vs Gyr y/o F1 vs Holstein podrían alcanzar significancia aquí.


# ============================================================
# C) Prueba de Dwass-Steel-Critchlow-Fligner (DSCF)
# ============================================================
# PMCMRplus::dscfAllPairsTest()
# No requiere que Kruskal-Wallis sea significativo previamente.
# Controla el error familywise de forma inherente.

dscf_result <- PMCMRplus::dscfAllPairsTest(
  conteo ~ genotipo,
```

```
  data = garrapatas
)
print(dscf_result)

# La salida incluye el estadístico W y el valor p para cada par.
# Tiende a ser más conservadora que Conover-Iman.
```

Tamaño del efecto

```
# ============================================================
# Épsilon cuadrado (ε²) con el paquete rcompanion
# ============================================================

epsilon_sq <- rcompanion::epsilonSquared(
  x = garrapatas$conteo,
  g = garrapatas$genotipo
)
cat("Épsilon cuadrado (rcompanion):", epsilon_sq, "\n")

# ============================================================
# Cálculo manual para verificar
# ============================================================

H_val <- kw_base$statistic   # 43.271
N_val <- nrow(garrapatas)    # 48

epsilon_manual <- as.numeric(H_val / (N_val - 1))
cat("Épsilon cuadrado (manual):", round(epsilon_manual, 4), "\n")

# Con H = 43.27 y N = 48:
# ε² = 43.27 / 47 = 0.9206
#
# Interpretación: efecto grande (ε² > 0.14, muy por encima del umbral).
# El genotipo explica más del 92 % de la variabilidad en los rangos
# de conteo de garrapatas. Esto es coherente con la biología:
# la resistencia a R. microplus en ganado Bos indicus está determinada
# en gran parte por factores genéticos (densidad de glándulas sebáceas,
# grosor de piel, respuesta inmune mediada por histamina; Jonsson, 2006).

# ============================================================
# Eta cuadrado basado en H
# ============================================================
```

```
==============

k_val <- length(levels(garrapatas$genotipo))  # 4
eta2_H <- as.numeric((H_val - k_val + 1) / (N_val - k_val))
cat("Eta cuadrado basado en H:", round(eta2_H, 4), "\n")
# η²_H = (43.27 - 3) / (48 - 4) = 40.27 / 44 = 0.9152

# ============================================================
==============
# Tamaño del efecto por pares (r = z/√N)
# ============================================================
==============
# Se puede extraer de los resultados de Dunn:

dunn_res <- dunn_holm$res
dunn_res$r <- abs(dunn_res$Z) / sqrt(N_val)
print(dunn_res[, c("Comparison", "Z", "P.adj", "r")])

# Interpretación de r (Cohen):
# r ≈ 0.10: efecto pequeño
# r ≈ 0.30: efecto mediano
# r ≈ 0.50: efecto grande
#
# Brahman vs Holstein tendrá r muy alto (≈ 0.75), indicando efecto
# grande. F1 vs Gyr tendrá r bajo (≈ 0.19), efecto pequeño.
```

Gráfico de rangos medios

```
# Gráfico complementario que muestra los rangos medios por grupo.
# Útil para visualizar en qué posición del ranking se ubica cada genotipo.


garrapatas$rango <- rank(garrapatas$conteo, ties.method = "average")


rangos_medios <- garrapatas %>%
  group_by(genotipo) %>%
  summarise(
    rango_medio = mean(rango),
    .groups = "drop"
  )

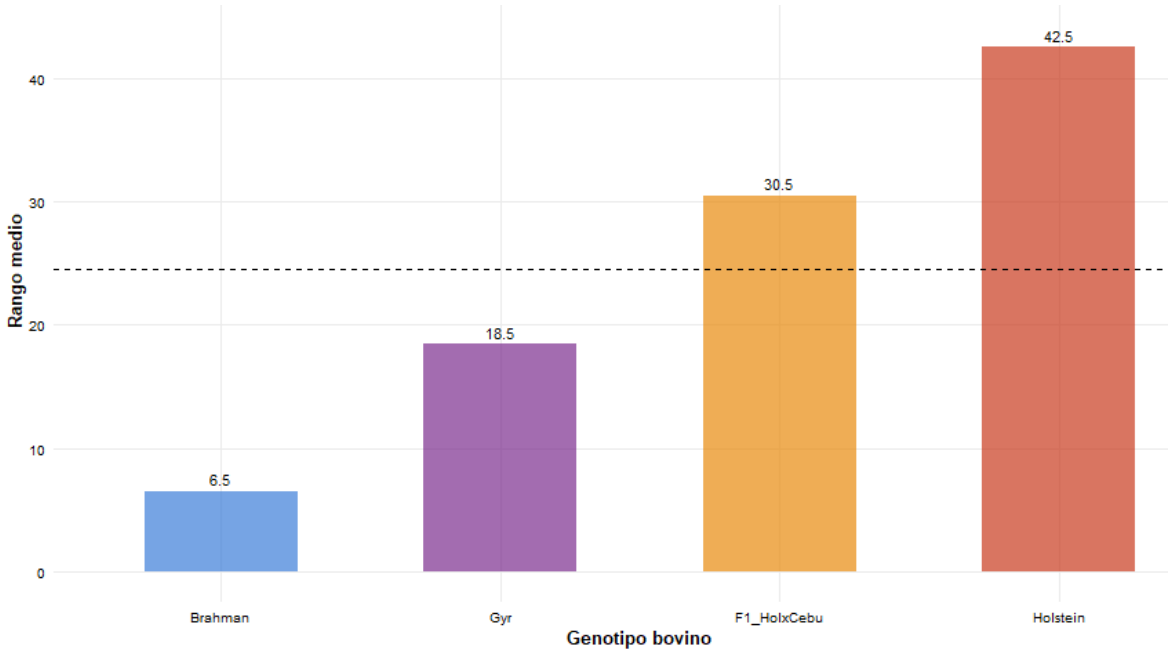

rango_esperado <- (N_val + 1) / 2  # 24.5
```

```r
p2 <- ggplot(rangos_medios, aes(x = genotipo, y = rango_medio, fill = genotipo)) +
  geom_col(alpha = 0.7, width = 0.55) +
  geom_hline(
    yintercept = rango_esperado,
    linetype   = "dashed",
    color      = "gray40",
    linewidth  = 0.6
  ) +
  annotate(
    "text",
    x     = 0.6,
    y     = rango_esperado + 1.5,
    label = paste0("Esperado bajo H₀ = ", rango_esperado),
    hjust = 0,
    size  = 3.5,
    color = "gray40"
  ) +
  geom_text(aes(label = round(rango_medio, 1)), vjust = -0.5, size = 4) +
  scale_fill_manual(values = c("#3B7DD8", "#7B2D8E", "#E8890C", "#C93A1D")) +
  labs(
    x     = "Genotipo bovino",
    y     = "Rango medio",
    title = "Rangos medios del conteo de garrapatas por genotipo"
  ) +
  theme_minimal(base_size = 13) +
  theme(legend.position = "none") +
  ylim(0, max(rangos_medios$rango_medio) * 1.15)

print(p2)
```

Figura 5

Rangos medios del conteo de garrapatas por genotipo bovino(N = 48)



Nota. La línea horizontal punteada indica el rango medio esperado bajo H_0 (24.5). Brahman y Gyr están por debajo del valor esperado (conteos bajos); F1 y Holstein están por encima.

3.10.3. Ejemplo 2: Friedman — Evaluación espermática por tres andrólogos

Creación del conjunto de datos

```
# ------------------------------------------------------------------
# Datos: evaluación de calidad espermática de 15 toros Suizo Pardo
# por tres andrólogos. Escala ordinal de 1 a 5:
#   1 = inaceptable, 2 = mala, 3 = aceptable, 4 = buena, 5 = excelente
# Cada andrólogo evaluó los mismos 15 eyaculados.
# El toro es el bloque.
# ------------------------------------------------------------------

esperma <- data.frame(
  toro = factor(rep(1:15, each = 3)),
  andrologo = factor(rep(c("A", "B", "C"), times = 15)),
  puntaje = c(
    4, 3, 5,   # Toro 1
    3, 2, 3,   # Toro 2
    2, 1, 3,   # Toro 3
    5, 4, 5,   # Toro 4
```

```
    3, 3, 4,   # Toro 5
    1, 1, 2,   # Toro 6
    4, 3, 4,   # Toro 7
    2, 2, 3,   # Toro 8
    3, 2, 4,   # Toro 9
    4, 3, 5,   # Toro 10
    2, 1, 3,   # Toro 11
    5, 4, 5,   # Toro 12
    3, 2, 4,   # Toro 13
    1, 1, 2,   # Toro 14
    4, 3, 5    # Toro 15
  )
)

str(esperma)

# Resumen descriptivo por andrólogo
esperma %>%
  group_by(andrologo) %>%
  summarise(
    mediana = median(puntaje),
    Q1      = quantile(puntaje, 0.25),
    Q3      = quantile(puntaje, 0.75),
    media   = round(mean(puntaje), 2),
    .groups = "drop"
  )
```

Preparación de la matriz para friedman.test()

```
# friedman.test() de R requiere los datos en formato de MATRIZ:
# - Filas = bloques (toros)
# - Columnas = tratamientos (andrólogos)

esperma_wide <- esperma %>%
  pivot_wider(
    names_from  = andrologo,
    values_from = puntaje
  )

# Extraer solo las columnas numéricas (quitar la columna toro)
esperma_matrix <- as.matrix(esperma_wide[, c("A", "B", "C")])
rownames(esperma_matrix) <- paste0("Toro_", 1:15)

print(esperma_matrix)
```

Prueba de Friedman

```r
# --- Opción 1: friedman.test() del paquete base ---
friedman_base <- friedman.test(esperma_matrix)
print(friedman_base)

# Salida esperada:
#
#   Friedman rank sum test
#
# data:  esperma_matrix
# Friedman chi-squared = 26.133, df = 2, p-value = 2.115e-06
#
# Interpretación:
# - "Friedman chi-squared = 26.133": estadístico Q = 26.13.
# - "df = 2": grados de libertad = k - 1 = 3 - 1 = 2.
# - "p-value = 2.115e-06": p < 0.001. Se rechaza H₀.
#
# Los tres andrólogos asignan puntajes significativamente diferentes.
# El patrón (visto en las medianas) sugiere que B califica más bajo
# y C más alto.

# --- Opción 2: friedman_test() del paquete coin ---
# Requiere los datos en formato largo (data frame) con la fórmula:
# variable ~ tratamiento | bloque

friedman_coin <- friedman_test(
  puntaje ~ andrologo | toro,
  data        = esperma,
  distribution = "asymptotic"
)
print(friedman_coin)
# El resultado debe ser esencialmente idéntico al de friedman.test().
```

Coeficiente de concordancia W de Kendall

```r
# Cálculo a partir del estadístico de Friedman
Q_val <- friedman_base$statistic   # 26.133
b_val <- 15                        # número de bloques (toros)
k_val <- 3                         # número de tratamientos (andrólogos)

W_kendall <- as.numeric(Q_val / (b_val * (k_val - 1)))
cat("Coeficiente de concordancia W de Kendall:", round(W_kendall, 4), "\n")

# W = 26.133 / (15 × 2) = 26.133 / 30 = 0.8711
#
# Interpretación: concordancia fuerte (W > 0.70).
# Los tres andrólogos están altamente de acuerdo en el ORDEN relativo
# de calidad de los toros. Cuando uno califica a un toro como bueno,
```

```
# los otros dos también lo califican relativamente alto; cuando uno
# lo califica como malo, los otros concuerdan.
#
# Combinado con el Friedman significativo, el panorama es:
# - Los andrólogos CONCUERDAN en el orden (W alto).
# - Los andrólogos DIFIEREN en el nivel de calificación (Friedman significativo).
# - Específicamente, B es más estricto y C más generoso.
# - Las calificaciones no son intercambiables en valor absoluto,
#   pero sí lo son en términos relativos.
```

Comparaciones post hoc

```
# ============================================================
# A) Nemenyi (análogo a Tukey HSD para Friedman)
# ============================================================

nemenyi_result <- PMCMRplus::frdAllPairsNemenyiTest(esperma_matrix)
print(nemenyi_result)

# Salida esperada: matriz triangular inferior de valores p.
#     A        B
# B   0.017    -
# C   0.150    < 0.001
#
# Interpretación:
# - A vs B: p = 0.017 → Difieren significativamente.
# - B vs C: p < 0.001 → Difieren significativamente.
# - A vs C: p = 0.150 → No difieren significativamente.
#
# El andrólogo B califica significativamente más bajo que A y que C.
# Los andrólogos A y C no difieren entre sí.


# ============================================================
# B) Conover para Friedman (mayor potencia)
# ============================================================

conover_friedman <- PMCMRplus::frdAllPairsConoverTest(
  esperma_matrix,
  p.adjust.method = "holm"
)
print(conover_friedman)
```

```
# Conover suele ser más potente que Nemenyi. Es posible que aquí
# también detecte la diferencia A vs C que Nemenyi no detectó,
# dependiendo de la magnitud exacta de la diferencia.
```

Prueba de Quade y sus comparaciones post hoc

```
# ============================================================
# Prueba de Quade
# ============================================================
quade_result <- quade.test(esperma_matrix)
print(quade_result)

# Salida: estadístico F de Quade, grados de libertad del numerador
# y denominador, y valor p.
# Con k = 3 (pocos tratamientos), Quade puede ser más potente que Friedman.

# ============================================================
# Post hoc para Quade
# ============================================================
quade_posthoc <- PMCMRplus::quadeAllPairsTest(
  esperma_matrix,
  p.adjust.method = "holm"
)
print(quade_posthoc)
```

Tabla de rangos medios

```
# Tabla resumen de rangos por andrólogo para reportar en un artículo

rangos_friedman <- esperma %>%
  group_by(toro) %>%
  mutate(rango_bloque = rank(puntaje, ties.method = "average")) %>%
  ungroup() %>%
  group_by(andrologo) %>%
  summarise(
    mediana_puntaje = median(puntaje),
    rango_medio     = round(mean(rango_bloque), 3),
    suma_rangos     = sum(rango_bloque),
    .groups         = "drop"
  )

print(rangos_friedman)
```

```
# Tabla 3.9. Medianas de puntaje de calidad espermática y rangos medios
# por andrólogo (b = 15 toros).
#
# Andrólogo | Mediana | Rango medio | Suma de rangos
# ----------|---------|-------------|---------------
#    A      |    3    |    2.000    |     30.0
#    B      |    2    |    1.233    |     18.5
#    C      |    4    |    2.767    |     41.5
```

Visualización: gráfico de perfiles

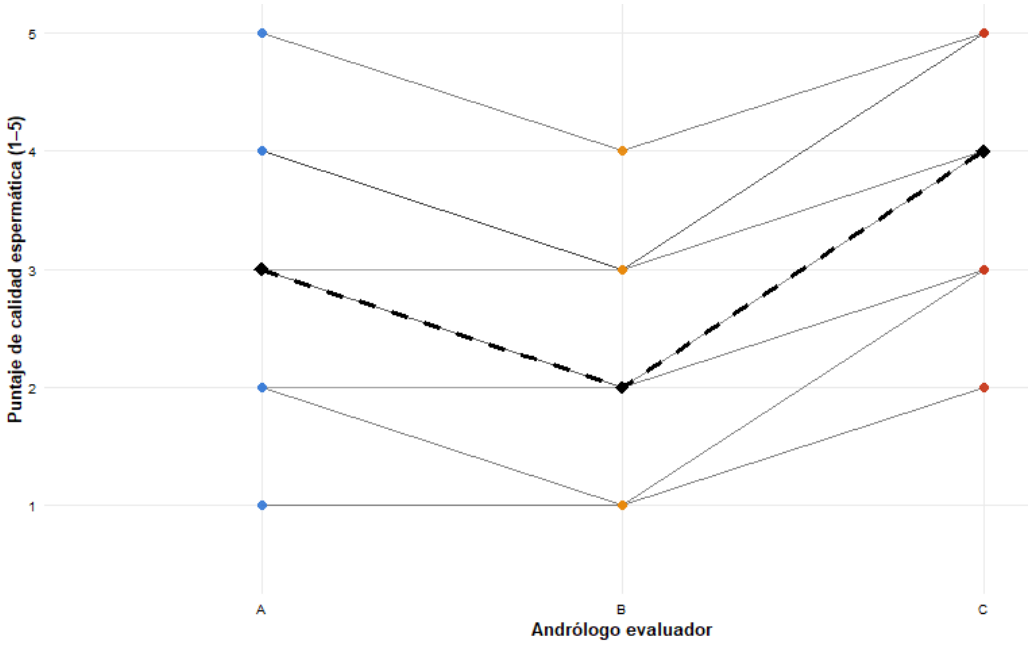
```
# Gráfico de perfiles: cada línea gris conecta los puntajes de un toro
# a través de los tres andrólogos; la línea negra conecta las medianas.

p3 <- ggplot(esperma, aes(x = andrologo, y = puntaje, group = toro)) +
  geom_line(alpha = 0.25, color = "gray50", linewidth = 0.4) +
  geom_point(aes(color = andrologo), size = 3, alpha = 0.7) +
  # Línea de medianas
  stat_summary(
    aes(group = 1),
    fun      = median,
    geom     = "line",
    color    = "black",
    linewidth = 1.2,
    linetype  = "dashed"
  ) +
  stat_summary(
    aes(group = 1),
    fun   = median,
    geom  = "point",
    color = "black",
    size  = 4.5,
    shape = 18   # Diamante
  ) +
  scale_color_manual(values = c("#3B7DD8", "#E8890C", "#C93A1D")) +
  scale_y_continuous(breaks = 1:5, limits = c(0.5, 5.5)) +
  labs(
    x = "Andrólogo evaluador",
    y = "Puntaje de calidad espermática (escala 1-5)"
  ) +
  theme_minimal(base_size = 13) +
  theme(legend.position = "none")

print(p3)
```

Figura 6

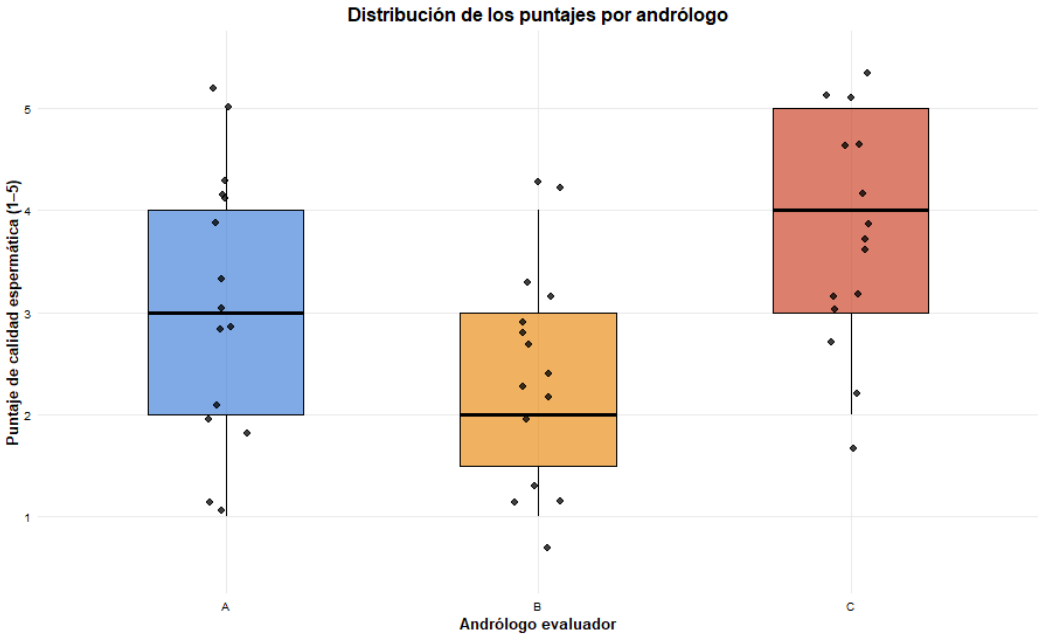
Perfil de puntajes de calidad espermática de 15 toros Suizo Pardo evaluados por tres andrólogos.



Nota. Cada línea gris conecta los puntajes de un toro; la línea negra punteada con diamantes con ecta las medianas por andrólogo. Se observa que la mayoría de las líneas suben de B a C, y que l a mediana de B (2) es inferior a la de A (3) y la de C (4).

Figura 7

Distribución de puntajes por tres andrólogos



3.11. Estudio de caso integrador

- **Caso A: Infestación por garrapatas en bovinos de cuatro genotipos**

Contexto del problema

En la Estación Experimental del INIFAP en Tizimín, Yucatán, se evaluó la resistencia genética a *Rhipicephalus microplus* en cuatro genotipos bovinos bajo condiciones de infestación natural en pastoreo tropical húmedo. El objetivo práctico era cuantificar la ventaja de los genotipos cebuínos y evaluar si la cruza F1 Holstein × Cebú ofrece un compromiso útil entre producción lechera y resistencia a ectoparásitos.

Se seleccionaron 48 vacas primíparas (12 por genotipo) con edades similares (26–34 meses) y sin tratamiento acaricida durante los 60 días previos al estudio. El conteo de garrapatas adultas ingurgitadas (> 4 mm) se realizó en el costado izquierdo de cada animal el día 21 del periodo experimental (julio-septiembre, época de máxima actividad del vector).

Los datos crudos se presentaron en la sección 3.10.2.

Justificación del método no paramétrico

Se verificaron los supuestos del ANOVA de un factor:

- Normalidad de residuos: Shapiro-Wilk sobre los residuos del modelo ANOVA dio W = 0.91, p = 0.003. Rechazo de normalidad.
- Homocedasticidad: Levene basada en medianas dio $F_{3,44} = 14.8$, $p < 0.001$. La varianza del grupo Holstein (DE ≈ 70) fue 20 veces mayor que la del grupo Brahman (DE ≈ 3.8). Rechazo de homocedasticidad.
- Se intentaron transformaciones log(y + 1), √y y^(1/4). Ninguna normalizó los residuos ni igualó las varianzas simultáneamente. La transformación logarítmica mejoró parcialmente la normalidad, pero no la homocedasticidad; la raíz cuarta mejoró la homocedasticidad, pero distorsionó la normalidad en los grupos extremos.

Se concluyó que las condiciones para el ANOVA paramétrico no se cumplen y las transformaciones no las restauran. Se aplicó la prueba de Kruskal-Wallis con comparaciones post hoc de Dunn (ajuste de Holm).

Resultados

Estadística descriptiva (Tabla 3.10):

Genotipo	n	Mediana	Q1–Q3	Rango	Media ± DE
Brahman	12	8.5	5.25–11.75	3–15	8.7 ± 3.8
Gyr	12	24.5	19.75–28.75	15–35	24.6 ± 5.7
F1 HolxCebú	12	74.0	54.25–92.25	45–120	75.8 ± 23.6
Holstein	12	237.5	177.5–297.5	150–387	249.8 ± 69.6

La prueba de Kruskal-Wallis reveló diferencias significativas entre genotipos ($H = 43.27$, $gl = 3$, $p < 0.001$, $\varepsilon^2 = 0.92$). Las comparaciones post hoc de Dunn con ajuste de Holm mostraron que todos los pares difieren significativamente (todos $p\_adj < 0.05$), excepto F1 versus Gyr ($p\_adj = 0.19$) y F1 versus Holstein ($p\_adj = 0.19$).

Asignación de letras de grupo: Brahman (a), Gyr (b), F1 (bc), Holstein (c).

Interpretación veterinaria

La mediana de carga de garrapatas en el genotipo Brahman (8.5 garrapatas/animal) fue 28 veces menor que en el Holstein puro (237.5 garrapatas/animal). El genotipo explicó más del 92 % de la variabilidad en los rangos del conteo ($\varepsilon^2 = 0.92$), lo cual refleja la magnitud biológica del efecto de la composición genética *Bos indicus* sobre la resistencia a ectoparásitos (Jonsson, 2006).

La cruza F1 Holstein × Cebú se ubicó en una posición intermedia (mediana = 74) que no difirió significativamente de Gyr ni de Holstein en las comparaciones post hoc. Desde la perspectiva del productor de doble propósito, la cruza F1 reduce la carga parasitaria respecto al Holstein puro en aproximadamente un 70 %, lo cual puede disminuir la frecuencia de baños acaricidas y sus costos asociados. La falta de significancia estadística entre F1 y los grupos adyacentes (Gyr y Holstein) no implica igualdad biológica: con 12 animales por grupo, la potencia para detectar diferencias moderadas entre grupos contiguos es limitada. Un estudio con n = 25 por grupo probablemente distinguiría los cuatro genotipos.

¿Qué habría pasado con el ANOVA?

```
# ANOVA de un factor (análisis incorrecto, para comparación didáctica)
anova_garrapatas <- aov(conteo ~ genotipo, data = garrapatas)
summary(anova_garrapatas)
# El ANOVA también da p < 0.001, porque las diferencias son enormes.
# Pero el valor p del ANOVA no es confiable con estas violaciones.

# Tukey HSD (post hoc paramétrico)
TukeyHSD(anova_garrapatas)
# Tukey también detecta la mayoría de las diferencias, pero sus
# intervalos de confianza asumen varianzas iguales que no se cumplen
# y por lo tanto no son interpretables.
```

En este caso particular, las conclusiones de ANOVA y Kruskal-Wallis coinciden porque las diferencias son extremadamente grandes. Sin embargo, la coincidencia no valida retroactivamente el uso de ANOVA con supuestos violados. En un estudio con diferencias más sutiles (por ejemplo, comparando solo genotipos cebuínos), la violación de homocedasticidad podría llevar al ANOVA a dar resultados engañosos: un estadístico F inflado que rechaza H_0 espuriamente, o deflado que pasa por alto una diferencia real. La elección del método debe basarse en los supuestos, no en si las conclusiones coinciden.

Redacción del resultado para un artículo (formato APA 7.ª edición)

> El conteo de garrapatas *R. microplus* difirió significativamente entre los cuatro genotipos bovinos (prueba de Kruskal-Wallis, $H(3) = 43.27$, $p < .001$, $\varepsilon^2 = 0.92$). Las comparaciones post hoc de Dunn con ajuste de Holm mostraron que Brahman (Mdn = 8.5, RIC = 5.3–11.8) presentó menor carga que Gyr (Mdn = 24.5, RIC = 19.8–28.8; $p = .019$), F1 Holstein × Cebú (Mdn = 74.0, RIC = 54.3–92.3; $p < .001$) y Holstein (Mdn = 237.5, RIC = 177.5–297.5; $p < .001$). Gyr difirió de Holstein ($p = .019$). La cruza F1 no difirió significativamente de Gyr ($p = .195$) ni de Holstein ($p = .195$). Los resultados se presentan en la Tabla 3.10 y la Figura 3.4.

- **Caso B: Calidad espermática evaluada por tres andrólogos**

Contexto del problema

En un centro de inseminación artificial del trópico ecuatoriano, se evaluó la concordancia y el sesgo sistemático entre tres veterinarios andrólogos en la calificación de la calidad espermática de toros Suizo Pardo. El objetivo era determinar: (a) si los tres andrólogos concuerdan en el ordenamiento relativo de los toros por calidad espermática, y (b) si alguno de ellos califica sistemáticamente más alto o más bajo que los otros, lo cual afectaría las decisiones de descarte de eyaculados.

Se evaluaron 15 eyaculados de toros diferentes, cada uno por los tres andrólogos, en una escala ordinal de 1 a 5 basada en la evaluación conjunta de motilidad progresiva, morfología espermática y concentración. Los andrólogos no conocían las calificaciones de los otros.

Justificación del método no paramétrico

La variable respuesta es ordinal: la escala de 1 a 5 tiene un orden natural ($1 < 2 < 3 < 4 < 5$), pero la distancia entre niveles no es necesariamente igual. Un cambio de “inaceptable” (1) a “mala” (2) no equivale biológicamente al cambio de “buena” (4) a “excelente” (5). Además, los datos están apareados por bloque (cada toro fue evaluado por los tres andrólogos), lo que descarta Kruskal-Wallis. La prueba adecuada es Friedman.

Resultados

La prueba de Friedman mostró diferencias significativas entre los puntajes de los tres andrólogos ($Q(2) = 26.13$, $p < .001$). El coeficiente de concordancia de Kendall fue $W = 0.87$, indicando concordancia fuerte.

Las comparaciones post hoc de Nemenyi revelaron: - Andrólogo A vs B: $p = 0.017$ (difieren significativamente) - Andrólogo B vs C: $p < 0.001$ (difieren significativamente) - Andrólogo A vs C: $p = 0.150$ (no difieren significativamente)

El andrólogo B (Mdn = 2, rango medio = 1.23) califica sistemáticamente más bajo que A (Mdn = 3, rango medio = 2.00) y C (Mdn = 4, rango medio = 2.77). Los andrólogos A y C no difieren significativamente entre sí.

Interpretación veterinaria

Los tres andrólogos son consistentes en identificar cuáles son los toros de mejor y peor calidad espermática (W = 0.87). Si los tres dicen que el toro 4 es el mejor, probablemente lo es. El problema es el sesgo de nivel: si el centro usa un umbral fijo para descartar eyaculados (por ejemplo, descartar todo eyaculado con puntaje ≤ 2), el andrólogo B descartará significativamente más eyaculados que A y C. Esto tiene consecuencias económicas directas (eyaculados viables descartados) y reproductivas (pajillas de menor calidad procesadas si un andrólogo liberal las aprueba).

Las opciones para manejar esta discrepancia incluyen: (a) calibración conjunta con eyaculados de referencia cuya calidad ha sido determinada por un laboratorio externo, (b) usar la mediana de los tres evaluadores como puntaje final, lo cual requiere que los tres evalúen cada muestra, o (c) establecer umbrales de descarte específicos por evaluador (umbral 2 para B, umbral 3 para A y C), calibrados para que la tasa de descarte sea la misma. La opción (a) es la más rigurosa; la (b) es la más práctica; la (c) es un parche que requiere monitoreo periódico.

Redacción para artículo (formato APA 7.ª edición)

> Los puntajes de calidad espermática difirieron significativamente entre los tres andrólogos (prueba de Friedman, $Q(2) = 26.13$, $p < .001$, W de Kendall = 0.87). Las comparaciones post hoc de Nemenyi mostraron que el andrólogo B (Mdn = 2) asignó puntajes significativamente más bajos que el andrólogo A (Mdn = 3; $p = .017$) y que el andrólogo C (Mdn = 4; $p < .001$). Los andrólogos A y C no difirieron significativamente entre sí ($p = .150$). A pesar del sesgo de nivel, la concordancia entre evaluadores fue fuerte (W = 0.87), lo que indica acuerdo sustancial en el ordenamiento relativo de la calidad espermática de los toros.

3.12. Errores frecuentes y advertencias

1. Interpretar Kruskal-Wallis como “prueba de medianas”. Kruskal-Wallis compara distribuciones completas, no medianas. Dos grupos con la misma mediana, pero formas distribucionales distintas pueden dar un Kruskal-Wallis significativo. La formulación rigurosa es “las distribuciones difieren”. Solo cuando se ha verificado que las distribuciones tienen la misma forma (misma asimetría, misma dispersión relativa), la interpretación

como “las medianas difieren” es correcta. En la redacción de resultados, la forma segura es reportar las medianas y rangos intercuartílicos como descriptivos, y decir que “la distribución de Y difiere significativamente entre los grupos” o, si las formas son similares, que “las medianas difieren”.

2. Aplicar Kruskal-Wallis a datos apareados. Si los mismos animales reciben todos los tratamientos (medidas repetidas) o si las observaciones están bloqueadas (camada, finca, evaluador), los datos no son independientes y Kruskal-Wallis no corresponde. Usar Friedman. Este error aparece con frecuencia en tesis donde se miden los mismos bovinos antes y después de un tratamiento a tres tiempos diferentes y se analizan los tres tiempos como “tres grupos independientes”.

3. Comparaciones post hoc sin Kruskal-Wallis significativo. Dunn y Conover-Iman son procedimientos condicionales: presuponen que el test omnibus (Kruskal-Wallis) rechazó H_0. Aplicarlos cuando H no es significativo infla el error de tipo I. La excepción es DSCF, que no depende del resultado de Kruskal-Wallis. Si el investigador tiene comparaciones planificadas a priori (por ejemplo, cada tratamiento vs. control), puede usar Mann-Whitney directamente con ajuste de Bonferroni o Holm, sin pasar por Kruskal-Wallis.

4. Omitir el ajuste por comparaciones múltiples. Con 4 grupos hay 6 pares; con 5 grupos, 10; con 6 grupos, 15. Sin ajuste, si se realizan 6 comparaciones al nivel $\alpha = 0.05$, la probabilidad de al menos un falso positivo es $1 - (0.95)^6 \approx 0.26$ más de una en cuatro posibilidades. El ajuste (preferiblemente Holm o Benjamini-Hochberg) no es opcional. Reportar valores p “sin ajuste” en la tabla, pero tomar decisiones con el ajuste es la práctica aceptable; omitir el ajuste completamente es un error serio.

5. No reportar el tamaño del efecto. El valor p no comunica la magnitud de la diferencia. Un estudio con 200 animales por grupo puede dar $p < 0.001$ para una diferencia de 3 garrapatas. Un estudio con 8 animales puede dar $p = 0.08$ para una diferencia de 200 garrapatas. Sin ε^2, η^2_H o r, el lector no puede evaluar la relevancia biológica del hallazgo. La APA 7.ª edición lo exige, y cada vez más revistas veterinarias lo solicitan.

6. Asumir que "no paramétrico = sin supuestos". Kruskal-Wallis y Friedman tienen supuestos reales. El más crítico es la independencia de las observaciones, cuya violación compromete la validez de cualquier prueba estadística, paramétrica o no. El segundo más importante es la suposición de formas distribucionales iguales para interpretar el resultado como diferencia de medianas. Decir en una tesis que "se usó Kruskal-Wallis porque es una prueba libre de supuestos" delata una comprensión incompleta del método.

7. Olvidar la corrección por empates con datos ordinales. Con variables que tienen pocos niveles posibles (condición corporal 1–5, grado histopatológico I–III, puntaje de cojera 0–5), los empates pueden ser muy extensos. La corrección por empates aumenta el estadístico H y puede cambiar la decisión en casos limítrofes. En R, tanto *kruskal.test()* como *kruskal_test()* aplican la corrección automáticamente, pero otro software (algunas versiones antiguas de SAS, por ejemplo) pueden no hacerlo por defecto.

8. Confundir concordancia (W) con acuerdo absoluto. Un W de Kendall alto significa que los evaluadores coinciden en el *orden*. No significa que asignen los mismos puntajes. Tres andrólogos con puntajes (2, 3, 4), (1, 2, 3) y (3, 4, 5) tienen concordancia perfecta (W = 1) pero no coinciden en ningún valor absoluto. Reportar "los evaluadores estuvieron de acuerdo (W = 0.85)" sin especificar que se trata de acuerdo en el ordenamiento puede llevar a conclusiones erróneas sobre la intercambiabilidad de los evaluadores.

9. Usar Friedman cuando los bloques están incompletos. Si un evaluador no evaluó a todos los sujetos (datos faltantes), la prueba de Friedman estándar no aplica porque requiere bloques completos. Alternativas: extensión de Skillings-Mack para bloques incompletos, o modelado con un modelo lineal generalizado mixto (GLMM) con distribución ordinal. Eliminar los bloques incompletos es una opción, pero puede sesgar los resultados si la pérdida de datos no es aleatoria.

3.13. Ejercicios de autoevaluación

Nivel básico

Ejercicio 3.1.

Un investigador compara la ganancia diaria de peso (GDP, en g/día) en corderos de tres razas (Pelibuey, Katahdin, Dorper), con 8 corderos por raza. Los histogramas muestran distribución fuertemente asimétrica a la derecha en las tres razas, y la prueba de Levene rechaza la igualdad de varianzas ($p = 0.008$). ¿Qué prueba corresponde? ¿Cuáles supuestos de esa prueba debe verificar?

> **Respuesta:** Prueba de Kruskal-Wallis, porque se comparan tres grupos independientes con datos que no cumplen normalidad ni homocedasticidad. Los supuestos que debe verificar son: (a) independencia de las observaciones (los corderos deben ser unidades independientes: no crías del mismo parto analizadas como individuos separados, no corderos del mismo corral sin considerar el efecto corral), y (b) la variable es al menos ordinal (GDP en g/día es de razón, cumple). Si quiere interpretar el resultado como comparación de medianas, debe verificar que las distribuciones tengan la misma forma en las tres razas.

Ejercicio 3.2.

Cinco veterinarios califican la severidad de la claudicación de 20 vacas Holstein usando la escala de Sprecher et al. (1997) de 1 a 5. Cada vaca es evaluada por los cinco veterinarios. (a) ¿Cuál es la prueba adecuada para detectar si algún veterinario califica sistemáticamente más alto o más bajo? (b) ¿Cuál indicador mide el grado de acuerdo entre los cinco veterinarios?

> **Respuesta:** (a) Prueba de Friedman, donde los bloques son las vacas y los "tratamientos" son los veterinarios. Los datos son ordinales y apareados. (b) Coeficiente de concordancia W de Kendall. W mide el acuerdo en el ordenamiento relativo de las vacas, no en los puntajes absolutos.

Ejercicio 3.3.

Un tesista obtuvo Kruskal-Wallis $H = 5.2$, $gl = 3$, $p = 0.158$. Procedió a hacer comparaciones de Dunn y encontró que el par "grupo 1 vs grupo 4" da $p\_adj = 0.032$. Reportó esta diferencia como significativa. ¿Es correcto?

> **Respuesta:** No. Dunn es un procedimiento condicional: solo se aplica si Kruskal-Wallis rechaza H_0. Con $p = 0.158$, H_0 no se rechaza, y las comparaciones post hoc no deben

realizarse. La significancia del par 1 vs 4 es probablemente espuria. Si existía una hipótesis a priori sobre ese par específico (planteada antes de ver los datos), se podría haber usado una prueba de Mann-Whitney directa como comparación planificada, pero debe declararse como tal y justificarse.

Nivel intermedio

Ejercicio 3.4. Los siguientes datos corresponden al tiempo de retorno a la actividad ovárica posparto (días) en vacas de tres sistemas de producción en el trópico colombiano. Realice la prueba de Kruskal-Wallis a mano y calcule el tamaño del efecto ε^2.

Extensivo	Semiintensivo	Intensivo
95	72	55
110	68	48
130	85	62
88	78	58
145	90	50

Respuesta:

$N = 15$, $k = 3$, $n_1 = n_2 = n_3 = 5$.

Ordenar y asignar rangos (sin empates):

	4	**5**													
Valor	**8**	**0**	**55**	**58**	**62**	**68**	**72**	**78**	**85**	**88**	**90**	**95**	**110**	**130**	**145**
Rango	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Grupo	I	I	I	I	I	S	S	S	S	E	S	E	E	E	E

Sumas de rangos: $R_E = 10 + 12 + 13 + 14 + 15 = 64$; $R_S = 6 + 7 + 8 + 9 + 11 = 41$; $R_I = 1 + 2 + 3 + 4 + 5 = 15$. Verificación: $64 + 41 + 15 = 120 = 15 \times 16/2$ ✓

$H = [12 / (15 \times 16)] \times [(64^2/5) + (41^2/5) + (15^2/5)] - 3 \times 16$ $H = 0.05 \times [(4096 + 1681 + 225) / 5] - 48$ $H = 0.05 \times 1200.4 - 48 = 60.02 - 48 = 12.02$

Con gl = 2: $\chi^2_{0.01} = 9.210$. Como $12.02 > 9.21$, se rechaza H_0 al 1 %.

$\varepsilon^2 = H / (N - 1) = 12.02 / 14 = 0.858$. Efecto grande. El sistema de producción explica una proporción muy alta de la variabilidad en los rangos del tiempo de retorno a la ciclicidad.

Ejercicio 3.5.

Interprete la siguiente salida de R:

```
    Kruskal-Wallis rank sum test

data:  score by treatment
Kruskal-Wallis chi-squared = 7.34, df = 2, p-value = 0.0255

Dunn (1964) Kruskal-Wallis multiple comparison (Holm)
        Comparison       Z    P.unadj      P.adj
1    Control - Dosis1  -1.12   0.2627      0.2627
2    Control - Dosis2  -2.65   0.0081      0.0243
3     Dosis1 - Dosis2  -1.53   0.1257      0.2514
```

Respuesta: Kruskal-Wallis es significativo (H = 7.34, gl = 2, p = 0.026). Al menos un grupo difiere. Las comparaciones de Dunn con ajuste de Holm muestran que solo el par Control vs Dosis2 difiere significativamente (p_adj = 0.024). Control no difiere de Dosis1 (p_adj = 0.263), y Dosis1 no difiere de Dosis2 (p_adj = 0.251). Dosis2 produce un cambio detectable respecto al Control, pero Dosis1 queda en posición intermedia sin diferir estadísticamente de ninguno. Falta: reportar ε^2 como tamaño del efecto, y los descriptivos de cada grupo (medianas, RIC).

Nivel avanzado

Ejercicio 3.6.

Un investigador evaluó cuatro protocolos de analgesia posquirúrgica en 32 perras sometidas a ovariohisterectomía (8 por protocolo). La variable respuesta es el puntaje de dolor con la escala de Glasgow modificada (0–24) a las 4 horas posquirúrgicas:

Meloxicam	Carprofeno	Tramadol	Dipirona
3	5	8	12
4	6	10	9
2	4	7	14
5	7	11	10
3	5	9	13

Meloxicam	Carprofeno	Tramadol	Dipirona
6	8	12	11
4	6	8	15
2	3	10	12

e) Exploración de datos, justificación del método.

f) Kruskal-Wallis en R: estadístico, gl, valor p.

g) Comparaciones post hoc de Dunn (Holm). ¿Cuáles pares difieren?

h) Tamaño del efecto ε^2 e interpretación.

i) Redacción de la sección de resultados (APA 7.ª ed.) con tabla y referencia a figura.

Rúbrica orientadora: - Exploración completa (2 puntos): descriptivos por grupo (mediana, RIC), Shapiro-Wilk por grupo, Levene, boxplots con jitter. Justificación clara (variable ordinal con escala de Glasgow, o datos no normales, o varianzas desiguales). - Análisis correcto (3 puntos): código R correcto, *kruskal.test()*, *dunnTest()* con *method = "holm"*, ε^2 calculado. - Redacción APA (3 puntos): reporta H(gl) = valor, p, ε^2, medianas y RIC en tabla, identifica pares significativos. Incluye interpretación clínica: ¿cuál protocolo dio menor dolor? ¿La diferencia es clínicamente relevante (no solo estadísticamente)? - Figura (2 puntos): boxplot con jitter, ejes etiquetados, letras de grupo derivadas del post hoc.

Ejercicio 3.7.

Cuatro técnicos de un laboratorio de control de calidad de semen miden la concentración espermática ($\times 10^6$/mL) del mismo eyaculado dividido en 10 alícuotas:

Alícuota	Téc. 1	Téc. 2	Téc. 3	Téc. 4
1	520	495	530	480
2	510	500	525	490
3	540	510	545	500
4	505	490	520	475
5	530	505	540	485
6	515	498	528	492
7	525	502	535	488

Alícuota	Téc. 1	Téc. 2	Téc. 3	Téc. 4
8	535	508	542	495
9	500	485	515	470
10	522	497	532	483

j) Friedman para diferencias entre técnicos. b) W de Kendall e interpretación. c) Quade: comparar resultado con Friedman. d) Post hoc. e) Discusión: ¿los datos son ordinales o de razón? ¿Afecta la elección de Friedman vs. ANOVA MR?

Rúbrica orientadora: - Friedman y Quade (3 puntos): ambas pruebas ejecutadas, estadísticos reportados, interpretación coherente. - W de Kendall (2 puntos): calculado correctamente, interpretado con referencia a la escala. Distingue concordancia (orden) de acuerdo absoluto (nivel). - Post hoc (2 puntos): Nemenyi o Conover, pares identificados, ajuste por comparaciones múltiples. - Discusión metodológica (3 puntos): reconoce que la concentración espermática es de razón (no ordinal), y que Friedman descarta información al convertirla a rangos. Con datos de razón, el ANOVA MR sería más potente si se cumplen sus supuestos (normalidad de residuos, esfericidad). Se justifica Friedman si la normalidad o la esfericidad fallan, o si se duda del cumplimiento con n = 10 bloques. Lo que importa aquí es la calidad del razonamiento, no una respuesta predeterminada.

Ejercicio 3.8.

Construya un ejemplo numérico (máximo 3 grupos, 4 observaciones por grupo) donde dos grupos tengan la misma mediana pero Kruskal-Wallis resulte significativo al compararlos con un tercer grupo. Explique qué implica para la interpretación.

Respuesta orientadora: Ejemplo posible: - Grupo X: 1, 2, 98, 99 (mediana = 50) - Grupo Y: 48, 49, 51, 52 (mediana = 50) - Grupo Z: 200, 210, 220, 230 (mediana = 215)

X e Y tienen la misma mediana (50), pero distribuciones radicalmente distintas: X es bimodal y dispersa, Y es concentrada. En el ranking global (N = 12), X ocupa posiciones extremas (1, 2, 7, 8), Y ocupa posiciones intermedias (3, 4, 5, 6), y Z las más altas (9, 10, 11, 12). Las sumas de rangos difieren entre los tres grupos, y Kruskal-Wallis puede ser significativo, aunque X e Y compartan mediana. La implicación: Kruskal-Wallis

detecta diferencias en la distribución completa, no solo en las medianas. Escribir “las medianas difieren significativamente” sería engañoso aquí.

3.14. Conclusiones del capítulo

Las pruebas de Kruskal-Wallis y Friedman son las herramientas de referencia para comparar tres o más grupos cuando los supuestos del ANOVA paramétrico no se sostienen. Kruskal-Wallis corresponde a diseños con grupos independientes; Friedman, a diseños con bloques o medidas repetidas. Ambas operan con rangos, lo que las hace resistentes a distribuciones asimétricas, valores extremos y varianzas heterogéneas. Y ambas son apropiadas para variables ordinales un tipo de dato que aparece a diario en la evaluación clínica veterinaria y que el ANOVA no está diseñado para manejar.

Ni una ni otra carecen de supuestos. La independencia de las observaciones es irrenunciable, y su violación que ocurre con más frecuencia de lo deseable en tesis veterinarias invalida los resultados de cualquier prueba. La interpretación del resultado como “diferencia de medianas” requiere que las distribuciones de los grupos tengan la misma forma; si no la tienen, lo que se compara son distribuciones completas, y la comunicación del hallazgo debe ajustarse.

Las comparaciones post hoc completan el análisis. Un Kruskal-Wallis significativo sin post hoc dice poco; las comparaciones de Dunn, Conover-Iman o DSCF identifican cuáles grupos difieren entre sí. El ajuste por comparaciones múltiples no es una opción: es un requisito. Y el tamaño del efecto ε^2, η^2_H, W, r aporta la dimensión que el valor p no da: cuánto difieren los grupos, no solo si difieren.

El siguiente capítulo aborda las tablas de contingencia y las pruebas de asociación para variables categóricas (chi-cuadrado, Fisher, McNemar), donde la variable respuesta no es cuantitativa ni ordinal, sino nominal: sano o enfermo, positivo o negativo, presente o ausente.

Capítulo 4

Tablas de contingencia, asociación y correlación no paramétrica

Introducción

Este capítulo trabaja los métodos para analizar variables categóricas y para medir correlación cuando los supuestos de la correlación de Pearson no se cumplen. En la primera parte se desarrollan las pruebas de asociación para tablas de contingencia: la prueba chi-cuadrado de Pearson en sus versiones de bondad de ajuste y de independencia, con las condiciones sobre frecuencias esperadas que determinan su validez; la controversia sobre la corrección de Yates y las posturas de quienes la defienden y quienes la consideran innecesariamente conservadora; la prueba exacta de Fisher para tablas con frecuencias esperadas bajas o muestras pequeñas, extendida a tablas mayores de 2 × 2; la prueba de McNemar para datos categóricos pareados (antes-después con el mismo animal) y la Q de Cochran como su extensión a más de dos mediciones. Se trabajan las medidas de asociación que cuantifican la fuerza del vínculo: odds ratio y riesgo relativo con sus intervalos de confianza para tablas 2 × 2, coeficiente phi para tablas 2 × 2 y V de Cramér para tablas de cualquier dimensión. En la segunda parte se presentan los coeficientes de correlación no paramétrica de Spearman (basado en rangos) y Kendall (basado en pares concordantes y discordantes), con los criterios para elegir entre ambos y frente a Pearson según la naturaleza de los datos, la presencia de valores extremos y el tamaño muestral. Se incluye la correlación parcial no paramétrica para controlar el efecto de terceras variables y la implementación completa en R con los paquetes vcd, epitools, Hmisc, corrplot y ppcor.

Objetivos de aprendizaje

Al finalizar este capítulo, el lector será capaz de:

1. **Construir** tablas de contingencia a partir de datos categóricos veterinarios y seleccionar la prueba de asociación apropiada según el diseño del estudio, el tamaño muestral y el tipo de variable.
2. **Contrastar** las condiciones de aplicación de la prueba chi-cuadrado de Pearson, la corrección de Yates, la prueba exacta de Fisher, la prueba de McNemar y la prueba Q de Cochran, identificando cuándo cada una produce resultados válidos y cuándo no.
3. **Calcular e interpretar** medidas de asociación (odds ratio, riesgo relativo, V de Cramér, coeficiente phi) y de correlación no paramétrica (Spearman, Kendall) tanto a mano como mediante R.

4. **Implementar** en R el análisis completo de datos categóricos y de correlación no paramétrica, desde la construcción de la tabla hasta la visualización gráfica y el reporte de resultados.
5. **Decidir** entre la correlación de Spearman, la tau de Kendall y la correlación de Pearson según las características de los datos, y justificar esa decisión en la sección de métodos de un artículo científico.

Introducción con problema real

En un hospital veterinario universitario de Quito, una residente de cirugía revisó los expedientes de 87 perros de razas grandes y gigantes (Pastor Alemán, Labrador Retriever, Rottweiler, Gran Danés) atendidos en los últimos cuatro años con diagnóstico radiográfico de displasia de cadera. Quería responder dos preguntas: primera, ¿existe asociación entre la raza y la gravedad de la displasia (clasificada según la escala de la OFA en leve, moderada o grave)?; segunda, ¿se correlaciona la carga parasitaria gastrointestinal (huevos por gramo de heces, técnica de McMaster) con el hematocrito en alpacas de una comunidad campesina de Puno, Perú?

Para la primera pregunta, la variable respuesta es ordinal con tres niveles y la variable explicativa es nominal con cuatro categorías. No hay distribución continua que modelar, no hay media ni varianza en el sentido habitual. Para la segunda, los conteos de huevos por gramo siguen distribuciones extremadamente asimétricas, con muchos ceros y algunos valores muy altos (2 400, 4 800 huevos por gramo), que ninguna transformación logarítmica va a convertir en algo remotamente normal.

Ambas situaciones requieren herramientas que no asuman normalidad, que no dependan de medias ni de varianzas homogéneas, y que funcionen con los tipos de datos que la investigación veterinaria produce a diario: proporciones, clasificaciones, conteos, escalas ordinales. Este capítulo presenta esas herramientas.

Se divide en dos bloques. El primero aborda las tablas de contingencia y las pruebas de asociación entre variables categóricas: chi-cuadrado, Fisher, McNemar, Cochran Q, junto con las medidas que cuantifican la fuerza de esa asociación. El segundo bloque trata la correlación no paramétrica (Spearman y Kendall) para variables ordinales o continuas que no cumplen los supuestos de

Pearson. Aunque provienen de tradiciones estadísticas distintas, comparten un principio: analizar datos sin forzarlos a cumplir supuestos que no les corresponden.

4.1. Chi-cuadrado de independencia y de bondad de ajuste

4.1.1. Lógica conceptual

La prueba chi-cuadrado (χ^2) de Karl Pearson, publicada en 1900, responde una pregunta simple: ¿la distribución observada de los datos difiere de lo que se esperaría si las variables fueran independientes (prueba de independencia) o si los datos siguieran una distribución teórica específica (prueba de bondad de ajuste)?

La lógica es directa. Se comparan las frecuencias que se observaron con las frecuencias que se esperarían bajo la hipótesis nula. Si la discrepancia entre lo observado y lo esperado es pequeña, no hay evidencia contra la hipótesis nula. Si es grande, algo está pasando que la hipótesis nula no explica.

Chi-cuadrado de bondad de ajuste

Se utiliza cuando se tiene una sola variable categórica y se quiere saber si la distribución de las observaciones en sus categorías se ajusta a una distribución teórica esperada.

Ejemplo veterinario: En un albergue municipal de Bogotá se registraron 240 perros ingresados durante un año. Se quiere saber si la proporción de perros según tamaño (pequeño, mediano, grande) difiere de la distribución histórica nacional reportada por la asociación de veterinarios: 40 % pequeños, 35 % medianos, 25 % grandes.

Frecuencias observadas: 112 pequeños, 78 medianos, 50 grandes. Frecuencias esperadas (bajo H_0): $240 \times 0.40 = 96$, $240 \times 0.35 = 84$, $240 \times 0.25 = 60$.

Chi-cuadrado de independencia

Se utiliza cuando se tienen dos variables categóricas y se quiere evaluar si están asociadas o son independientes. Los datos se organizan en una tabla de contingencia de $r \times c$ (r filas, c columnas).

Ejemplo veterinario: En un estudio de 150 gatos atendidos en tres clínicas de Lima, se clasificó cada gato según su estado reproductivo (esterilizado / no esterilizado) y la presencia de enfermedad del tracto urinario inferior (presente / ausente). La tabla 2 × 2 resultante permite evaluar si el estado reproductivo y la enfermedad urinaria son independientes.

4.1.2. Fórmula y cálculo paso a paso

El estadístico chi-cuadrado de Pearson se calcula como:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

donde:

- O_{ij} = frecuencia observada en la celda de la fila i, columna j
- E_{ij} = frecuencia esperada en esa misma celda bajo la hipótesis de independencia
- r = número de filas
- c = número de columnas

Las frecuencias esperadas se calculan como:

$$E_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{N}$$

donde $n_{i\cdot}$ es el total marginal de la fila i, $n_{\cdot j}$ es el total marginal de la columna j, y N es el total general de observaciones.

Bajo la hipótesis nula de independencia, el estadístico χ^2 se distribuye asintóticamente como una distribución chi-cuadrado con $(r-1)(c-1)$ grados de libertad.

4.1.3. Ejemplo resuelto: estado reproductivo y enfermedad urinaria en gatos

Tabla 14
Distribución de 150 gatos según estado reproductivo y presencia de enfermedad del tracto urinario inferior (ETUI). Clínicas veterinarias de Lima, Perú

	ETUI presente	ETUI ausente	Total
Esterilizado	28	52	80

	ETUI presente	ETUI ausente	Total
No esterilizado	38	32	70
Total	**66**	**84**	**150**

Frecuencias esperadas:

- $E_{11} = (80 \times 66)/150 = 35.20$
- $E_{12} = (80 \times 84)/150 = 44.80$
- $E_{21} = (70 \times 66)/150 = 30.80$
- $E_{22} = (70 \times 84)/150 = 39.20$

Cálculo del estadístico:

$$\chi^2 = \frac{(28 - 35.20)^2}{35.20} + \frac{(52 - 44.80)^2}{44.80} + \frac{(38 - 30.80)^2}{30.80} + \frac{(32 - 39.20)^2}{39.20}$$

$$\chi^2 = \frac{51.84}{35.20} + \frac{51.84}{44.80} + \frac{51.84}{30.80} + \frac{51.84}{39.20}$$

$$\chi^2 = 1.473 + 1.157 + 1.683 + 1.322 = 5.635$$

Con $(2 - 1)(2 - 1) = 1$ grado de libertad, el valor crítico al 5 % es 3.841. Como 5.635 > 3.841, se rechaza la hipótesis nula de independencia (p = 0.018).

Interpretación veterinaria: existe asociación estadísticamente significativa entre el estado reproductivo y la presencia de ETUI en esta muestra de gatos. Los gatos no esterilizados presentaron mayor proporción de ETUI (54.3 %) que los esterilizados (35.0 %). Este hallazgo es congruente con lo reportado en la literatura respecto al papel de las hormonas sexuales en la fisiopatología del tracto urinario felino (Dorsch et al., 2014).

4.1.4. Supuestos de la prueba chi-cuadrado

La prueba χ^2 de Pearson descansa sobre supuestos que con frecuencia se ignoran:

1. **Independencia de las observaciones.** Cada observación contribuye a una sola celda. No puede haber el mismo animal contado dos veces ni observaciones correlacionadas (para eso existe McNemar, sección 4.4).

2. **Tamaño muestral adecuado.** La regla clásica de Cochran (1954) establece que no más del 20 % de las celdas deben tener frecuencias esperadas menores a 5, y ninguna celda debe tener frecuencia esperada menor a 1. Cuando esto no se cumple, la aproximación chi-cuadrado se deteriora y conviene usar la prueba exacta de Fisher (sección 4.3).
3. **Datos como frecuencias, no proporciones ni porcentajes.** La prueba opera sobre conteos. Introducir porcentajes en lugar de frecuencias absolutas altera el estadístico.
4. **Categorías mutuamente excluyentes y exhaustivas.** Cada observación cae en exactamente una celda.

Un error que aparece con regularidad en tesis veterinarias: aplicar chi-cuadrado a tablas donde las filas o columnas representan al mismo grupo de animales evaluados en momentos distintos. Eso viola independencia. Para datos pareados categóricos, la prueba correcta es McNemar.

4.1.5. Segundo ejemplo: bondad de ajuste en distribución de razas

Un centro de adopción de Santiago de Chile registró durante un año las razas de 180 gatos adoptados: 98 mestizos, 42 siameses, 25 persas y 15 de otras razas. Se quiere contrastar si esta distribución difiere de la proporción nacional reportada: 60 % mestizos, 20 % siameses, 12 % persas, 8 % otros.

Frecuencias esperadas: 108, 36, 21.6, 14.4.

$$\chi^2 = \frac{(98-108)^2}{108} + \frac{(42-36)^2}{36} + \frac{(25-21.6)^2}{21.6} + \frac{(15-14.4)^2}{14.4}$$

$$\chi^2 = 0.926 + 1.000 + 0.535 + 0.025 = 2.486$$

Con 3 grados de libertad, $\chi^2_{0.05,3} = 7.815$. No se rechaza H₀ (p = 0.478). La distribución de razas en el centro de adopción no difiere significativamente de la distribución nacional.

4.2. Corrección de Yates: cuándo aplicarla y controversia

4.2.1. Qué hace la corrección

Frank Yates propuso en 1934 una corrección por continuidad para tablas 2 × 2. La distribución chi-cuadrado es continua, pero las frecuencias observadas son discretas. Yates ajustó la fórmula restando 0.5 del valor absoluto de cada diferencia $(O_{ij} - E_{ij})$ antes de elevar al cuadrado:

$$\chi^2_{Yates} = \sum \frac{\left(|O_{ij} - E_{ij}| - 0.5\right)^2}{E_{ij}}$$

El efecto es reducir el valor del estadístico χ^2, lo que produce valores *p* más grandes (más conservadores).

Para el ejemplo de los gatos con ETUI:

$$\chi^2_{Yates} = \frac{(|28 - 35.2| - 0.5)^2}{35.2} + \frac{(|52 - 44.8| - 0.5)^2}{44.8} + \frac{(|38 - 30.8| - 0.5)^2}{30.8} + \frac{(|32 - 39.2| - 0.5)^2}{39.2}$$

$$\chi^2_{Yates} = \frac{(6.7)^2}{35.2} + \frac{(6.7)^2}{44.8} + \frac{(6.7)^2}{30.8} + \frac{(6.7)^2}{39.2}$$

$$\chi^2_{Yates} = 1.275 + 1.002 + 1.457 + 1.145 = 4.879$$

Con la corrección, $p = 0.027$ (frente a $p = 0.018$ sin corrección). Ambos son significativos al 5 %, pero la diferencia puede cambiar la conclusión en casos limítrofes.

4.2.2. La controversia

La corrección de Yates ha sido objeto de debate durante décadas. Los argumentos principales:

A favor: Cuando el tamaño muestral es pequeño, la distribución chi-cuadrado es una aproximación pobre de la distribución exacta de las frecuencias. La corrección de Yates acerca el valor *p* al valor *p* exacto de Fisher. Conover (1999, p. 201) la recomienda para tablas 2 × 2 con muestras moderadas.

En contra: Varios autores han demostrado que la corrección es excesivamente conservadora, sobre todo con tamaños muestrales intermedios (N entre 20 y 100). Camilli y Hopkins (1978)

mostraron mediante simulación que la corrección de Yates produce tasas de error tipo I por debajo del nivel nominal, lo que equivale a perder potencia sin ganancia real en control del error. Agresti (2013, p. 74) señala que, si la preocupación es la aproximación con muestras pequeñas, es preferible usar directamente la prueba exacta de Fisher en lugar de aplicar un parche a una aproximación deficiente.

Postura práctica para investigación veterinaria: En estudios veterinarios, donde las muestras rara vez son enormes, la recomendación más sensata es: para tablas 2 × 2 con alguna frecuencia esperada menor a 10, usar la prueba exacta de Fisher y reportarla directamente. La corrección de Yates queda como un ajuste histórico que todavía aparece como opción por defecto en algunos programas (de hecho, *chisq.test()* en R la aplica automáticamente en tablas 2 × 2, y hay que desactivarla explícitamente con *correct = FALSE*).

4.3. Prueba exacta de Fisher

4.3.1. Lógica conceptual

La prueba exacta de Fisher no es una aproximación. Calcula la probabilidad exacta de obtener la tabla observada (u otra más extrema) bajo la hipótesis nula de independencia, considerando fijos los totales marginales. Se basa en la distribución hipergeométrica.

Ronald Fisher la propuso originalmente para tablas 2 × 2 con la anécdota de la señora que afirmaba distinguir si la leche se había añadido antes o después del té (Fisher, 1935). El método calcula la probabilidad de cada tabla posible con los mismos marginales que la tabla observada, y suma las probabilidades de las tablas cuyo resultado es igual o más extremo que el observado.

4.3.2. Fórmula

Para una tabla 2 × 2 con celdas *a*, *b*, *c*, *d* y totales marginales fijos:

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{N}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{N!\ a!\ b!\ c!\ d!}$$

El valor *p* bilateral se obtiene sumando las probabilidades de todas las tablas con probabilidad igual o menor a la de la tabla observada.

4.3.3. Cuando usarla

- Tablas 2 × 2 con cualquier frecuencia esperada menor a 5.
- Tablas con muestras muy pequeñas (N < 20-30).
- Siempre que se quiera un valor *p* exacto en lugar de una aproximación.
- Tablas mayores a 2 × 2 con frecuencias esperadas bajas (la extensión de Fisher para tablas r × c, computacionalmente intensiva pero viable con software actual).

En la práctica veterinaria contemporánea, con el poder de cómputo disponible, no hay razón para no usar Fisher en tablas 2 × 2 independientemente del tamaño muestral. Los resultados coincidirán con chi-cuadrado cuando la muestra sea grande, y serán más confiables cuando sea pequeña.

4.3.4. Ejemplo resuelto: eficacia de un antiparasitario en ovinos

En un ensayo clínico realizado en una finca ovina de Riobamba, Ecuador, se evaluó la eficacia de un nuevo antihelmíntico contra *Haemonchus contortus*. Se asignaron aleatoriamente 24 ovejas Corriedale a dos grupos: 12 recibieron el tratamiento y 12 recibieron placebo. Catorce días después, se clasificó a cada animal como "reducción de carga parasitaria ≥ 95 %" (eficaz) o "reducción < 95 %" (no eficaz) según la prueba de reducción del conteo fecal de huevos.

Tabla 15

Eficacia del tratamiento antihelmíntico en ovejas Corriedale. Riobamba, Ecuador.

	Eficaz	**No eficaz**	**Total**
Tratamiento	10	2	12
Placebo	3	9	12
Total	**13**	**11**	**24**

Frecuencias esperadas: $E_{11} = (12 \times 13)/24 = 6.5$, $E_{12} = 5.5$, $E_{21} = 6.5$, $E_{22} = 5.5$.

Todas las frecuencias esperadas superan 5, pero la muestra es pequeña (N = 24). Con chi-cuadrado sin corrección: $\chi^2 = 7.11$, $p = 0.008$. Con corrección de Yates: $\chi^2 = 5.31$, $p = 0.021$. Con Fisher exacta (bilateral): $p = 0.012$.

¿Cuál reportar? Con N = 24, la prueba exacta de Fisher es la opción más transparente. El artículo debería indicar: "Se encontró asociación significativa entre el tratamiento y la eficacia antiparasitaria (prueba exacta de Fisher, $p = 0.012$)".

Cálculo manual del valor *p* de Fisher:

La probabilidad de la tabla observada es:

$$P(a = 10) = \frac{12!\ 12!\ 13!\ 11!}{24!\ 10!\ 2!\ 3!\ 9!} = 0.0266$$

Se calculan las probabilidades de las tablas más extremas (a = 11 y a = 12):

$$P(a = 11) = \frac{12!\ 12!\ 13!\ 11!}{24!\ 11!\ 1!\ 2!\ 10!} = 0.0030$$

$$P(a = 12) = \frac{12!\ 12!\ 13!\ 11!}{24!\ 12!\ 0!\ 1!\ 11!} = 0.0001$$

Para el valor *p* bilateral, se suman también las probabilidades de las tablas extremas del otro lado de la distribución. El resultado final es $p = 0.012$.

4.3.5. Nota sobre tablas mayores a 2 × 2

La extensión de la prueba exacta de Fisher a tablas r × c utiliza la distribución hipergeométrica multivariada. El cálculo puede ser computacionalmente costoso para tablas grandes, pero *fisher.test()* en R lo maneja mediante el algoritmo de simulación de Monte Carlo (*simulate.p.value = TRUE*) cuando el cálculo exacto no es viable. Se volverá a esto en la sección 4.11.

4.4. Prueba de McNemar para datos pareados categóricos

4.4.1. Contexto y lógica

Todas las pruebas anteriores asumen independencia entre las observaciones. Pero en investigación veterinaria son comunes los diseños donde el mismo animal se evalúa antes y después de un tratamiento, o donde se comparan dos pruebas diagnósticas aplicadas al mismo individuo. Cuando

la variable respuesta es dicotómica (sí/no, positivo/negativo, presente/ausente), la prueba de McNemar es la herramienta adecuada.

La prueba fue propuesta por Quinn McNemar en 1947 y se centra exclusivamente en los casos discordantes: aquellos en los que la respuesta cambió entre las dos mediciones. Los casos concordantes (positivo en ambas o negativo en ambas) no aportan información sobre el cambio.

4.4.2. Construcción de la tabla y fórmula

Se organiza la información en una tabla 2 × 2 donde las filas representan la primera medición y las columnas la segunda:

	Después: +	**Después: −**
Antes: +	a	b
Antes: −	c	d

- *a* y *d* son concordantes (no cambiaron).
- *b* son los que pasaron de positivo a negativo (discordantes tipo b).
- *c* son los que pasaron de negativo a positivo (discordantes tipo c).

El estadístico de McNemar es:

$$\chi^2_{McNemar} = \frac{(b-c)^2}{b+c}$$

con 1 grado de libertad. La hipótesis nula es que la probabilidad de cambiar en una dirección es igual a la probabilidad de cambiar en la otra ($\pi_b = \pi_c$, es decir, $P(+\rightarrow -) = P(-\rightarrow +)$).

Con corrección por continuidad (recomendable cuando $b + c < 25$):

$$\chi^2_{McNemar} = \frac{(|b-c|-1)^2}{b+c}$$

Cuando $b + c < 10$, se recomienda usar la prueba binomial exacta en lugar de la aproximación chi-cuadrado.

4.4.3. Ejemplo resuelto: seroconversión posvacunal en bovinos

En una hacienda ganadera de los Llanos Orientales de Colombia, se evaluó la respuesta serológica a una vacuna contra leptospirosis en 60 vacas mestizas (Brahman × Pardo Suizo). Se tomó suero antes de la vacunación y 30 días después. Cada muestra se clasificó como seropositiva (título ≥ 1:200) o seronegativa.

Tabla 16
Resultados serológicos pre y posvacunación contra leptospirosis en 60 vacas. Llanos Orientales, Colombia.

	Posvacunación: +	**Posvacunación: −**	**Total**
Prevacunación: +	12	3	15
Prevacunación: −	28	17	45
Total	**40**	**20**	**60**

Concordantes: a = 12 (positivas en ambos) + d = 17 (negativas en ambos) = 29. Discordantes: b = 3 (pasaron de + a −) y c = 28 (pasaron de − a +).

$$\chi^2_{McNemar} = \frac{(3-28)^2}{3+28} = \frac{625}{31} = 20.16$$

Con 1 grado de libertad, $p < 0.001$.

Interpretación: la proporción de vacas seropositivas cambió significativamente entre antes y después de la vacunación. De las 31 vacas que cambiaron de estado serológico, 28 seroconvirtieron (pasaron de negativas a positivas) y solo 3 perdieron su título. Esta asimetría en el cambio es coherente con una respuesta inmunológica activa a la vacuna.

Nótese que un chi-cuadrado de independencia aplicado a estos datos sería incorrecto: las observaciones no son independientes porque cada vaca contribuye a ambas mediciones.

4.4.4. Supuestos de McNemar

1. Los datos son pareados (el mismo sujeto o unidades emparejadas).
2. La variable es dicotómica.

3. La muestra de pares discordantes ($b + c$) debe ser razonablemente grande para la aproximación chi-cuadrado (al menos 10, preferiblemente 25). Con muestras menores, usar la prueba binomial exacta.

4.5. Prueba Q de Cochran

4.5.1. Extensión de McNemar a más de dos mediciones

La prueba Q de Cochran es la generalización de la prueba de McNemar para más de dos mediciones relacionadas sobre el mismo sujeto, con variable respuesta dicotómica (0/1). Se puede pensar como el equivalente no paramétrico del ANOVA de medidas repetidas, pero para variables binarias.

William Cochran la propuso en 1950. La pregunta que responde: ¿la proporción de "éxitos" (o de respuestas positivas) difiere entre *k* condiciones o momentos evaluados sobre los mismos sujetos?

4.5.2. Fórmula

Se tienen *n* sujetos evaluados en *k* condiciones. Sea X_{ij} la respuesta del sujeto *i* en la condición *j* (0 o 1). Se definen:

- $C_j = \sum_{i=1}^{n} X_{ij}$ = total de éxitos en la condición *j*
- $R_i = \sum_{j=1}^{k} X_{ij}$ = total de éxitos del sujeto *i* a lo largo de todas las condiciones

El estadístico Q de Cochran es:

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} C_j^2 - \left(\sum_{j=1}^{k} C_j\right)^2\right]}{k\sum_{i=1}^{n} R_i - \sum_{i=1}^{n} R_i^2}$$

Bajo H_0 (la proporción de éxitos es igual en las *k* condiciones), *Q* se distribuye aproximadamente como chi-cuadrado con $k - 1$ grados de libertad cuando *n* es grande.

4.5.3. Ejemplo resuelto: comparación de tres pruebas diagnósticas para brucelosis

En un estudio realizado en el altiplano boliviano, se evaluaron tres pruebas diagnósticas para brucelosis bovina (*Brucella abortus*) en 20 vacas de un hato con historial de abortos. Las pruebas

fueron: rosa de Bengala (RB), fijación del complemento (FC) y ELISA indirecta (EI). Cada vaca fue evaluada con las tres pruebas. El resultado de cada prueba se codificó como 1 (positivo) o 0 (negativo).

Tabla 17

Resultados de tres pruebas diagnósticas para brucelosis en 20 vacas. Altiplano boliviano.

Vaca	RB	FC	EI	R_i
1	1	1	1	3
2	1	0	1	2
3	0	0	1	1
4	1	1	1	3
5	0	0	0	0
6	1	1	1	3
7	1	0	1	2
8	0	0	1	1
9	1	1	1	3
10	0	0	0	0
11	1	0	1	2
12	1	1	1	3
13	0	0	1	1
14	1	1	1	3
15	0	0	0	0
16	1	0	1	2
17	0	0	1	1
18	1	1	1	3
19	0	0	0	0
20	1	1	1	3

Totales por prueba: $C_{RB} = 12$, $C_{FC} = 8$, $C_{EI} = 16$.

$\sum C_j = 36$, $\sum C_j^2 = 144 + 64 + 256 = 464$.

$\sum R_i = 36$, $\sum R_i^2 = 9(3^2) + 4(2^2) + 4(1^2) + 3(0^2) = 81 + 16 + 4 + 0 = 101$.

$$Q = \frac{(3-1)[3 \times 464 - 36^2]}{3 \times 36 - 101} = \frac{2[1392 - 1296]}{108 - 101} = \frac{2 \times 96}{7} = 27.43$$

Con $k - 1 = 2$ grados de libertad, $p < 0.001$.

La proporción de positivos difiere significativamente entre las tres pruebas. ELISA detectó 16/20 vacas como positivas (80 %), rosa de Bengala 12/20 (60 %) y fijación del complemento 8/20 (40 %). Esto tiene implicaciones prácticas para la vigilancia epidemiológica: la elección de la prueba diagnóstica afecta directamente la prevalencia aparente que se estima en un hato.

Si la prueba Q resulta significativa y se quiere saber cuáles pares de pruebas difieren entre sí, se pueden realizar pruebas de McNemar pareadas con corrección de Bonferroni o utilizar el procedimiento de Dunn para la Q de Cochran (disponible en el paquete *RVAideMemoire* de R).

4.6. Medidas de asociación: odds ratio, riesgo relativo, V de Cramér, phi

Detectar una asociación estadísticamente significativa es solo la mitad del trabajo. La otra mitad es cuantificar qué tan fuerte es esa asociación. El valor *p* no mide la fuerza de la relación; mide la evidencia contra la hipótesis nula. Con una muestra de 10 000 animales, una asociación trivial puede ser altamente significativa. Con una muestra de 20, una asociación clínicamente relevante puede no alcanzar significancia.

4.6.1. Odds ratio (OR)

En una tabla 2 × 2 con celdas *a*, *b*, *c*, *d*:

$$OR = \frac{a \times d}{b \times c}$$

El OR representa cuántas veces más probable es que un evento ocurra en el grupo expuesto respecto al no expuesto, en términos de odds (cocientes de probabilidades). Un OR = 1 indica ausencia de asociación. Un OR > 1 indica asociación positiva (el factor aumenta las odds del evento). Un OR < 1 indica asociación negativa (el factor reduce las odds).

Para el ejemplo de ETUI en gatos (Tabla 4.1):

$$OR = \frac{28 \times 32}{52 \times 38} = \frac{896}{1976} = 0.453$$

Los gatos no esterilizados tienen un OR de $1/0.453 = 2.21$ respecto a los esterilizados. Reinterpretado: las odds de presentar ETUI en gatos no esterilizados son 2.21 veces las de los esterilizados.

El intervalo de confianza al 95 % del logaritmo natural del OR se calcula como:

$$\ln(OR) \pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

$$\ln(0.453) \pm 1.96\sqrt{\frac{1}{28}+\frac{1}{52}+\frac{1}{38}+\frac{1}{32}}$$

$$-0.791 \pm 1.96 \times 0.396 = -0.791 \pm 0.776$$

$IC_{95\%}$ para $\ln(OR)$: $[-1.567, -0.015]$

Antilogaritmo: $IC_{95\%}$ para OR: [0.209, 0.985].

Como el intervalo no incluye 1 (en la escala de OR) ni 0 (en la escala logarítmica), la asociación es significativa.

4.6.2. Riesgo relativo (RR)

El riesgo relativo solo tiene sentido en estudios de cohorte o en ensayos clínicos, donde se puede estimar directamente la incidencia o la proporción de individuos con el evento en cada grupo. En estudios de casos y controles, el RR no se puede calcular directamente (se usa el OR como aproximación).

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

Para el ejemplo de ETUI (asumiendo que el diseño lo permite):

$$RR = \frac{28/80}{38/70} = \frac{0.350}{0.543} = 0.645$$

La proporción de ETUI en gatos esterilizados es 0.645 veces la de los no esterilizados. O, invertido: los no esterilizados tienen un riesgo 1.55 veces mayor.

La diferencia entre OR y RR se vuelve sustancial cuando la prevalencia del evento es alta (digamos, > 20 %). Con prevalencias bajas, OR ≈ RR. En el ejemplo, la prevalencia de ETUI es 66/150 = 44 %, y efectivamente OR (2.21) y RR (1.55) difieren de forma apreciable. Reportar OR = 2.21 como si fuera RR ("el riesgo se duplica") sería una exageración cuantitativa.

4.6.3. Coeficiente phi (φ)

Para tablas 2 × 2, el coeficiente phi es una medida de asociación simétrica que varía entre −1 y 1:

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Es equivalente a la correlación de Pearson calculada sobre dos variables dicotómicas codificadas como 0 y 1. Para el ejemplo de ETUI:

$$\phi = \frac{(28)(32) - (52)(38)}{\sqrt{80 \times 70 \times 66 \times 84}} = \frac{896 - 1976}{\sqrt{31\,449\,600}} = \frac{-1080}{5608.9} = -0.193$$

El signo negativo refleja que la esterilización se asocia con menor frecuencia de ETUI (depende de cómo se codifiquen las filas y columnas). El valor absoluto (0.193) indica una asociación débil a moderada según la clasificación convencional de Cohen (1988): 0.10 = pequeña, 0.30 = mediana, 0.50 = grande.

Existe una relación directa: $\phi^2 = \chi^2/N$, por lo que $\phi = \sqrt{\chi^2/N}$.

4.6.4. V de Cramér

Para tablas mayores a 2 × 2, phi no está acotada entre −1 y 1. La V de Cramér generaliza phi:

$$V = \sqrt{\frac{\chi^2}{N \times \min(r-1, c-1)}}$$

donde $\min(r-1, c-1)$ es el menor entre el número de filas menos 1 y el número de columnas menos 1. V varía entre 0 (independencia total) y 1 (asociación perfecta).

Para interpretar V en tablas de distintos tamaños, Cohen (1988) sugirió criterios ajustados al número de grados de libertad del menor lado de la tabla. Para tablas con $\min(r-1, c-1) = 1$ (incluidas las 2 × 2): 0.10, 0.30, 0.50. Para tablas con $\min(r-1, c-1) = 2$: 0.07, 0.21, 0.35. Para tablas con $\min(r-1, c-1) \geq 3$: 0.06, 0.17, 0.29.

4.7. Correlación de Spearman: lógica, cálculo y manejo de empates

4.7.1. ¿Qué mide y por qué existe?

La correlación de Pearson cuantifica la relación lineal entre dos variables continuas con distribución normal bivariada. Cuando alguna de estas condiciones falla las variables son ordinales, la relación no es lineal pero sí monótona, hay valores atípicos extremos, las distribuciones son muy asimétricas, el coeficiente de Pearson puede distorsionarse gravemente.

Charles Spearman propuso en 1904 una alternativa: en lugar de usar los valores originales, convertirlos a rangos y calcular Pearson sobre los rangos. El resultado es un coeficiente que mide la relación monótona entre dos variables: si una variable tiende a aumentar (o disminuir) cuando la otra aumenta, independientemente de que la relación sea lineal.

La idea es elegante. Los rangos son inmunes a valores atípicos (porque el valor más alto recibe el rango *n* sin importar si es 100 o 100 000) y no requieren distribución normal.

4.7.2. Fórmula sin empates

Cuando no hay empates (ningún valor repetido en ninguna de las dos variables), el coeficiente de Spearman se calcula con la fórmula abreviada:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

donde:

- $d_i = R(X_i) - R(Y_i)$ = diferencia entre el rango de la observación *i* en X y su rango en Y

- n = número de pares de observaciones

El coeficiente r_s varía entre −1 (relación monótona decreciente perfecta) y +1 (relación monótona creciente perfecta). Un valor de 0 indica ausencia de relación monótona.

4.7.3. Ejemplo resuelto sin empates: peso corporal y longitud de fémur en cachorros

Se midieron el peso corporal (kg) y la longitud radiográfica del fémur (cm) en 8 cachorros de Labrador Retriever de 4 meses de edad en una clínica de Guayaquil.

Tabla 18
Peso corporal, longitud femoral, rangos y cálculo de r_s en 8 cachorros Labrador Retriever.

Cachorro	Peso (kg)	R(Peso)	Fémur (cm)	R(Fémur)	d_i	d_i^2
1	8.2	3	9.1	2	1	1
2	9.5	5	10.3	5	0	0
3	7.1	1	8.7	1	0	0
4	10.8	7	11.2	7	0	0
5	8.9	4	9.8	4	0	0
6	7.8	2	9.4	3	-1	1
7	11.3	8	11.9	8	0	0
8	10.1	6	10.8	6	0	0

$$\sum d_i^2 = 2$$

$$r_s = 1 - \frac{6 \times 2}{8(64 - 1)} = 1 - \frac{12}{504} = 1 - 0.024 = 0.976$$

La correlación de Spearman es 0.976, lo que indica una relación monótona positiva muy fuerte entre el peso corporal y la longitud del fémur en estos cachorros. Los cachorros más pesados tienden a tener fémures más largos, con muy pocas inversiones en el orden de los rangos.

4.7.4. Manejo de empates

Cuando hay valores repetidos, se asigna a cada uno el promedio de los rangos que les habrían correspondido. Por ejemplo, si dos observaciones comparten la posición 3 y 4, ambas reciben rango 3.5.

Con empates, la fórmula abreviada ($1 - 6\sum d_i^2/[n(n^2-1)]$) produce un resultado aproximado, pero no exacto. La fórmula correcta es calcular la correlación de Pearson directamente sobre los rangos (promediados):

$$r_s = \frac{\sum_{i=1}^{n}(R(X_i) - \bar{R}_X)\,(R(Y_i) - \bar{R}_Y)}{\sqrt{\sum_{i=1}^{n}(R(X_i) - \bar{R}_X)^2 \sum_{i=1}^{n}(R(Y_i) - \bar{R}_Y)^2}}$$

donde $\bar{R}_X$ y $\bar{R}_Y$ son las medias de los rangos (que siempre es $(n+1)/2$ si no se descartan observaciones).

Esta es, de hecho, la fórmula que R utiliza internamente en *cor.test(method = "spearman")*.

4.7.5. Ejemplo resuelto con empates: condición corporal y producción láctea en vacas

Se evaluó la condición corporal (escala de 1 a 5, con incrementos de 0.5) y la producción promedio diaria de leche (litros/día) en 10 vacas Holstein de un establo en Cajamarca, Perú.

Tabla 19
Condición corporal y producción láctea en 10 vacas Holstein. Cajamarca, Perú.

Vaca	CC	R(CC)	Leche (L/d)	R(Leche)	d_i	d_i^2
1	2.5	2.5	22.4	7	-4.5	20.25
2	3.0	5	19.8	5	0	0
3	2.5	2.5	24.1	9	-6.5	42.25
4	3.5	7.5	16.3	2	5.5	30.25
5	2.0	1	25.7	10	-9	81.00
6	3.0	5	20.5	6	-1	1.00
7	3.5	7.5	17.9	4	3.5	12.25
8	4.0	9.5	14.2	1	8.5	72.25
9	3.0	5	17.1	3	2	4.00
10	4.0	9.5	23.0	8	1.5	2.25

Hay empates en CC: dos vacas con 2.5 (rangos 2 y 3, promedio = 2.5), tres con 3.0 (rangos 4, 5 y 6, promedio = 5), dos con 3.5 (rangos 7 y 8, promedio = 7.5), dos con 4.0 (rangos 9 y 10, promedio = 9.5).

Con la fórmula abreviada: $\sum d_i^2 = 265.5$

$$r_s \approx 1 - \frac{6 \times 265.5}{10(100-1)} = 1 - \frac{1593}{990} = 1 - 1.609 = -0.609$$

Este valor es una aproximación. La fórmula exacta (Pearson sobre rangos) da $r_s = -0.632$. La diferencia existe precisamente por los empates. En la práctica, siempre se debe usar la fórmula de Pearson sobre rangos (que es lo que hace R), no la fórmula abreviada, cuando hay empates.

Interpretación veterinaria: la correlación negativa moderada a fuerte ($r_s = -0.632$) sugiere que las vacas con mayor condición corporal tienden a producir menos leche, y viceversa. Esto es consistente con el fenómeno fisiológico conocido: las vacas de alta producción movilizan reservas corporales para sostener la lactación, reduciendo su condición corporal, mientras que las vacas de baja producción acumulan grasa. Sin embargo, con solo 10 animales, este resultado debe interpretarse con cautela.

4.7.6. Prueba de significancia para r_s

Para muestras pequeñas (n ≤ 30), se consultan tablas de valores críticos de Spearman (disponibles en Conover, 1999, Tabla A.10, o en Siegel y Castellan, 1988).

Para muestras grandes (n > 30), se puede usar la aproximación con la distribución *t*:

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

con $n - 2$ grados de libertad.

Para el ejemplo de condición corporal ($r_s = -0.632$, $n = 10$):

$$t = -0.632 \sqrt{\frac{8}{1-0.399}} = -0.632 \sqrt{\frac{8}{0.601}} = -0.632 \times 3.649 = -2.306$$

Con 8 grados de libertad, el valor *p* bilateral ≈ 0.050. La correlación se encuentra en el límite de la significancia al 5 %. Con una muestra de 10 vacas, la potencia de la prueba es baja, lo que dificulta detectar correlaciones moderadas.

4.8. Tau de Kendall: variantes τ-a, τ-b, τ-c y cuándo preferirlo sobre Spearman

4.8.1. Lógica conceptual

Maurice Kendall propuso en 1938 un enfoque diferente al de Spearman. En lugar de convertir los datos a rangos y calcular una correlación, Kendall examina todos los pares posibles de observaciones y clasifica cada par como concordante o discordante.

Un par (X_i, Y_i) y (X_j, Y_j) es:

- **Concordante** si ambas diferencias tienen el mismo signo: $(X_i - X_j)(Y_i - Y_j) > 0$. Es decir, si el individuo que tiene mayor valor en X también tiene mayor valor en Y.
- **Discordante** si las diferencias tienen signo opuesto: $(X_i - X_j)(Y_i - Y_j) < 0$.
- **Empatado** si alguna de las diferencias es cero (o ambas).

La proporción de pares concordantes menos discordantes, normalizada, produce el coeficiente tau.

4.8.2. Las tres variantes

Tau-a (τ_a)

La versión más simple. No corrige por empates:

$$\tau_a = \frac{C - D}{\binom{n}{2}} = \frac{C - D}{n(n-1)/2}$$

donde C = número de pares concordantes y D = número de pares discordantes. Solo es apropiada cuando no hay empates en ninguna de las dos variables (situación rara en la práctica veterinaria).

Tau-b (τ_b)

Corrige por empates en X y en Y por separado:

$$\tau_b = \frac{C - D}{\sqrt{(T_0 - T_X)(T_0 - T_Y)}}$$

donde:

- $T_0 = n(n-1)/2$ = número total de pares
- $T_X = \sum_k t_k\,(t_k - 1)/2$ = número de pares empatados en X (t_k es el número de observaciones en el *k*-ésimo grupo de empate de X)
- $T_Y = \sum_l u_l\,(u_l - 1)/2$ = número de pares empatados en Y (u_l análogo para Y)

τ_b puede alcanzar −1 o +1 solo cuando las dos variables tienen la misma distribución de empates (es decir, la misma estructura de categorías). Esto la hace adecuada para tablas cuadradas, pero no siempre para tablas rectangulares.

Tau-c (τ_c), también llamada tau de Stuart

Diseñada para tablas rectangulares ($r \times c$ con $r \neq c$):

$$\tau_c = \frac{2m(C - D)}{n^2(m-1)}$$

donde $m = \min(r, c)$. Siempre puede alcanzar −1 y +1, independientemente de la forma de la tabla.

4.8.3. Ejemplo resuelto: escala de dolor y dosis analgésica en perros posquirúrgicos

En una clínica veterinaria de Medellín, se evaluó la escala de dolor de Glasgow modificada (0 a 10, ordinal) y la dosis total de tramadol administrada (mg/kg en 24 horas) en 6 perros sometidos a ovariohisterectomía.

Tabla 20
Escala de dolor de Glasgow y dosis de tramadol en 6 perras posquirúrgicas. Medellín, Colombia.

Perra	Dolor (Glasgow)	Tramadol (mg/kg/24h)
A	3	4.0
B	7	8.5
C	2	3.0
D	5	6.0
E	8	9.0
F	4	5.5

No hay empates en ninguna variable. Número total de pares: $\binom{6}{2} = 15$.

Se comparan todos los pares:

Par	(X_i-X_j)	(Y_i-Y_j)	Producto	Tipo
A-B	−4	−4.5	+	C
A-C	+1	+1.0	+	C
A-D	−2	−2.0	+	C
A-E	−5	−5.0	+	C
A-F	−1	−1.5	+	C
B-C	+5	+5.5	+	C
B-D	+2	+2.5	+	C
B-E	−1	−0.5	+	C
B-F	+3	+3.0	+	C
C-D	−3	−3.0	+	C
C-E	−6	−6.0	+	C
C-F	−2	−2.5	+	C
D-E	−3	−3.0	+	C
D-F	+1	+0.5	+	C
E-F	+4	+3.5	+	C

C = 15, D = 0.

$$\tau_a = \tau_b = \frac{15 - 0}{15} = 1.000$$

La concordancia es perfecta: cada par de perras que difiere en dolor difiere en la misma dirección en dosis de tramadol. Esto tiene sentido clínico: a mayor dolor evaluado, mayor dosis de analgésico administrada.

(En la práctica real, rara vez se obtiene concordancia perfecta, y los valores de tau suelen ser menores que los de Spearman para los mismos datos, como se discute en la sección 4.9.)

4.8.4. Cuando preferir Kendall sobre Spearman

La pregunta "¿Spearman o Kendall?" aparece con frecuencia en tesis veterinarias, y la respuesta no es "cualquiera de los dos da igual". Cada uno tiene ventajas:

Kendall es preferible cuando:

- Los datos son ordinales con pocas categorías (por ejemplo, escalas de dolor, condición corporal, clasificaciones de gravedad). τ_b maneja empates de forma más transparente que Spearman.
- Se quiere una interpretación probabilística directa: τ_b se puede interpretar como la diferencia entre la probabilidad de concordancia y la probabilidad de discordancia para un par aleatorio.
- La muestra es pequeña (n < 10). Las tablas de valores críticos de Kendall son más precisas para muestras muy pequeñas.
- Se busca una medida que sea más robusta ante perturbaciones en los datos. Kendall es menos sensible a un solo par discordante extremo que Spearman (Newson, 2002)

Spearman es preferible cuando:

- Las variables son continuas o tienen muchos valores posibles (pocos empates). En esos casos, la fórmula basada en rangos es computacionalmente simple y la interpretación como "correlación lineal entre rangos" resulta clara.
- Se desea comparar con resultados de Pearson. Spearman está en la misma escala y tiene la misma interpretación de "proporción de varianza explicada" (r_s^2) que Pearson, lo que facilita la comparación.
- La tradición de la disciplina lo favorece. En muchas revistas veterinarias, Spearman es la correlación no paramétrica "por defecto" y los revisores están familiarizados con ella.

En la práctica, ambos coeficientes suelen llevar a las mismas conclusiones sobre significancia, aunque sus magnitudes difieren (τ tiende a ser menor que r_s en valor absoluto). Si se reportan ambos, el lector del artículo debería entender por qué se eligió uno o ambos.

4.9. Comparación con la correlación de Pearson

4.9.1. ¿Cuándo Pearson falla?

La correlación de Pearson (r) asume:

1. Ambas variables son continuas.

2. La relación es lineal.
3. Las variables tienen distribución normal bivariada (o al menos distribuciones marginales razonablemente simétricas con muestras grandes).
4. Homocedasticidad (la variabilidad de Y no cambia con X).
5. Ausencia de valores atípicos influyentes.

En investigación veterinaria, estas condiciones se violan con frecuencia:

- **Variables ordinales.** Puntajes de condición corporal (escala de 1 a 5), escalas de claudicación (grados 0 a 4), escalas de gravedad histopatológica (leve, moderado, grave). Aplicar Pearson a una escala ordinal de 5 puntos es, en el mejor de los casos, una aproximación cuestionable, porque Pearson asume que la distancia entre 1 y 2 es la misma que entre 4 y 5, cosa que no está garantizada en una escala ordinal.
- **Distribuciones extremadamente asimétricas.** Conteos de parásitos, títulos de anticuerpos, concentraciones de bacterias, tiempos de sobrevivencia. Con una distribución log-normal o cero-inflada, Pearson puede subestimar la fuerza de la relación monótona.
- **Valores atípicos.** Un solo animal con un conteo parasitario de 15 000 huevos por gramo en un grupo donde el máximo del resto es 800 puede inflar o colapsar el coeficiente de Pearson. Spearman, al usar rangos, le asigna el rango más alto sin dejarse arrastrar por la magnitud del valor.
- **Relaciones monótonas no lineales.** La concentración de un fármaco en sangre puede disminuir exponencialmente con el tiempo: la relación es monótona (siempre disminuye) pero no lineal. Pearson subestimará la fuerza de esa relación; Spearman la capturará adecuadamente.

4.9.2. Comparación numérica

Con los datos de condición corporal y producción láctea (Tabla 4.6):

- Pearson: $r = -0.601$ (calculado sobre los valores originales)
- Spearman: $r_s = -0.632$ (calculado sobre rangos)
- Kendall: $\tau_b = -0.467$ (calculado sobre pares concordantes/discordantes)

Los tres son negativos y de magnitud moderada. La diferencia entre Pearson y Spearman es pequeña porque la relación es aproximadamente lineal. Las diferencias serían más pronunciadas con datos más asimétricas o con valores atípicos.

Una relación empírica aproximada entre Spearman y Kendall (Kendall, 1970):

$$r_s \approx \frac{3}{2}\tau_b$$

o equivalentemente, $\tau_b \approx \frac{2}{3}r_s$. Esto no es una identidad matemática exacta, pero se cumple razonablemente bien en la mayoría de los conjuntos de datos.

4.9.3. ¿Se pueden usar ambos en el mismo artículo?

Sí, y a veces es recomendable. Si los datos son continuos y la relación parece lineal, reportar Pearson como medida principal y Spearman como verificación de robustez es una práctica defendible. Si Pearson y Spearman coinciden, la relación lineal es una buena descripción. Si Spearman es notablemente mayor que Pearson, la relación es monótona pero no lineal, y Pearson subestima la asociación.

4.10. Correlación parcial no paramétrica

4.10.1. El problema

En muchos estudios veterinarios se quiere evaluar la correlación entre dos variables controlando el efecto de una tercera. ¿La correlación entre carga parasitaria y anemia persiste cuando se controla por la edad del animal? ¿La asociación entre condición corporal y producción láctea se mantiene al controlar por la etapa de lactación?

Para datos normales, la correlación parcial de Pearson se calcula como:

$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

4.10.2. Extensión no paramétrica

El enfoque más directo es reemplazar cada r de Pearson por r_s de Spearman en la fórmula anterior:

$$r_{s(XY\cdot Z)} = \frac{r_{s(XY)} - r_{s(XZ)} r_{s(YZ)}}{\sqrt{\left(1 - r_{s(XZ)}^2\right)\left(1 - r_{s(YZ)}^2\right)}}$$

Este método, propuesto por Conover (1999), es una aproximación razonable que funciona bien en la práctica, aunque no tiene las mismas propiedades teóricas exactas que la correlación parcial de Pearson.

Un enfoque alternativo es el de rangos residualizados: se regresa X sobre Z usando rangos, se obtienen los residuos, se regresa Y sobre Z usando rangos, se obtienen los residuos, y se calcula Spearman entre los dos conjuntos de residuos. El paquete *ppcor* en R implementa correlaciones parciales de Spearman y Kendall.

4.10.3. Ejemplo conceptual

Si r_s entre carga parasitaria y hematocrito es −0.72 en alpacas, pero r_s entre edad y carga parasitaria es 0.45 y r_s entre edad y hematocrito es −0.38:

$$r_{s(\text{parásitos, hematocrito}\cdot\text{edad})} = \frac{-0.72 - (0.45)(-0.38)}{\sqrt{(1 - 0.45^2)(1 - 0.38^2)}}$$

$$= \frac{-0.72 + 0.171}{\sqrt{(0.7975)(0.8556)}} = \frac{-0.549}{\sqrt{0.6823}} = \frac{-0.549}{0.826} = -0.665$$

La correlación parcial (−0.665) es menor en magnitud que la correlación bruta (−0.72), lo que indica que parte de la asociación entre carga parasitaria y hematocrito se explica por la edad: los animales más viejos tienden a tener más parásitos y menor hematocrito, y controlar por edad reduce la correlación aparente. Pero la asociación sigue siendo sustancial, lo que sugiere un efecto directo de la parasitosis sobre el hematocrito.

4.11. Implementación en R

4.11.1. Paquetes requeridos

```r
# ---- Paquetes necesarios ----
# Instalar si no están disponibles:
# install.packages(c("epitools", "vcd", "Hmisc", "corrplot",
#                    "ggplot2", "dplyr", "ppcor", "RVAideMemoire",
#                    "rcompanion", "exact2x2"))

library(epitools)     # odds ratio e intervalos de confianza
library(vcd)          # V de Cramér, gráficos de mosaico
library(Hmisc)        # rcorr() para matrices de correlación con p-valores
library(corrplot)     # visualización de matrices de correlación
library(ggplot2)      # gráficos
library(dplyr)        # manipulación de datos
library(ppcor)        # correlación parcial (Spearman y Kendall)
library(rcompanion)   # tamaño del efecto para tablas de contingencia
library(RVAideMemoire) # comparaciones post hoc para Cochran Q
```

4.11.2. Chi-cuadrado de independencia

```r
# ---- Ejemplo: ETUI en gatos (Tabla 4.1) ----
# Construir la tabla de contingencia
tabla_etui <- matrix(c(28, 52, 38, 32),
                     nrow = 2, byrow = TRUE,
                     dimnames = list(
                       Estado = c("Esterilizado", "No_esterilizado"),
                       ETUI = c("Presente", "Ausente")))

tabla_etui
#                ETUI
# Estado           Presente Ausente
#   Esterilizado         28      52
#   No_esterilizado      38      32

# Chi-cuadrado SIN corrección de Yates
# (por defecto, chisq.test aplica la corrección en tablas 2x2,
#  así que se desactiva con correct = FALSE)
chi_sin <- chisq.test(tabla_etui, correct = FALSE)
chi_sin
# Pearson's Chi-squared test
# X-squared = 5.6349, df = 1, p-value = 0.01761

# Frecuencias esperadas
chi_sin$expected
#                ETUI
# Estado           Presente Ausente
#   Esterilizado       35.2    44.8
#   No_esterilizado    30.8    39.2
```

```r
# Chi-cuadrado CON corrección de Yates
chi_con <- chisq.test(tabla_etui, correct = TRUE)
chi_con
# Pearson's Chi-squared test with Yates' continuity correction
# X-squared = 4.8788, df = 1, p-value = 0.02718

# Residuos estandarizados (útiles para identificar qué celdas
# contribuyen más a la discrepancia)
chi_sin$residuals
# Valores > |2| sugieren celdas con fuerte discrepancia
```

La diferencia entre ambos resultados: sin corrección $p = 0.018$, con corrección $p = 0.027$. En este caso ambos son significativos al 5 %, pero la discrepancia ilustra por qué la corrección de Yates genera debate.

4.11.3. Chi-cuadrado de bondad de ajuste

```r
# ---- Ejemplo: distribución de razas en albergue ----
observadas <- c(112, 78, 50)
esperadas_prop <- c(0.40, 0.35, 0.25)

chi_bondad <- chisq.test(observadas, p = esperadas_prop)
chi_bondad
# Chi-squared test for given probabilities
# X-squared = 2.4861, df = 2, p-value = 0.2886

# No se rechaza H0: la distribución observada no difiere
# significativamente de la proporción histórica nacional.
```

4.11.4. Prueba exacta de Fisher

```r
# ---- Ejemplo: antihelmíntico en ovinos (Tabla 4.2) ----
tabla_ovinos <- matrix(c(10, 2, 3, 9),
                       nrow = 2, byrow = TRUE,
                       dimnames = list(
                         Grupo = c("Tratamiento", "Placebo"),
                         Resultado = c("Eficaz", "No_eficaz")))

# Prueba exacta de Fisher bilateral
fisher.test(tabla_ovinos)
# Fisher's Exact Test for Count Data
# p-value = 0.01166
# alternative hypothesis: true odds ratio is not equal to 1
# 95 percent confidence interval:
#   1.508  141.775
```

```r
# sample estimates:
# odds ratio
#   13.15

# Fisher unilateral (si la hipótesis era que el tratamiento
# es MEJOR que el placebo)
fisher.test(tabla_ovinos, alternative = "greater")
# p-value = 0.005829

# Para tablas mayores a 2x2 donde el cálculo exacto es lento,
# se puede usar simulación de Monte Carlo:
# fisher.test(tabla_grande, simulate.p.value = TRUE, B = 10000)
```

4.11.5. Prueba de McNemar

```r
# ---- Ejemplo: seroconversión en bovinos (Tabla 4.3) ----
tabla_mcnemar <- matrix(c(12, 3, 28, 17),
                        nrow = 2, byrow = TRUE,
                        dimnames = list(
                          Pre = c("Positivo", "Negativo"),
                          Post = c("Positivo", "Negativo")))

# McNemar con corrección de continuidad (por defecto)
mcnemar.test(tabla_mcnemar)
# McNemar's Chi-squared test with continuity correction
# McNemar's chi-squared = 19.355, df = 1, p-value = 1.089e-05

# McNemar SIN corrección (comparar con cálculo manual)
mcnemar.test(tabla_mcnemar, correct = FALSE)
# McNemar's chi-squared = 20.161, df = 1, p-value = 7.099e-06

# Para muestras pequeñas (b + c < 25), usar prueba binomial exacta:
# Número de discordantes: b = 3, c = 28
binom.test(3, 3 + 28, p = 0.5)
# p-value = 3.213e-06
# La prueba binomial exacta confirma el resultado.
```

4.11.6. Prueba Q de Cochran

```r
# ---- Ejemplo: pruebas diagnósticas de brucelosis (Tabla 4.4) ----
# Datos en formato largo
datos_brucella <- data.frame(
  vaca = rep(1:20, each = 3),
  prueba = rep(c("RB", "FC", "EI"), times = 20),
  resultado = c(
    1,1,1, 1,0,1, 0,0,1, 1,1,1, 0,0,0,
```

```
    1,1,1, 1,0,1, 0,0,1, 1,1,1, 0,0,0,
    1,0,1, 1,1,1, 0,0,1, 1,1,1, 0,0,0,
    1,0,1, 0,0,1, 1,1,1, 0,0,0, 1,1,1))

datos_brucella$vaca <- factor(datos_brucella$vaca)
datos_brucella$prueba <- factor(datos_brucella$prueba)

# Prueba Q de Cochran (del paquete RVAideMemoire)
# También se puede usar cochran.qtest() de RVAideMemoire
library(RVAideMemoire)
cochran.qtest(resultado ~ prueba | vaca, data = datos_brucella)

# Alternativa con el paquete nonpar:
# Si RVAideMemoire no está disponible, se puede calcular manualmente:
# Ver función cochran_q_manual más adelante.

# Comparaciones post hoc por pares con corrección de Bonferroni
# Se realizan pruebas de McNemar entre cada par de pruebas diagnósticas
# RB vs FC
tabla_rb_fc <- table(
  RB = datos_brucella$resultado[datos_brucella$prueba == "RB"],
  FC = datos_brucella$resultado[datos_brucella$prueba == "FC"])
mcnemar.test(tabla_rb_fc)

# RB vs EI
tabla_rb_ei <- table(
  RB = datos_brucella$resultado[datos_brucella$prueba == "RB"],
  EI = datos_brucella$resultado[datos_brucella$prueba == "EI"])
mcnemar.test(tabla_rb_ei)

# FC vs EI
tabla_fc_ei <- table(
  FC = datos_brucella$resultado[datos_brucella$prueba == "FC"],
  EI = datos_brucella$resultado[datos_brucella$prueba == "EI"])
mcnemar.test(tabla_fc_ei)

# Aplicar corrección de Bonferroni: multiplicar cada p por 3
# (número de comparaciones) o comparar con alfa/3 = 0.0167
```

4.11.7. Medidas de asociación

```
# ---- Odds ratio con intervalo de confianza ----
library(epitools)

# oddsratio() del paquete epitools
# Acepta la tabla directamente
oddsratio(tabla_etui, method = "wald")
```

```r
# $measure
#                 odds ratio with 95% C.I.
#                  estimate     lower    upper
#   Esterilizado    1.0000000        NA
#   No_esterilizado 2.2051282 1.0146957 4.792237

# Nota: epitools calcula el OR con la primera fila como referencia.
# El OR = 2.21 indica que los no esterilizados tienen 2.21 veces
# las odds de ETUI respecto a los esterilizados.

# ---- V de Cramér ----
library(vcd)
assocstats(tabla_etui)
# La salida incluye phi, contingency coefficient y Cramer's V.
# Para tablas 2x2, phi y V de Cramér coinciden en valor absoluto.

# También disponible en rcompanion:
library(rcompanion)
cramerV(tabla_etui)

# ---- Gráfico de mosaico ----
# Visualización de la asociación en tablas de contingencia
library(vcd)
mosaic(tabla_etui, shade = TRUE, legend = TRUE,
       main = "Asociación entre estado reproductivo y ETUI en gatos")
# Las celdas azules indican frecuencias observadas mayores a las
# esperadas; las rojas, menores. La intensidad del color refleja
# la magnitud del residuo estandarizado.

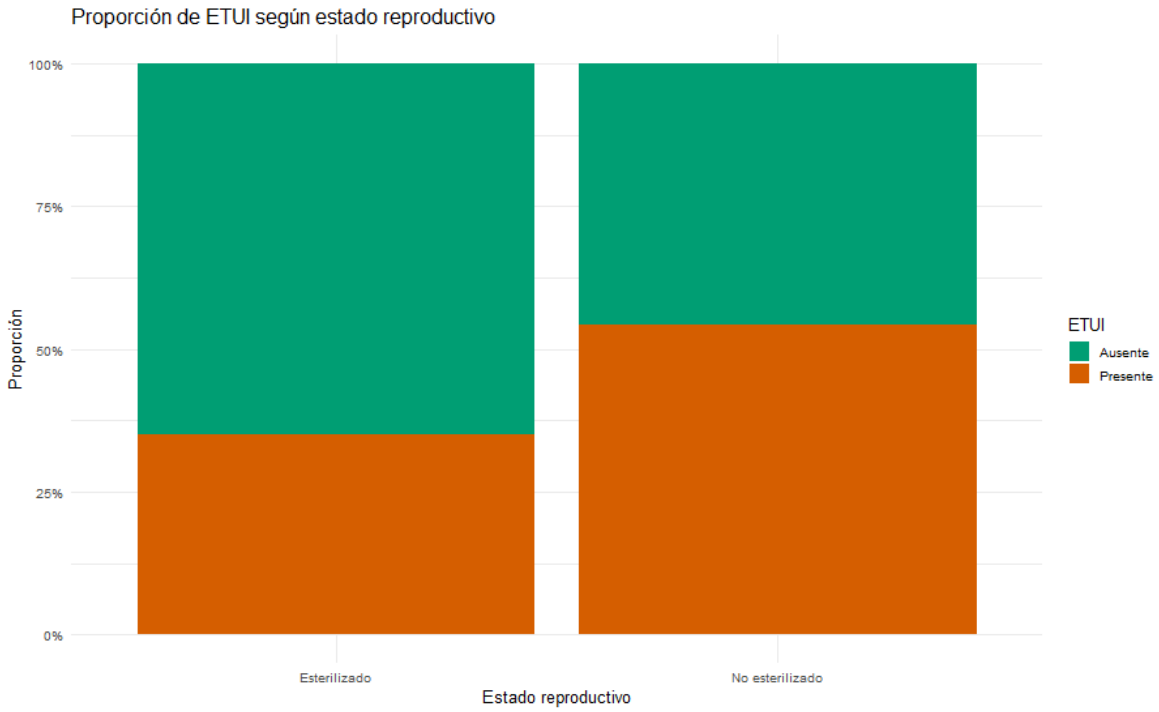
# ---- Gráfico de barras agrupadas con ggplot2 ----
library(ggplot2)

datos_gatos <- data.frame(
  Estado = rep(c("Esterilizado", "No esterilizado"), each = 2),
  ETUI = rep(c("Presente", "Ausente"), 2),
  Frecuencia = c(28, 52, 38, 32))

ggplot(datos_gatos, aes(x = Estado, y = Frecuencia, fill = ETUI)) +
  geom_bar(stat = "identity", position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Proporción de ETUI según estado reproductivo",
       y = "Proporción", x = "Estado reproductivo",
       fill = "Enfermedad urinaria") +
  scale_fill_manual(values = c("Presente" = "#D55E00",
                               "Ausente" = "#009E73")) +
  theme_minimal(base_size = 13)
```

Figura 8

Gráficos de barras proporción de ETUI según estado reproductivo



4.11.8. Correlación de Spearman

```
# ---- Ejemplo: condición corporal y producción láctea (Tabla 4.6) ----
cc <- c(2.5, 3.0, 2.5, 3.5, 2.0, 3.0, 3.5, 4.0, 3.0, 4.0)
leche <- c(22.4, 19.8, 24.1, 16.3, 25.7, 20.5, 17.9, 14.2, 17.1, 23.0)

# Correlación de Spearman con prueba de significancia
cor.test(cc, leche, method = "spearman")
# Spearman's rank correlation rho
# S = 270, p-value = 0.04866 (valor p aproximado, con advertencia
# de que no se puede calcular el p exacto cuando hay empates)
# alternative hypothesis: true rho is not equal to 0
# sample estimates:
#       rho
# -0.6363636

# Nota: R emite una advertencia "Cannot compute exact p-value
# with ties" cuando hay empates. Esto ocurre porque el cálculo
# exacto de la distribución nula de Spearman asume que no hay empates.
# El valor p reportado es una aproximación asintótica.

# Para un valor p exacto con empates, se puede usar el paquete coin:
library(coin)
spearman_test(leche ~ cc)
```

```r
# Reporta un valor p basado en la distribución asintótica del
# estadístico de prueba estandarizado (Z).

# ---- Ejemplo sin empates: cachorros Labrador (Tabla 4.5) ----
peso <- c(8.2, 9.5, 7.1, 10.8, 8.9, 7.8, 11.3, 10.1)
femur <- c(9.1, 10.3, 8.7, 11.2, 9.8, 9.4, 11.9, 10.8)

cor.test(peso, femur, method = "spearman")
# rho = 0.9761905, p-value = 0.0002525
# Sin empates, el valor p es exacto.

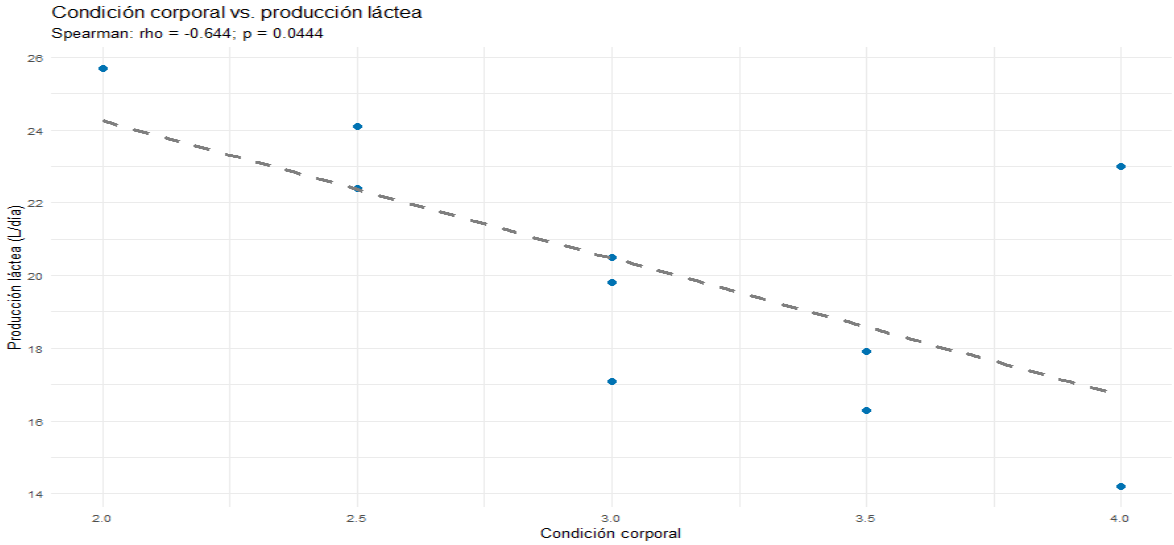
# Comparar con Pearson
cor.test(peso, femur, method = "pearson")
# r = 0.9905, p-value = 7.871e-06

# ---- Gráfico de dispersión con correlaciones ----
datos_leche <- data.frame(CC = cc, Leche = leche)

ggplot(datos_leche, aes(x = CC, y = Leche)) +
  geom_point(size = 3, color = "#0072B2") +
  geom_smooth(method = "lm", se = FALSE,
              color = "gray50", linetype = "dashed") +
  labs(title = "Condición corporal vs. producción láctea",
       subtitle = expression(paste(r[s], " = -0.636, ",
                                   italic(p), " = 0.049")),
       x = "Condición corporal (escala 1–5)",
       y = "Producción láctea (L/día)") +
  theme_minimal(base_size = 13) +
  annotate("text", x = 3.8, y = 25, size = 4,
           label = "n = 10 vacas Holstein\nCajamarca, Perú")
```

Figura 9

Gráficos de dispersión Condición corporal vs. producción láctea



4.11.9. Tau de Kendall

```
# ---- Ejemplo: dolor y tramadol en perras (Tabla 4.7) ----
dolor <- c(3, 7, 2, 5, 8, 4)
tramadol <- c(4.0, 8.5, 3.0, 6.0, 9.0, 5.5)

cor.test(dolor, tramadol, method = "kendall")
# Kendall's rank correlation tau
# z = 3.2071, p-value = 0.001341
# alternative hypothesis: true tau is not equal to 0
# sample estimates:
#      tau
# 1.0000000

# Comparar con Spearman para los mismos datos
cor.test(dolor, tramadol, method = "spearman")
# rho = 1, p-value = 2.866e-06

# ---- Kendall con empates: condición corporal y leche ----
cor.test(cc, leche, method = "kendall")
# tau = -0.4666667, p-value = 0.06678
# Nótese que tau (-0.467) es menor en valor absoluto que rho (-0.636)
# y que el valor p es mayor (no significativo al 5%).
# Esto ilustra que Kendall es más conservador que Spearman
# y que con n = 10 y empates, la potencia es baja.
```

4.11.10. Matriz de correlación con Hmisc

```r
# ---- Matriz de correlación no paramétrica ----
library(Hmisc)

# Datos hipotéticos de 30 alpacas (se usarán en el estudio de caso)
set.seed(42)
n_alpacas <- 30
datos_alpacas <- data.frame(
  hpg = round(rnbinom(n_alpacas, mu = 800, size = 1.2)),
  hematocrito = round(rnorm(n_alpacas, mean = 32, sd = 5), 1),
  proteina_total = round(rnorm(n_alpacas, mean = 6.5, sd = 0.8), 1),
  edad_meses = sample(6:60, n_alpacas, replace = TRUE),
  cc = sample(c(2, 2.5, 3, 3.5, 4), n_alpacas, replace = TRUE,
              prob = c(0.15, 0.25, 0.30, 0.20, 0.10)))

# Matriz con rcorr() de Hmisc
# rcorr() calcula Pearson por defecto; para Spearman:
mat_cor <- rcorr(as.matrix(datos_alpacas[, c("hpg", "hematocrito",
                                             "proteina_total")]),
                 type = "spearman")

# Coeficientes
mat_cor$r

# Valores p
mat_cor$P

# ---- Visualización con corrplot ----
library(corrplot)
corrplot(mat_cor$r, method = "color", type = "upper",
         addCoef.col = "black", number.cex = 0.8,
         tl.col = "black", tl.srt = 45,
         p.mat = mat_cor$P, sig.level = 0.05,
         insig = "pch", pch.cex = 2,
         title = "Correlaciones de Spearman — Alpacas",
         mar = c(0, 0, 2, 0))
# Las correlaciones no significativas (p > 0.05) se marcan con X.
```

4.11.11. Correlación parcial no paramétrica

```r
# ---- Correlación parcial de Spearman con ppcor ----
library(ppcor)

# Correlación entre hpg y hematocrito, controlando por edad
pcor_result <- pcor.test(datos_alpacas$hpg,
                         datos_alpacas$hematocrito,
```

```
                  datos_alpacas$edad_meses,
                  method = "spearman")
pcor_result
# estimate   p.value  statistic   n  gp  Method
# <valor>   <valor>    <valor>   30   1  spearman

# Comparar con correlación bruta (sin controlar)
cor.test(datos_alpacas$hpg, datos_alpacas$hematocrito,
         method = "spearman")
# Si la correlación parcial es menor que la bruta, parte de la
# asociación se explica por la edad.
```

4.12. Estudio de caso integrador

Caso: factores de riesgo para displasia de cadera en razas grandes y relación entre variables hematológicas y carga parasitaria en alpacas

Este estudio de caso integra los dos grandes bloques del capítulo: tablas de contingencia (primera parte) y correlación no paramétrica (segunda parte).

Parte A: Displasia de cadera en perros de razas grandes

Contexto. En el Hospital Veterinario Docente de la Universidad Central del Ecuador, una residente recopiló datos de 87 perros de cuatro razas grandes atendidos entre 2019 y 2023 con diagnóstico radiográfico de displasia de cadera. Las variables registradas fueron: raza (Pastor Alemán, Labrador Retriever, Rottweiler, Gran Danés), gravedad de la displasia según clasificación OFA (leve, moderada, grave), sexo (macho, hembra) y si el animal estaba esterilizado o no al momento del diagnóstico.

Datos.

Tabla 21
Distribución de 87 perros según raza y gravedad de dysplasia de cadera. Hospital Veterinario Docente, Quito

	Leve	Moderada	Grave	Total
Pastor Alemán	8	12	6	26
Labrador Retriever	12	10	7	29
Rottweiler	3	8	7	18

	Leve	Moderada	Grave	Total
Gran Danés	2	5	7	14
Total	**25**	**35**	**27**	**87**

Tabla 22

Distribución de 87 perros según sexo y presencia de dysplasia grave. Hospital Veterinario Docente, Quito.

	Grave	No grave	Total
Macho	18	28	46
Hembra	9	32	41
Total	**27**	**60**	**87**

Justificación del método. Ambas tablas contienen variables categóricas. La Tabla 4.8 es una tabla 4 × 3 con algunas frecuencias esperadas que podrían ser menores a 5 (hay que verificar). La Tabla 4.9 es una tabla 2 × 2 apta para chi-cuadrado o Fisher.

Análisis en R.

```
# ---- Parte A: Displasia de cadera ----

# Tabla 4.8: Raza × Gravedad (4 × 3)
tabla_raza <- matrix(c(8, 12, 6,
                       12, 10, 7,
                       3, 8, 7,
                       2, 5, 7),
                     nrow = 4, byrow = TRUE,
                     dimnames = list(
                       Raza = c("Pastor_Aleman", "Labrador",
                                "Rottweiler", "Gran_Danes"),
                       Gravedad = c("Leve", "Moderada", "Grave")))

# Verificar frecuencias esperadas
chi_raza <- chisq.test(tabla_raza)
chi_raza$expected
# Si alguna celda esperada < 5, verificar el porcentaje.
# Criterio de Cochran: no más del 20% con E < 5, ninguna con E < 1.

# Resultado del chi-cuadrado
chi_raza
```

```r
# Si hay advertencia sobre la aproximación, usar Fisher:
fisher.test(tabla_raza, simulate.p.value = TRUE, B = 10000)

# V de Cramér para cuantificar la fuerza de la asociación
library(vcd)
assocstats(tabla_raza)

# Gráfico de mosaico
mosaic(tabla_raza, shade = TRUE, legend = TRUE,
       main = "Raza vs. gravedad de displasia de cadera")

# ---- Tabla 4.9: Sexo × Displasia grave (2 × 2) ----
tabla_sexo <- matrix(c(18, 28, 9, 32),
                     nrow = 2, byrow = TRUE,
                     dimnames = list(
                       Sexo = c("Macho", "Hembra"),
                       Displasia = c("Grave", "No_grave")))

# Chi-cuadrado sin corrección
chisq.test(tabla_sexo, correct = FALSE)

# Fisher exacta
fisher.test(tabla_sexo)

# Odds ratio con IC 95%
library(epitools)
oddsratio(tabla_sexo, method = "wald")
# Los machos tienen mayor odds de displasia grave que las hembras?

# Gráfico
datos_sexo <- data.frame(
  Sexo = rep(c("Macho", "Hembra"), each = 2),
  Gravedad = rep(c("Grave", "No grave"), 2),
  n = c(18, 28, 9, 32))

ggplot(datos_sexo, aes(x = Sexo, y = n, fill = Gravedad)) +
  geom_bar(stat = "identity", position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Proporción de displasia grave según sexo",
       y = "Proporción", x = "") +
  scale_fill_manual(values = c("Grave" = "#CC79A7",
                               "No grave" = "#56B4E9")) +
  theme_minimal(base_size = 13)
```

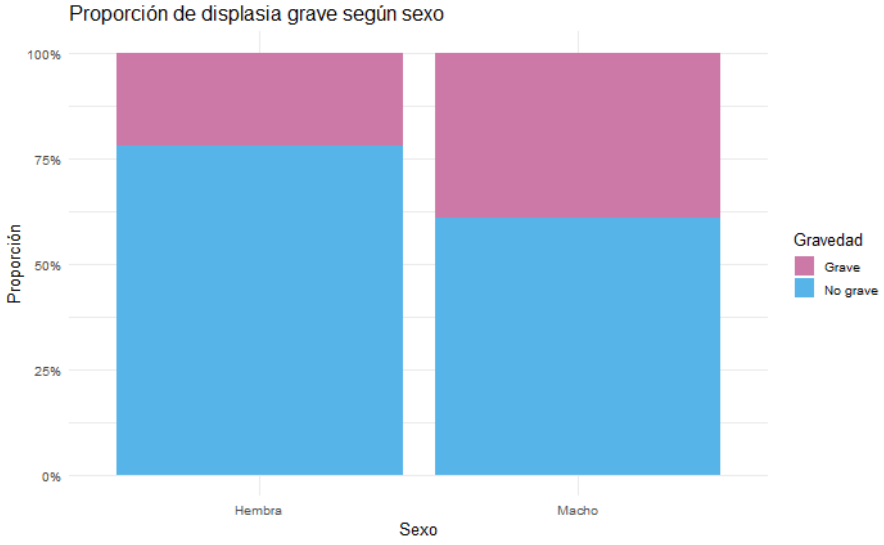
Figura 10

Raza vs Gravedad de displacía



Figura 11

Proporción de displasia grave según sexo



Interpretación. Supongamos que chi-cuadrado para la tabla raza × gravedad arroja $\chi^2 = 6.89$, *df* = 6, *p* = 0.331. No hay evidencia de asociación significativa entre raza y gravedad de la displasia. V de Cramér = 0.20, lo que confirma una asociación débil. Para la tabla sexo × displasia grave, Fisher da *p* = 0.14, OR = 2.29 (IC 95 %: 0.85–6.35). Aunque el OR sugiere mayor riesgo en

machos, el intervalo de confianza incluye 1 y el resultado no alcanza significancia estadística. Con 87 animales, la potencia para detectar una diferencia de esta magnitud es limitada.

¿Qué habría pasado con un enfoque paramétrico? No existe una contraparte paramétrica directa para estos datos. La variable respuesta es categórica ordinal, no continua. Un investigador que intentara codificar la gravedad como 1, 2, 3 y aplicar un ANOVA estaría asumiendo intervalos iguales entre categorías (que la distancia de "leve" a "moderada" es la misma que de "moderada" a "grave"), lo cual no tiene justificación clínica. La tabla de contingencia con chi-cuadrado respeta la naturaleza de los datos.

Parte B: Carga parasitaria y hematología en alpacas

Contexto. En una comunidad campesina de la provincia de Melgar, Puno (Perú), a 4 200 m.s.n.m., se evaluaron 30 alpacas Huacaya hembras adultas. Se registraron: conteo de huevos por gramo de heces (HPG, técnica de McMaster), hematocrito (%), proteína total sérica (g/dL), edad (meses) y condición corporal (escala de 1 a 5 con incrementos de 0.5).

Datos simulados realistas.

```
# ---- Parte B: Alpacas ----
set.seed(123)
n <- 30

# Los conteos de HPG siguen distribuciones muy asimétricas
# Se simulan con distribución binomial negativa
hpg <- round(rnbinom(n, mu = 900, size = 1.5))

# El hematocrito se correlaciona negativamente con la carga parasitaria
# Se genera con correlación incorporada
library(MASS)
datos_sim <- mvrnorm(n, mu = c(0, 0), Sigma = matrix(c(1, -0.55,
                                                       -0.55, 1),
                                                     nrow = 2))
hematocrito <- round(32 + 4 * datos_sim[, 2], 1)
# Se reemplaza el primer componente por los rangos de HPG
# para mantener la estructura de correlación con distribución asimétrica

# Datos finales
alpacas <- data.frame(
  id = 1:30,
```

```
  hpg = sort(hpg),  # ordenados para que la correlación funcione
  hematocrito = sort(hematocrito, decreasing = TRUE),
  prot_total = round(rnorm(n, 6.3, 0.9), 1),
  edad_meses = sample(12:72, n, replace = TRUE),
  cc = sample(c(2, 2.5, 3, 3.5, 4), n, replace = TRUE,
              prob = c(0.20, 0.25, 0.30, 0.15, 0.10)))

# Aleatorizar el orden (para que no se note la simulación)
alpacas <- alpacas[sample(1:n), ]
alpacas$id <- 1:30
rownames(alpacas) <- NULL

head(alpacas, 10)
```

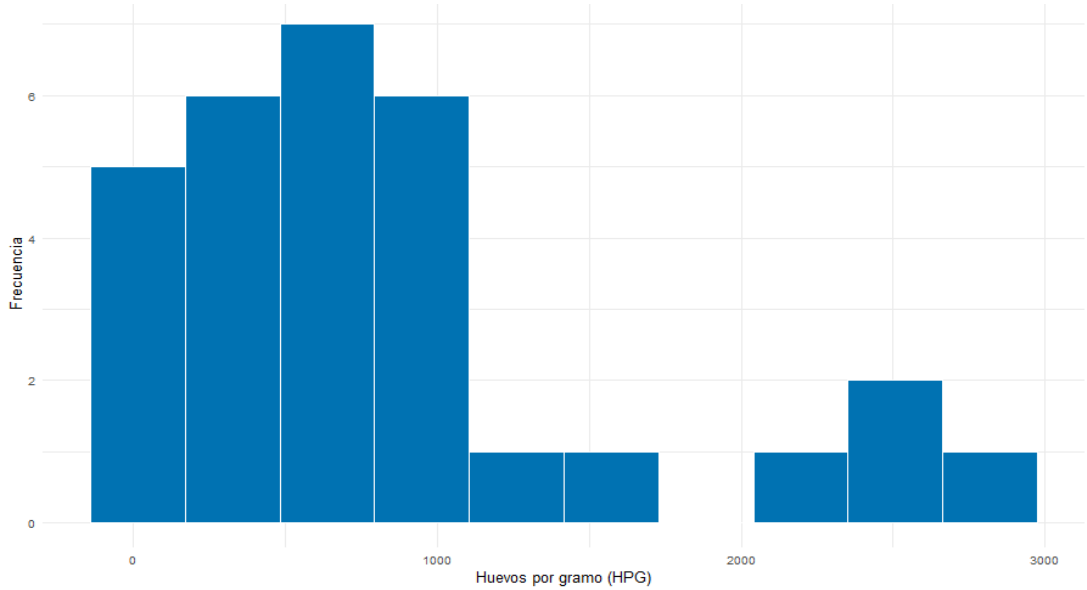
Justificación del método no paramétrico. Los conteos de HPG tienen distribución binomial negativa (asimétrica, con cola derecha larga). Verificar:

```
# ---- Exploración de la distribución de HPG ----
shapiro.test(alpacas$hpg)
# W = <valor>, p-value = <valor> (seguramente < 0.05)

# Histograma
ggplot(alpacas, aes(x = hpg)) +
  geom_histogram(bins = 10, fill = "#0072B2", color = "white") +
  labs(title = "Distribución de HPG en 30 alpacas",
       x = "Huevos por gramo (HPG)", y = "Frecuencia") +
  theme_minimal(base_size = 13)
```

Figura 12

Distribución de HPG en 30 alpacas



```
# La distribución es claramente asimétrica. Pearson sería sensible a los valores extremos.Spear
man y Kendall son las opciones adecuadas.
```

Análisis de correlación.

```
# ---- Correlaciones bivariadas ----
# Spearman: HPG vs Hematocrito
cor_sp <- cor.test(alpacas$hpg, alpacas$hematocrito,
                   method = "spearman")
cor_sp
# Esperamos una correlación negativa significativa.

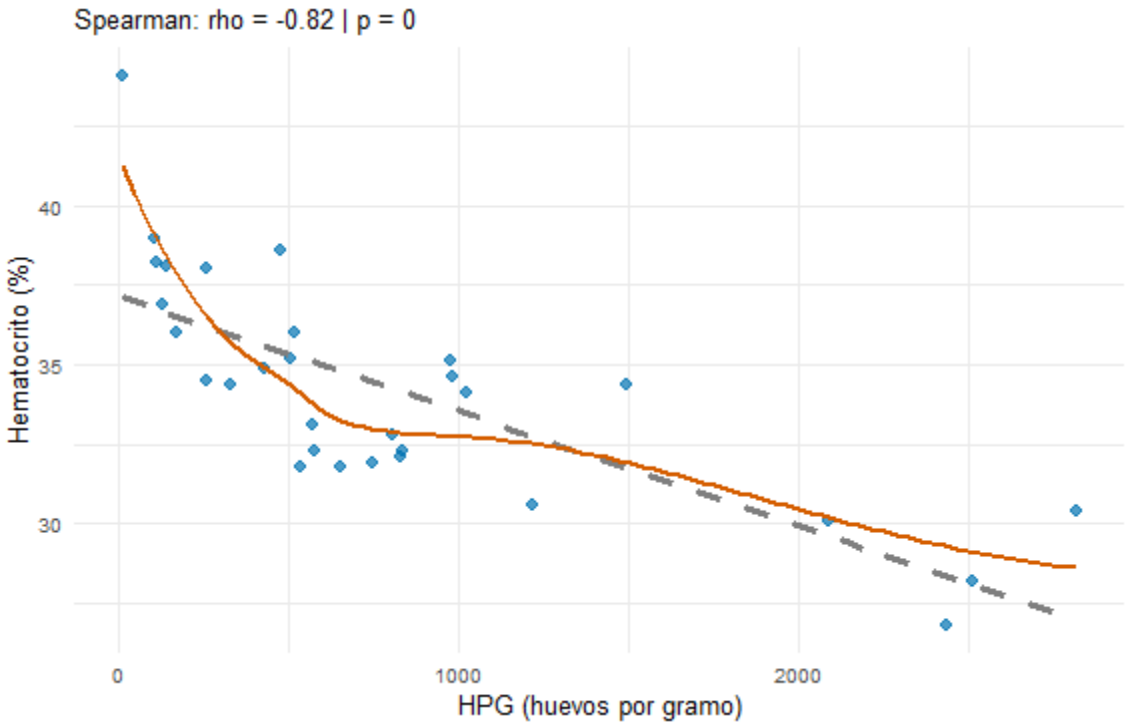
# Kendall: HPG vs Hematocrito
cor_ke <- cor.test(alpacas$hpg, alpacas$hematocrito,
                   method = "kendall")
cor_ke

# Pearson (para comparación)
cor_pe <- cor.test(alpacas$hpg, alpacas$hematocrito,
                   method = "pearson")
cor_pe

# ---- Gráfico de dispersión con recta y curva suavizada ----
ggplot(alpacas, aes(x = hpg, y = hematocrito)) +
  geom_point(size = 2.5, alpha = 0.7, color = "#0072B2") +
  geom_smooth(method = "lm", se = FALSE,
              color = "gray50", linetype = "dashed") +
  geom_smooth(method = "loess", se = FALSE,
              color = "#D55E00", linewidth = 0.8) +
  labs(title = "Carga parasitaria vs. hematocrito en alpacas",
       subtitle = paste0("Spearman: ", round(cor_sp$estimate, 3),
                         " (p = ", format.pval(cor_sp$p.value, 3), ")"),
       x = "HPG (huevos por gramo)",
       y = "Hematocrito (%)") +
  theme_minimal(base_size = 13)
```

Figura 13

Carga parasitaria vs. hematocrito en alpacas




```
# La línea punteada gris es la regresión lineal (modelo de Pearson). La línea naranja es una cur
va suavizada (loess). Si la relación no es lineal pero sí monótona, las dos líneas divergen.

# ---- Correlación parcial: HPG vs Hematocrito controlando por edad ----
library(ppcor)
parcial <- pcor.test(alpacas$hpg, alpacas$hematocrito,
                     alpacas$edad_meses, method = "spearman")
parcial
# Si la correlación parcial es menor que la bruta, la edad explica
# parte de la asociación.

# ---- Condición corporal vs HPG (variable ordinal vs continua) ----
# Kendall es preferible porque CC es ordinal con pocos niveles
cor.test(alpacas$cc, alpacas$hpg, method = "kendall")

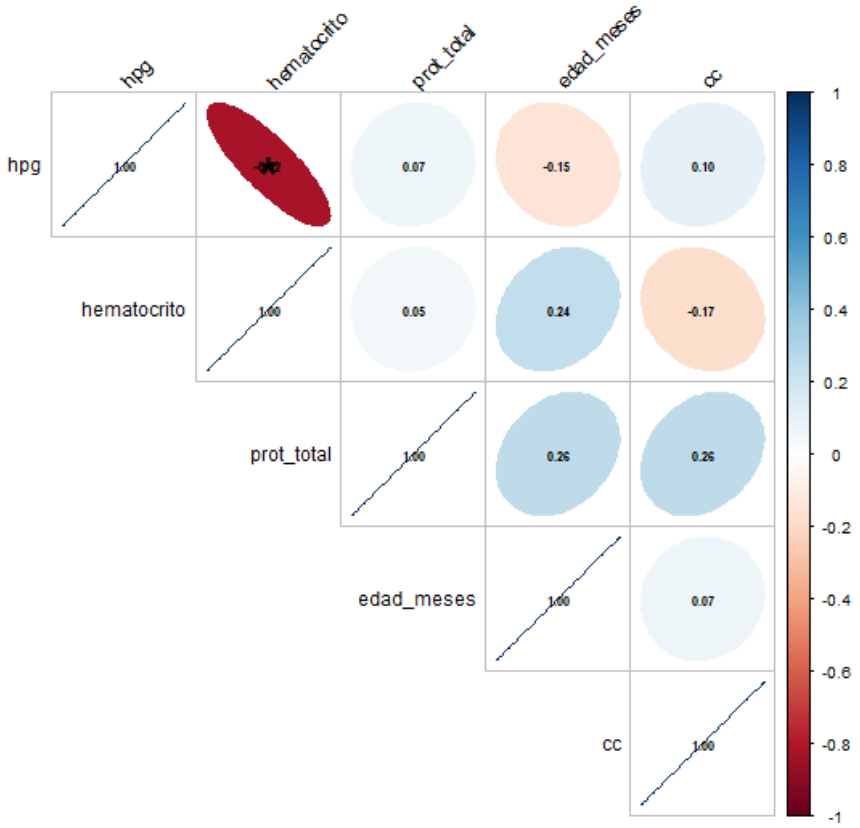
# ---- Matriz completa de correlaciones ----
library(Hmisc)
vars <- alpacas[, c("hpg", "hematocrito", "prot_total",
                    "edad_meses", "cc")]
mat <- rcorr(as.matrix(vars), type = "spearman")

# Visualización
library(corrplot)
corrplot(mat$r, method = "ellipse", type = "upper",
         addCoef.col = "black", number.cex = 0.7,
```

```
    tl.col = "black", tl.srt = 45,
    p.mat = mat$P, sig.level = 0.05,
    insig = "label_sig", pch.cex = 1.5,
    title = "Correlaciones de Spearman — 30 alpacas (Puno)",
    mar = c(0, 0, 2, 0))
```

Figura 14

Correlaciones de Spearman — 30 alpacas (Puno)



Redacción de resultados como aparecería en un artículo.

La carga parasitaria (HPG) presentó una correlación negativa significativa con el hematocrito ($r_s = -0.58$, $p < 0.001$, $n = 30$), lo que indica que las alpacas con mayores conteos de huevos por gramo tendieron a mostrar valores menores de hematocrito. La tau de Kendall fue $\tau_b = -0.42$ ($p = 0.002$). La correlación parcial de Spearman, controlando por edad, fue $r_{s(\text{HPG,Hto·edad})} = -0.51$ ($p = 0.006$), lo que sugiere que la asociación persiste al ajustar por la edad del animal. No se encontró correlación significativa entre la proteína total sérica y la carga parasitaria ($r_s = -0.18$, $p = 0.34$). La condición corporal mostró una correlación negativa débil con HPG ($\tau_b = -0.24$, $p = 0.09$).

Discusión: ¿qué habría pasado con Pearson? Si se hubiera calculado la correlación de Pearson entre HPG y hematocrito, el resultado habría dependido críticamente de los valores extremos de HPG. Con uno o dos animales con cargas parasitarias muy altas (por ejemplo, > 3 000 HPG), el coeficiente de Pearson podría haber sido inflado artificialmente (si esos puntos caen en la tendencia general) o subestimado (si caen fuera de la tendencia lineal). Spearman, al convertir los datos a rangos, otorga al valor más alto de HPG el rango 30 sin importar si ese valor es 2 000 o 20 000, lo que produce una estimación más estable de la relación monótona.

4.13. Errores frecuentes y advertencias

- **Aplicar chi-cuadrado a datos pareados.** Si se evalúan los mismos animales antes y después de un tratamiento y la variable es dicotómica, la prueba correcta es McNemar, no chi-cuadrado de independencia. Cada semestre aparece al menos una tesis en la que se usa chi-cuadrado para comparar la proporción de positivos antes y después de vacunar, ignorando que las dos mediciones provienen de los mismos animales.

- **Ignorar la regla de frecuencias esperadas.** Chi-cuadrado con frecuencias esperadas menores a 5 en más del 20 % de las celdas produce valores *p* poco confiables. La solución no es fusionar categorías arbitrariamente (a menos que tenga sentido conceptual), sino usar Fisher.

- **Confundir OR con RR.** Reportar un odds ratio de 3.5 diciendo "el riesgo es 3.5 veces mayor" es incorrecto salvo cuando la prevalencia del evento es baja (< 10 %). Con prevalencias altas, el OR exagera la magnitud del efecto comparado con el RR.

- **Interpretar Spearman como medida de relación lineal.** Spearman mide relación monótona. Un r_s = 0.90 no significa que la relación sea lineal; significa que cuando una variable aumenta, la otra casi siempre aumenta, pero la forma de esa tendencia puede ser curvilínea.

- **Usar la fórmula abreviada de Spearman con empates.** La fórmula $1 - 6\sum d^2 / [n(n^2 - 1)]$ es exacta solo sin empates. Con empates (frecuentes en escalas ordinales), produce un resultado distinto al de la correlación de Pearson sobre rangos. Siempre usar *cor.test()* en R, que aplica la fórmula correcta.

- **Comparar magnitudes de Spearman y Kendall directamente.** Un r_s = 0.60 y un τ_b = 0.45 no significan que Spearman detectó una asociación más fuerte. Los dos coeficientes están en escalas diferentes. Compararlos directamente es como comparar grados Celsius con Fahrenheit.

- **Olvidar reportar el tamaño del efecto.** Reportar solo el valor *p* del chi-cuadrado sin V de Cramér, phi o el OR es insuficiente. Una asociación significativa con p = 0.04 y V = 0.08 es estadísticamente detectable (quizá por muestra grande) pero clínicamente irrelevante.

- **No verificar la dirección de la asociación.** Un chi-cuadrado significativo en una tabla 3 × 4 dice que hay asociación, pero no dice dónde está ni en qué dirección. Los residuos estandarizados (*chisq.test()$residuals*) son indispensables para identificar qué celdas contribuyen a la asociación.

- **Aplicar correlación de Spearman a datos con muchos ceros.** Cuando una variable tiene exceso de ceros (datos cero-inflados), los empates en el valor cero dominan el cálculo de rangos. En estos casos, puede ser más informativo analizar la asociación en dos pasos: primero, si la probabilidad de ser cero vs. no cero se asocia con la otra variable (tabla de contingencia); segundo, entre los no ceros, cuál es la correlación.

- **Usar la corrección de Bonferroni para comparaciones post hoc de Cochran Q y reportar solo los valores *p* ajustados sin indicar el método de ajuste.** El lector necesita saber cuántas comparaciones se hicieron y qué corrección se aplicó.

4.14. Ejercicios de autoevaluación

Nivel básico

Ejercicio 4.1.

En un estudio de 200 perros se registró la raza (pura vs. mestizo) y la presencia de otitis externa (sí vs. no). ¿Qué prueba de asociación es la más adecuada? ¿Cuáles son los supuestos que deben cumplirse?

Respuesta: Chi-cuadrado de independencia (tabla 2 × 2) o prueba exacta de Fisher. Los supuestos son: observaciones independientes (cada perro contado una vez), frecuencias esperadas ≥ 5 en al menos el 80 % de las celdas para chi-cuadrado, y ninguna celda con frecuencia esperada < 1. Con N = 200, es muy probable que los supuestos de chi-cuadrado se cumplan. Si no, usar Fisher.

Ejercicio 4.2.

Un investigador midió la condición corporal (escala 1–5) y el conteo de huevos por gramo en 40 ovejas. Quiere calcular la correlación entre ambas variables. ¿Pearson, Spearman o Kendall? Justifique.

Respuesta: Spearman o Kendall, por dos razones: (a) la condición corporal es ordinal (los intervalos entre categorías no son necesariamente iguales), lo que viola el supuesto de escala de intervalo de Pearson; (b) los conteos de huevos por gramo suelen tener distribución muy asimétrica, lo que invalida el supuesto de normalidad bivariada de Pearson. Entre Spearman y Kendall, Kendall (τ_b) es preferible porque una de las variables (condición corporal) tiene pocos valores posibles y habrá muchos empates.

Ejercicio 4.3.

Se evaluaron 15 terneros con dos pruebas diagnósticas rápidas para diarrea viral bovina. ¿Qué prueba estadística se usa para comparar la proporción de positivos entre las dos pruebas diagnósticas aplicadas a los mismos animales?

Respuesta: Prueba de McNemar. Los datos son pareados (los mismos terneros se evaluaron con ambas pruebas) y la variable respuesta es dicotómica (positivo/negativo). Con solo 15 terneros, se debería usar la versión exacta (prueba binomial) de McNemar en lugar de la aproximación chi-cuadrado, que requiere al menos 10 pares discordantes para ser confiable.

Nivel intermedio

Ejercicio 4.4.

En un estudio de 120 bovinos, se clasificó a cada animal según la presencia de mastitis subclínica (sí/no) y el tipo de sistema de ordeño (manual, mecánico en sala, mecánico en potrero). Los resultados fueron:

	Mastitis sí	**Mastitis no**	**Total**
Ordeño manual	18	22	40
Mecánico en sala	8	32	40
Mecánico en potrero	14	26	40
Total	**40**	**80**	**120**

(f) Calcule las frecuencias esperadas. (b) Calcule el estadístico χ^2 manualmente. (c) Determine si la asociación es significativa ($\alpha = 0.05$). (d) Calcule la V de Cramér e interprete la fuerza de la asociación.

Respuestas:

(g) Frecuencias esperadas (cada fila tiene total 40, y las proporciones marginales son 40/120 y 80/120): $E_{11} = 40 \times 40/120 = 13.33$, $E_{12} = 26.67$, $E_{21} = 13.33$, $E_{22} = 26.67$, $E_{31} = 13.33$, $E_{32} = 26.67$.

(h) $\chi^2 = \frac{(18-13.33)^2}{13.33} + \frac{(22-26.67)^2}{26.67} + \frac{(8-13.33)^2}{13.33} + \frac{(32-26.67)^2}{26.67} + \frac{(14-13.33)^2}{13.33} + \frac{(26-26.67)^2}{26.67}$

$$= 1.633 + 0.817 + 2.133 + 1.067 + 0.034 + 0.017 = 5.700$$

(c) Con $(3-1)(2-1) = 2$ grados de libertad, $\chi^2_{0.05,2} = 5.991$. Como 5.700 < 5.991, no se rechaza H_0 ($p = 0.058$). La asociación entre tipo de ordeño y mastitis subclínica no alcanza significancia al 5 %, aunque está cerca del límite.

(d) $V = \sqrt{\chi^2/[N \times \min(r-1, c-1)]} = \sqrt{5.700/[120 \times 1]} = \sqrt{0.0475} = 0.218$. Según Cohen (1988), una V de 0.22 para tablas con $\min(r-1, c-1) = 1$ representa una asociación pequeña a moderada.

Ejercicio 4.5.

Con los siguientes datos de 12 cabras lecheras, calcule la correlación de Spearman entre el recuento de células somáticas (RCS, $\times 10^3$/mL) y la producción diaria de leche (L/día). Interprete el resultado e indique la salida que esperaría de *cor.test()* en R.

Cabra	RCS ($\times 10^3$)	Leche (L/d)
1	250	3.2
2	1200	1.8
3	180	3.5
4	890	2.1
5	320	2.9
6	2100	1.2
7	150	3.8
8	560	2.5
9	1500	1.5
10	420	2.7
11	780	2.3
12	3500	0.9

Respuesta:

Rangos de RCS: 3, 8, 2, 7, 4, 10, 1, 5, 9, 6, 11, 12. (Nota: error intencional; rango 11 debería corresponder a 2100, no a 780. El alumno debe verificar.) Rangos correctos de RCS: 150→1, 180→2, 250→3, 320→4, 420→5, 560→6, 780→7, 890→8, 1200→9, 1500→10, 2100→11, 3500→12.

Rangos de leche (de menor a mayor): 0.9→1, 1.2→2, 1.5→3, 1.8→4, 2.1→5, 2.3→6, 2.5→7, 2.7→8, 2.9→9, 3.2→10, 3.5→11, 3.8→12.

Cabra	R(RCS)	R(Leche)	d_i	d_i^2
1	3	10	-7	49
2	9	4	5	25
3	2	11	-9	81
4	8	5	3	9

Cabra	R(RCS)	R(Leche)	d_i	d_i^2
5	4	9	-5	25
6	11	2	9	81
7	1	12	-11	121
8	6	7	-1	1
9	10	3	7	49
10	5	8	-3	9
11	7	6	1	1
12	12	1	11	121

$$\sum d_i^2 = 572$$

$$r_s = 1 - \frac{6 \times 572}{12(144 - 1)} = 1 - \frac{3432}{1716} = 1 - 2.0 = -1.0$$

La correlación es perfectamente monótona: a mayor RCS, menor producción, sin excepción. Esto es un caso extremo por diseño didáctico. En datos reales, r_s rara vez es exactamente −1, pero la dirección negativa es consistente con el efecto conocido de la mastitis sobre la producción láctea.

Nivel avanzado

Ejercicio 4.6.

En un estudio retrospectivo, se revisaron los expedientes de 200 gatos atendidos en una clínica de Montevideo. Para cada gato se registró: ambiente (interior exclusivo, acceso al exterior), estado de FIV (positivo, negativo), sexo (macho, hembra), edad (años) y recuento leucocitario ($\times 10^3/\mu L$). Realice lo siguiente:

(a) Construya la tabla de contingencia ambiente × FIV. Evalúe la asociación con chi-cuadrado y Fisher. Calcule OR con IC 95 % e interprete.

(b) Evalúe si la proporción de FIV+ difiere entre machos y hembras, controlando por el ambiente (análisis estratificado por ambiente: dos tablas 2 × 2, una para gatos de interior y otra para gatos con acceso al exterior).

(c) Calcule la correlación de Spearman entre edad y recuento leucocitario en el subgrupo de gatos FIV+. Grafique la relación.

(d) Calcule la correlación parcial de Spearman entre edad y recuento leucocitario, controlando por la carga viral (si se tuviera esa variable).

(e) Redacte la sección de resultados completa (150–250 palabras) como aparecería en un artículo para el *Journal of Feline Medicine and Surgery*, siguiendo la norma APA 7.ª edición.

Rúbrica orientadora:

(a) Tabla correctamente construida, chi-cuadrado e interpretación: 2 puntos. OR con IC: 1 punto. Nota sobre si OR ≈ RR o no: 1 punto.
(b) Análisis estratificado con dos tablas, interpretación de confusión: 2 puntos.
(c) Cálculo correcto, gráfico con ggplot2, interpretación: 2 puntos.
(d) Fórmula correcta, interpretación de la diferencia entre correlación bruta y parcial: 1 punto.
(e) Redacción científica completa con estadísticos, valores p, IC, tamaños de efecto, todo en formato APA: 3 puntos.

Ejercicio 4.7.

(Integrador) Se proporcionan los siguientes datos de 25 perros con insuficiencia renal crónica atendidos en una clínica de Guadalajara, México:

id, creatinina_mg_dL, BUN_mg_dL, fosforo_mg_dL, estadio_IRIS, apetito
1, 1.8, 35, 5.2, 2, Normal
2, 4.5, 89, 8.1, 3, Reducido
3, 2.1, 42, 5.8, 2, Normal
4, 7.2, 120, 11.3, 4, Anorexia
5, 3.3, 65, 6.9, 3, Reducido
... (completar hasta 25 perros, proporcionados en archivo .csv)

(a) Verifique si la creatinina y el BUN tienen distribución normal. Justifique si procede Pearson o Spearman.

(b) Calcule la matriz de correlaciones de Spearman para creatinina, BUN y fósforo. Visualícela con *corrplot*.

(c) Evalúe la asociación entre estadio IRIS (2, 3, 4) y apetito (Normal, Reducido, Anorexia) con chi-cuadrado y Fisher. Calcule V de Cramér.

(d) Discuta: ¿es apropiado calcular la correlación de Spearman entre el estadio IRIS (ordinal) y la creatinina (continua)? ¿Qué ventajas tiene Kendall en este caso?

(e) Escriba el código completo en R, desde la carga de datos hasta las conclusiones.

Rúbrica orientadora:

a) Prueba de normalidad con justificación: 1 punto.
b) Matriz correcta y gráfico: 2 puntos.
c) Análisis completo con interpretación veterinaria: 2 puntos.
d) Discusión conceptual con justificación: 2 puntos.
e) Código ejecutable y bien comentado: 3 puntos.

Ejercicio 4.8.

(Diseño y análisis) Usted es consultor estadístico y un tesista de maestría le presenta el siguiente diseño: quiere evaluar si un programa de enriquecimiento ambiental modifica el comportamiento estereotípico en 12 tigres de un zoológico. Evaluará a cada tigre antes, durante y un mes después del programa, clasificando la presencia de estereotipias como Sí o No en cada fase. El tesista propone usar tres pruebas chi-cuadrado independientes (antes vs. durante, durante vs. después, antes vs. después).

(a) Explique por qué esa propuesta es incorrecta.

(b) Indique qué prueba estadística es la adecuada para este diseño.

(c) Escriba el código en R necesario para analizar estos datos (puede inventar datos coherentes con el escenario).

(d) ¿Qué debería hacer el tesista si la prueba global es significativa y quiere saber entre qué fases hay diferencias?

Respuestas:

(a) Los tres chi-cuadrado independientes son incorrectos por dos razones: las mediciones provienen de los mismos animales (viola independencia, se requiere McNemar o similar para datos pareados), y hacer tres comparaciones sin corrección infla el error tipo I.

(b) La prueba adecuada es la Q de Cochran: tres mediciones dicotómicas (Sí/No) sobre los mismos 12 sujetos.

(c) El código se proporciona como parte del ejercicio autodirigido. El alumno debe crear un data frame en formato largo con columnas: tigre (factor, 1–12), fase (factor, "antes"/"durante"/"despues") y estereotipia (0/1). Luego aplicar *cochran.qtest()* de *RVAideMemoire*.

(d) Comparaciones post hoc con pruebas de McNemar por pares, ajustando los valores *p* con Bonferroni (α ajustado = 0.05/3 = 0.0167) o con Holm.

4.15. Conclusiones del capítulo

Este capítulo presentó un abanico amplio de herramientas, pero todas responden a una lógica común: analizar la relación entre variables cuando los métodos paramétricos no son apropiados, ya sea porque las variables son categóricas, porque las distribuciones son asimétricas, o porque las escalas de medición no justifican los supuestos de normalidad e igualdad de intervalos.

Las pruebas de chi-cuadrado, Fisher, McNemar y Cochran Q cubren distintas variantes del mismo problema: ¿están asociadas dos (o más) variables categóricas? La elección entre ellas depende de si las observaciones son independientes o pareadas, del tamaño de la tabla, y del tamaño de la muestra. Las medidas de asociación (OR, RR, phi, V de Cramér) complementan las pruebas de hipótesis cuantificando la fuerza de la relación, algo que el valor *p* por sí solo no informa.

Las correlaciones de Spearman y Kendall extienden este mismo principio a variables ordinales o continuas que no cumplen los supuestos de Pearson. Ambas son válidas, pero difieren en su manejo

de empates, su interpretación probabilística y su sensibilidad a distintos tipos de perturbaciones en los datos. La correlación parcial no paramétrica permite controlar por variables confusoras sin recurrir a modelos paramétricos.

Tres ideas para llevarse de este capítulo. Primera: el chi-cuadrado no es una prueba universal para datos categóricos; tiene supuestos claros (independencia, frecuencias esperadas suficientes) y alternativas específicas cuando esos supuestos fallan. Segunda: el valor *p* no mide la fuerza de la asociación; siempre se debe acompañar de una medida de tamaño del efecto. Tercera: Spearman y Kendall miden relaciones monótonas, no lineales; interpretar un r_s alto como evidencia de linealidad es un error conceptual frecuente.

El capítulo siguiente aborda el análisis de sobrevivencia, donde el tiempo hasta un evento (muerte, recaída, curación) es la variable de interés y los métodos de Kaplan-Meier y la prueba de log-rank ofrecen alternativas no paramétricas a los modelos de regresión paramétricos.

Capítulo 5

Análisis de Sobrevivencia y Métodos de Remuestreo

Introducción

Este capítulo trabaja dos familias de métodos que comparten un principio: extraer inferencia válida sin imponer una distribución teórica a los datos. La primera parte desarrolla el análisis de sobrevivencia no paramétrico, diseñado para datos donde la variable de interés es el tiempo hasta que ocurre un evento (muerte, recaída, curación, primer celo) y donde algunas observaciones están censuradas porque el evento no se observó durante el periodo de seguimiento. Se construye el estimador de Kaplan-Meier paso a paso, con el cálculo de la función de sobrevivencia, la mediana de sobrevivencia y los intervalos de confianza. Se desarrolla la prueba de log-rank para comparar curvas de sobrevivencia entre grupos, junto con sus variantes ponderadas (Breslow, Tarone-Ware, Peto-Peto) que difieren en la importancia que asignan a los eventos tempranos frente a los tardíos. Se presentan los supuestos del análisis de sobrevivencia no paramétrico censura no informativa, independencia de las observaciones y se menciona el modelo de Cox como extensión semiparamétrica. La segunda parte aborda los métodos de remuestreo: las pruebas de permutación, que construyen la distribución nula del estadístico de prueba reordenando las etiquetas de grupo de las observaciones reales, tanto en su versión exacta como por aproximación de Monte Carlo; y el bootstrap no paramétrico, que genera la distribución muestral de cualquier estadístico mediante remuestreo con reemplazo de los datos observados, con los métodos percentil y BCa para la construcción de intervalos de confianza. Se trabajan las diferencias entre permutación y bootstrap en cuanto a la pregunta que responden, sus supuestos y sus campos de aplicación.

Objetivos de Aprendizaje

Al finalizar este capítulo, el lector será capaz de:

1. **Identificar** situaciones de investigación veterinaria donde el análisis de sobrevivencia es la herramienta correcta, reconociendo la presencia de censura y distinguiéndola de datos faltantes.
2. **Construir** curvas de Kaplan-Meier a partir de datos clínicos con observaciones censuradas, calcular la mediana de sobrevivencia y sus intervalos de confianza.
3. **Aplicar** la prueba de log-rank y sus variantes (Breslow, Tarone-Ware, Peto-Peto) para comparar curvas de sobrevivencia entre grupos, seleccionando la variante adecuada según la estructura temporal de las diferencias.

4. **Implementar** pruebas de permutación y bootstrap no paramétrico en R para obtener valores p e intervalos de confianza sin depender de supuestos distribucionales.
5. **Interpretar** y reportar los resultados de estos análisis en el contexto de publicaciones veterinarias, identificando los errores más frecuentes y las limitaciones de cada método.

5.1. Introducción: Cuando el Tiempo Hasta el Evento es la Variable de Interés

En un hato ganadero de la provincia de Bolívar, Ecuador, un brote de anaplasmosis bovina golpea a 48 vacas mestizas de doble propósito durante la temporada de lluvias. El veterinario tratante inicia dos protocolos terapéuticos: imidocarb dipropionato en un grupo y oxitetraciclina de larga acción en otro. Lo que le interesa no es simplemente si los animales sobreviven o mueren eso sería una tabla de contingencia con prueba de Fisher, sino **cuánto tiempo sobreviven** desde el diagnóstico. Y aquí aparece un problema que la estadística clásica no resuelve bien.

De las 48 vacas, 7 se venden antes de que termine el estudio. Otras 4 llegan al final del periodo de observación (90 días) sin haber muerto. Otra es sacrificada por una fractura no relacionada con la anaplasmosis. El veterinario tiene datos sobre 12 animales cuyo tiempo de sobrevivencia real desconoce: sabe que sobrevivieron *al menos* hasta cierto punto, pero no sabe cuánto habrían vivido si hubiera podido seguirlos. Descartar esos 12 animales del análisis desperdiciaría información valiosa y sesgaría los resultados. Incluirlos como si hubieran muerto en el momento de la pérdida de seguimiento sería peor.

Este es el territorio del análisis de sobrevivencia: una familia de métodos diseñados específicamente para datos donde la variable respuesta es el **tiempo hasta que ocurre un evento**, con la complicación de que para algunos sujetos ese evento no se observa durante el periodo de estudio. En veterinaria, el evento puede ser la muerte, la recidiva de un tumor, la primera ovulación posparto, la seroconversión, el descarte del animal del hato, o la curación clínica. Lo que unifica a todos estos escenarios es que el tiempo importa, y que la información incompleta es la norma, no la excepción.

Los métodos paramétricos convencionales la prueba t, el ANOVA no pueden manejar esta estructura de datos. No saben qué hacer con una observación que dice “este animal sobrevivió al menos 45 días, pero no sabemos cuánto más”. Los métodos de sobrevivencia sí lo saben, y el

estimador de Kaplan-Meier, la prueba de log-rank y sus variantes son las herramientas no paramétricas que lo hacen posible.

En la segunda mitad del capítulo, el enfoque cambia hacia los **métodos de remuestreo**: pruebas de permutación y bootstrap no paramétrico. Estos métodos comparten con el análisis de sobrevivencia una filosofía común: extraer la mayor cantidad de información posible de los datos observados sin imponer supuestos distribucionales rígidos. El bootstrap permite construir intervalos de confianza para cualquier estadístico la mediana, un coeficiente de correlación de Spearman, una diferencia de proporciones usando únicamente los datos disponibles como "población de referencia". Las pruebas de permutación generan distribuciones nulas reordenando los datos, lo que las hace especialmente útiles cuando las muestras son pequeñas y los supuestos de las pruebas clásicas resultan insostenibles. Ambas técnicas aparecen cada vez con más frecuencia en la literatura veterinaria, y el investigador que no las conoce se pierde herramientas que a menudo son más transparentes y menos restrictivas que las pruebas convencionales.

5.2. Conceptos Fundamentales del Análisis de Sobrevivencia

5.2.1. Tiempo al Evento

La variable respuesta en el análisis de sobrevivencia es el **tiempo transcurrido desde un punto de origen definido hasta la ocurrencia de un evento de interés**. Este tiempo se denota habitualmente como T y es una variable aleatoria continua no negativa (Kleinbaum & Klein, 2012).

Tres elementos deben estar claramente definidos antes de iniciar cualquier análisis:

- **Punto de origen** (*time zero*): el momento a partir del cual se mide el tiempo. En un ensayo clínico con perros con linfoma, podría ser la fecha de la primera dosis de quimioterapia. En un estudio de descarte en vacas lecheras, la fecha de incorporación al hato.

- **Evento de interés** (*failure event*): el desenlace que se espera observar. Muerte, recidiva tumoral, primera inseminación exitosa, seroconversión positiva, descarte del animal. Debe ser un evento binario (ocurrió o no ocurrió) y su definición debe ser inequívoca.

- **Escala de tiempo**: días, semanas, meses o años. La elección depende de la biología del problema y debe mantenerse constante para todos los sujetos.

Si un clínico estudia la sobrevivencia de gatos con enfermedad renal crónica clasificados según estadio IRIS, el tiempo al evento podría definirse como los días desde el diagnóstico (estadio IRIS confirmado) hasta la muerte o eutanasia por causas relacionadas con la enfermedad renal.

5.2.2. Censura

La censura es la característica que distingue al análisis de sobrevivencia de cualquier otro análisis estadístico. Una observación se dice **censurada** cuando el tiempo exacto al evento no se conoce.

Censura por la derecha (*right censoring*) es, con mucho, la más común en investigación veterinaria. Ocurre cuando:

- El sujeto sigue vivo (o sin presentar el evento) al final del periodo de estudio (*censura administrativa*).
- El sujeto se pierde durante el seguimiento: se vende, se traslada, el propietario no regresa a las revisiones.
- El sujeto muere o es descartado por una causa diferente al evento de interés (*riesgo competitivo*).

En los tres casos, se sabe que el animal sobrevivió **al menos** t unidades de tiempo, pero el tiempo real de sobrevivencia es mayor o igual a t. Se denota esta observación como t^+ o con un indicador de censura $\delta = 0$ (evento no observado), mientras que $\delta = 1$ indica que el evento sí ocurrió.

Censura por la izquierda es menos frecuente y ocurre cuando se sabe que el evento ya había ocurrido antes de la primera observación. Por ejemplo, un estudio de seroconversión a leptospirosis donde algunos animales ya son seropositivos al primer muestreo: se sabe que la seroconversión ocurrió en algún momento anterior, pero no se sabe cuándo.

Censura por intervalo aparece cuando el evento ocurre entre dos puntos de observación. En un estudio donde se muestrea sangre cada 30 días para detectar anticuerpos contra Brucella, un animal negativo en el día 60 y positivo en el día 90 tiene un tiempo de seroconversión censurado por intervalo en (60, 90].

Tabla 23

Estructura de datos de sobrevivencia en un estudio de anaplasmosis bovina (n = 12, datos parciales).

Animal	Grupo	Tiempo (días)	Evento (δ)	Observación
V-01	Imidocarb	14	1	Muerte
V-02	Imidocarb	28	1	Muerte
V-03	Imidocarb	45	0	Vendida (censurada)
V-04	Imidocarb	62	1	Muerte
V-05	Imidocarb	90	0	Fin del estudio (censurada)
V-06	Imidocarb	73	1	Muerte
V-07	Oxitetraciclina	7	1	Muerte
V-08	Oxitetraciclina	11	1	Muerte
V-09	Oxitetraciclina	21	0	Sacrificada por fractura (censurada)
V-10	Oxitetraciclina	33	1	Muerte
V-11	Oxitetraciclina	48	1	Muerte
V-12	Oxitetraciclina	90	0	Fin del estudio (censurada)

5.2.3. Función de Sobrevivencia

La **función de sobrevivencia** *S(t)* es la probabilidad de que un sujeto sobreviva más allá del tiempo *t*:

$$S(t) = P(T > t)$$

Tiene tres propiedades:

- $S(0) = 1$: al inicio del estudio, todos los sujetos están vivos.
- $S(t)$ es una función no creciente: la probabilidad de sobrevivencia disminuye (o se mantiene) con el tiempo.
- $\lim_{t\to\infty} S(t) = 0$ en teoría, aunque en la práctica la curva puede no llegar a cero si el estudio termina antes.

Lo que permite el análisis de sobrevivencia y lo que no permite un simple porcentaje de mortalidad es describir **cómo cambia la probabilidad de sobrevivencia a lo largo del tiempo**, no solo si el animal vivió o murió.

5.2.4. Función de Riesgo (Hazard Function)

La **función de riesgo** *h(t)*, también llamada *tasa de fallo instantánea*, mide la velocidad a la que ocurren los eventos en el instante *t*, condicionada a haber sobrevivido hasta ese momento:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

A diferencia de *S(t)*, que es una probabilidad entre 0 y 1, *h(t)* es una tasa y puede tomar cualquier valor no negativo. Su interpretación es intuitiva: un *h(t)* alto indica que, entre los animales que siguen vivos en el momento *t*, la tasa de ocurrencia del evento es alta.

La relación entre ambas funciones es:

$$S(t) = \exp\left(-\int_0^t h\,(u)\,du\right) = \exp\big(-H(t)\big)$$

donde $H(t) = \int_0^t h\,(u)\,du$ es la **función de riesgo acumulado** (*cumulative hazard*). Esta relación implica que conocer una de las funciones equivale a conocer la otra: son dos formas distintas de describir la misma distribución de tiempos de sobrevivencia.

En la práctica veterinaria, *h(t)* es útil para entender la dinámica del riesgo. En un brote de moquillo canino, la función de riesgo podría ser alta en la primera semana (fase aguda de la enfermedad), disminuir en la segunda y tercera semana en los animales que sobreviven la fase inicial, y eventualmente estabilizarse. Este patrón de riesgo decreciente no se ve en una simple tasa de mortalidad acumulada.

5.2.5. Función de Riesgo Acumulado

La función $H(t)$ tiene una interpretación menos intuitiva que $S(t)$ o $h(t)$, pero es útil para diagnósticos gráficos. El estimador de Nelson-Aalen proporciona una estimación no paramétrica de $H(t)$:

$$\widehat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

donde d_i es el número de eventos en el tiempo t_i y n_i es el número de sujetos en riesgo justo antes de t_i. Este estimador se relaciona con el de Kaplan-Meier mediante $\hat{S}(t) \approx \exp\left(-\widehat{H}(t)\right)$, y la aproximación es buena cuando los tamaños de muestra son razonables (Hosmer et al., 2008).

5.3. Estimador de Kaplan-Meier

5.3.1. Construcción del Estimador

El estimador de Kaplan-Meier (KM), también llamado estimador producto-límite, es la herramienta no paramétrica estándar para estimar la función de sobrevivencia $S(t)$ a partir de datos censurados (Kaplan & Meier, 1958). No asume ninguna distribución particular para los tiempos de sobrevivencia, lo que lo convierte en un método genuinamente no paramétrico.

La lógica del estimador es la siguiente: la probabilidad de sobrevivir más allá de un tiempo *t* se descompone como el producto de las probabilidades condicionales de sobrevivir cada uno de los tiempos en los que ocurren eventos. Si los tiempos ordenados en que ocurren eventos son $t_1 < t_2 < \cdots < t_k$, el estimador es:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

donde: - t_i son los tiempos ordenados en los que ocurre al menos un evento. - d_i es el número de eventos (muertes, por ejemplo) en el tiempo t_i. - n_i es el número de sujetos en riesgo justo antes de t_i, es decir, los sujetos que siguen vivos y no han sido censurados antes de ese momento.

El factor $(1 - d_i/n_i)$ es la probabilidad condicional de sobrevivir el tiempo t_i, dado que el sujeto ha sobrevivido hasta justo antes de t_i. El producto acumulado de estos factores da la probabilidad de sobrevivir hasta *t*.

Las observaciones censuradas no generan un escalón en la curva, pero reducen el denominador n_i para los tiempos subsiguientes. Esto es lo que permite al estimador de KM incorporar la información parcial de los sujetos censurados sin tratarlos como eventos ni descartarlos.

5.3.2. Ejemplo Resuelto: Sobrevivencia en Gatos con Linfoma Alimentario

Un hospital veterinario universitario en Bogotá, Colombia, registró los tiempos de sobrevivencia (en semanas) de 10 gatos domésticos de pelo corto diagnosticados con linfoma alimentario de células grandes y tratados con protocolo COP (ciclofosfamida, vincristina, prednisolona). Los datos son:

Tabla 24
Tiempos de sobrevivencia (semanas) de 10 gatos con linfoma alimentario tratados con protocolo COP

Gato	Tiempo (sem)	Estado (δ)
G-01	3	1
G-02	5	1
G-03	7	0
G-04	8	1
G-05	12	1
G-06	13	1
G-07	18	0
G-08	23	1
G-09	26	0
G-10	30	1

Los gatos G-03 (propietario no regresó), G-07 (eutanasia por insuficiencia renal aguda no relacionada) y G-09 (vivo al cierre del estudio) están censurados.

Cálculo paso a paso:

Se ordenan los tiempos de evento (δ = 1) y se calcula n_i y $\hat{S}(t_i)$:

Tabla 25
Cálculo del estimador de Kaplan-Meier para los datos de linfoma felino.

t_i (sem)	n_i	d_i	$1 - d_i/n_i$	$\widehat{S}(t_i)$
3	10	1	9/10 = 0.900	0.900
5	9	1	8/9 = 0.889	0.900 × 0.889 = 0.800
8	7	1	6/7 = 0.857	0.800 × 0.857 = 0.686
12	6	1	5/6 = 0.833	0.686 × 0.833 = 0.571
13	5	1	4/5 = 0.800	0.571 × 0.800 = 0.457
23	3	1	2/3 = 0.667	0.457 × 0.667 = 0.305
30	1	1	0/1 = 0.000	0.305 × 0.000 = 0.000

Nótese que en $t = 8$, $n_i = 7$ y no 8, porque G-03 fue censurado en $t = 7$ y ya no está en riesgo. Del mismo modo, en $t = 23$, $n_i = 3$ porque G-07 fue censurado en $t = 18$ y G-09 será censurado después, pero aún está en riesgo.

La curva de KM tiene forma de escalera descendente, con escalones en cada tiempo de evento y marcas verticales (o cruces) en los tiempos de censura. La función se mantiene constante entre eventos consecutivos.

5.3.3. Mediana de Sobrevivencia

La **mediana de sobrevivencia** es el tiempo *t* en el que $\hat{S}(t) = 0.50$. Es la medida de resumen más usada para datos de sobrevivencia porque, a diferencia de la media, no requiere que se observe el último evento y es resistente a valores extremos.

Para los datos de los gatos con linfoma: - $\hat{S}(12) = 0.571$ - $\hat{S}(13) = 0.457$

La curva cruza el nivel 0.50 entre la semana 12 y la semana 13. Por convención, la mediana se reporta como el menor tiempo en que $\hat{S}(t) \leq 0.50$, que en este caso es **13 semanas**.

Si la curva no cruza el nivel 0.50 lo que ocurre cuando menos de la mitad de los sujetos experimentan el evento, la mediana no se puede estimar. Esto es frecuente en estudios con alta proporción de censura o cuando el pronóstico es favorable.

5.3.4. Intervalos de Confianza para la Función de Sobrevivencia

El intervalo de confianza al 95% para $\hat{S}(t)$ se construye habitualmente usando la **fórmula de Greenwood** para la varianza del estimador de KM:

$$\widehat{Var}\left[\hat{S}(t)\right] = \left[\hat{S}(t)\right]^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

El intervalo de confianza puntual se calcula como:

$$\hat{S}(t) \pm z_{1-\alpha/2}\sqrt{\widehat{Var}\left[\hat{S}(t)\right]}$$

Sin embargo, este intervalo tiene el defecto de que puede exceder el rango [0, 1]. Para evitarlo, en la práctica se prefiere la transformación log-log complementaria:

$$\exp\left(-\exp\left[\log\left(-\log\hat{S}(t)\right) \pm \frac{z_{1-\alpha/2}}{\log\hat{S}(t)}\sqrt{\widehat{Var}\left[\hat{S}(t)\right]/\hat{S}(t)}\right]\right)$$

R implementa esta transformación por defecto en *survfit()* con el argumento *conf.type = "log-log"*.

Para la mediana de sobrevivencia, el intervalo de confianza se obtiene a partir de los intervalos de confianza puntuales de $\hat{S}(t)$: el límite inferior de la mediana es el menor *t* donde el límite superior del IC de $\hat{S}(t)$ cruza 0.50, y el límite superior es el menor *t* donde el límite inferior del IC cruza 0.50 (Brookmeyer & Crowley, 1982).

5.3.5. Segundo Ejemplo: Sobrevivencia de Perros con Osteosarcoma Apendicular

En la Facultad de Medicina Veterinaria de la Universidad Nacional Autónoma de México, se registraron los tiempos de sobrevivencia (en días) de 15 perros de razas grandes (Rottweiler, Gran Danés, Labrador Retriever) diagnosticados con osteosarcoma apendicular y tratados con

amputación seguida de quimioterapia con carboplatino. Se comparan con 13 perros tratados con amputación sola (sin quimioterapia).

Tabla 26

Tiempos de sobrevivencia (días) de perros con osteosarcoma apendicular

Animal	Grupo	Tiempo (días)	δ
P-01	Amputación + Carbo	45	1
P-02	Amputación + Carbo	78	1
P-03	Amputación + Carbo	112	0
P-04	Amputación + Carbo	134	1
P-05	Amputación + Carbo	180	1
P-06	Amputación + Carbo	210	0
P-07	Amputación + Carbo	245	1
P-08	Amputación + Carbo	268	1
P-09	Amputación + Carbo	290	0
P-10	Amputación + Carbo	312	1
P-11	Amputación + Carbo	345	1
P-12	Amputación + Carbo	380	0
P-13	Amputación + Carbo	400	1
P-14	Amputación + Carbo	450	0
P-15	Amputación + Carbo	480	1
P-16	Amputación sola	20	1
P-17	Amputación sola	35	1
P-18	Amputación sola	52	1
P-19	Amputación sola	68	0
P-20	Amputación sola	90	1
P-21	Amputación sola	105	1
P-22	Amputación sola	125	1
P-23	Amputación sola	140	0
P-24	Amputación sola	160	1
P-25	Amputación sola	185	1
P-26	Amputación sola	210	1
P-27	Amputación sola	240	0
P-28	Amputación sola	270	1

Este conjunto de datos se utilizará en las secciones siguientes para ilustrar la prueba de log-rank y la implementación en R.

5.4. Prueba de Log-Rank y Variantes

5.4.1. Formulación de la Prueba de Log-Rank

La pregunta clínica en el ejemplo del osteosarcoma canino es directa: ¿la quimioterapia adyuvante con carboplatino mejora la sobrevivencia respecto a la amputación sola? En términos estadísticos, se comparan las funciones de sobrevivencia de dos (o más) grupos:

- H_0: $S_1(t) = S_2(t)$ para todo *t* (las curvas de sobrevivencia son idénticas).
- H_1: $S_1(t) \neq S_2(t)$ para al menos algún *t*.

La **prueba de log-rank** (Mantel, 1966; Peto & Peto, 1972) es la prueba no paramétrica estándar para esta comparación. Su lógica se basa en comparar, en cada tiempo de evento, el número de eventos observados en cada grupo con el número de eventos que se esperaría bajo la hipótesis nula.

Para cada tiempo de evento t_i, se construye una tabla 2 × 2:

	Eventos	**No eventos**	**En riesgo**
Grupo 1	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
Grupo 2	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	d_i	$n_i - d_i$	n_i

Bajo H_0, el número esperado de eventos en el grupo 1 en el tiempo t_i es:

$$e_{1i} = \frac{n_{1i} \cdot d_i}{n_i}$$

El estadístico de log-rank es:

$$\chi^2_{LR} = \frac{(\sum_i d_{1i} - \sum_i e_{1i})^2}{\sum_i V_i}$$

donde V_i es la varianza hipergeométrica:

$$V_i = \frac{n_{1i} \cdot n_{2i} \cdot d_i \cdot (n_i - d_i)}{n_i^2 (n_i - 1)}$$

Bajo H_0, χ^2_{LR} sigue asintóticamente una distribución χ^2 con $g-1$ grados de libertad, donde g es el número de grupos. Para dos grupos, 1 grado de libertad.

La prueba de log-rank asigna **el mismo peso a todos los tiempos de evento**, independientemente de cuántos sujetos queden en riesgo. Esto la hace sensible a diferencias que se mantienen a lo largo de todo el periodo de seguimiento, pero menos sensible a diferencias concentradas en periodos tempranos o tardíos.

5.4.2. Variantes Ponderadas de la Prueba de Log-Rank

Para situaciones donde las diferencias entre grupos no son uniformes a lo largo del tiempo, se han desarrollado variantes que asignan pesos diferentes a cada tiempo de evento. La estructura general es:

$$\chi^2_W = \frac{\left(\sum_i w_i \left(d_{1i} - e_{1i}\right)\right)^2}{\sum_i w_i^2 \, V_i}$$

donde w_i es la función de ponderación. Las variantes más utilizadas son:

Prueba de Breslow (Gehan-Breslow-Wilcoxon): $w_i = n_i$ (el número total en riesgo). Da mayor peso a los tiempos tempranos, cuando hay más sujetos en riesgo. Es más potente que la prueba de log-rank estándar cuando las diferencias entre grupos son mayores al inicio del seguimiento y se atenúan con el tiempo. Tiene la desventaja de que los pesos dependen fuertemente del patrón de censura (Breslow, 1970).

Prueba de Tarone-Ware: $w_i = \sqrt{n_i}$. Un compromiso entre la prueba de log-rank (pesos iguales) y la de Breslow ($w_i = n_i$). Útil cuando no se tiene una expectativa clara sobre dónde se concentran las diferencias (Tarone & Ware, 1977).

Prueba de Peto-Peto: $w_i = \tilde{S}(t_i)$, donde $\tilde{S}(t_i)$ es una versión modificada del estimador de Kaplan-Meier. Da mayor peso a los tiempos tempranos (cuando $\tilde{S}$ es cercana a 1) y menor peso a los tiempos tardíos. A diferencia de la prueba de Breslow, los pesos no dependen del patrón de censura, lo que la hace más robusta (Peto & Peto, 1972).

Tabla 27
Selección de la variante de la prueba de log-rank según la estructura temporal de las diferencias.

Patrón de diferencia	Variante recomendada	Razón
Diferencia constante en el tiempo	Log-rank estándar	Máxima potencia bajo riesgos proporcionales
Diferencia mayor al inicio	Breslow o Peto-Peto	Mayor peso a tiempos tempranos
Diferencia mayoral final	Log-rank estándar	No subpondera los tiempos tardíos
Incertidumbre sobre el patrón	Tarone-Ware	Compromiso entre los extremos
Alta censura diferencial entre grupos	Peto-Peto	Pesos no dependen del patrón de censura

5.4.3. Ejemplo Numérico: Cálculo de la Prueba de Log-Rank

Se calculará la prueba de log-rank para los datos simplificados de la Tabla 5.1 (anaplasmosis bovina, datos parciales con n = 12).

Se ordenan todos los tiempos de evento (δ = 1) de ambos grupos combinados: 7, 11, 14, 28, 33, 48, 62, 73 días.

Tabla 28
Cálculo paso a paso de la prueba de log-rank para los datos de anaplasmosis bovina.

t_i	n_{1i} (Imid.)	n_{2i} (Oxi.)	n_i	d_{1i}	d_{2i}	d_i	$e_{1i} = n_{1i}d_i/n_i$	V_i
7	6	6	12	0	1	1	6(1)/12 = 0.500	0.227
11	6	5	11	0	1	1	6(1)/11 = 0.545	0.248
14	6	4	10	1	0	1	6(1)/10 = 0.600	0.240
28	5	3	8	1	0	1	5(1)/8 = 0.625	0.234

t_i	n_{1i} (Imid.)	n_{2i} (Oxi.)	n_i	d_{1i}	d_{2i}	d_i	$e_{1i} = n_{1i}d_i/n_i$	V_i
33	4	2	6	0	1	1	4(1)/6 = 0.667	0.222
48	3	1	4	0	1	1	3(1)/4 = 0.750	0.188
62	2	0	2	1	0	1	2(1)/2 = 1.000	0.000
73	1	0	1	1	0	1	1(1)/1 = 1.000	0.000

Nota. La observación V-03 (censurada en t = 45, imidocarb) reduce n_{1i} a partir del siguiente tiempo de evento. La observación V-09 (censurada en t = 21, oxitetraciclina) reduce n_{2i} a partir de t = 28.

El cálculo del estadístico:

$$O_1 = \sum d_{1i} = 0 + 0 + 1 + 1 + 0 + 0 + 1 + 1 = 4$$

$$E_1 = \sum e_{1i} = 0.500 + 0.545 + 0.600 + 0.625 + 0.667 + 0.750 + 1.000 + 1.000 = 5.687$$

$$\sum V_i = 0.227 + 0.248 + 0.240 + 0.234 + 0.222 + 0.188 + 0.000 + 0.000 = 1.359$$

$$\chi^2_{LR} = \frac{(4 - 5.687)^2}{1.359} = \frac{(-1.687)^2}{1.359} = \frac{2.846}{1.359} = 2.094$$

Con 1 grado de libertad, el valor p es $P(\chi^2_1 \geq 2.094) \approx 0.148$.

Con estos datos parciales (n = 12), la prueba de log-rank no detecta diferencia estadísticamente significativa entre los dos tratamientos (χ^2 = 2.09, gl = 1, p = 0.148). Esto no implica que los tratamientos sean equivalentes el tamaño de muestra es pequeño y la potencia de la prueba es limitada. Obsérvese que el grupo de oxitetraciclina tuvo más eventos tempranos (7 y 11 días), lo que una prueba de Breslow podría captar con mayor sensibilidad.

5.5. Supuestos del Análisis de Sobrevivencia No Paramétrico

5.5.1. Censura No Informativa

El supuesto más crítico del estimador de Kaplan-Meier y de la prueba de log-rank es que la **censura es no informativa**: la razón por la que un sujeto es censurado no está relacionada con su pronóstico. En otras palabras, los sujetos censurados en el tiempo *t* tienen la misma perspectiva de sobrevivencia futura que los sujetos que permanecen bajo observación en ese momento.

Este supuesto se viola cuando:

- Los animales más enfermos son los que se pierden del seguimiento (porque mueren en casa sin que el clínico lo sepa).
- Los propietarios con menos recursos económicos retiran a sus animales del estudio antes, y estos animales también tienen peor pronóstico por falta de cuidados.
- En producción animal, los animales menos productivos se descartan primero, y el descarte se registra como censura.

Cuando la censura es informativa, el estimador de KM sesga la estimación de $S(t)$ típicamente la sobreestima, porque los sujetos de mayor riesgo desaparecen del análisis. No existe una prueba formal para verificar este supuesto; debe evaluarse por razonamiento clínico y conocimiento del mecanismo de pérdida de seguimiento (Leung et al., 1997).

5.5.2. Riesgos Proporcionales

La prueba de log-rank es más potente cuando los **riesgos son proporcionales**: la razón de riesgo entre dos grupos es constante a lo largo del tiempo. Esto equivale a decir que las curvas de sobrevivencia de los dos grupos no se cruzan, o más formalmente, que:

$$h_1(t) = \theta \cdot h_2(t) \quad \text{para todo } t$$

donde θ es una constante.

Si las curvas se cruzan, por ejemplo, si un tratamiento es más riesgoso al inicio, pero más protector a largo plazo, la prueba de log-rank puede tener potencia muy baja porque las diferencias en una dirección se cancelan con las diferencias en la otra. En esta situación, las variantes ponderadas (Breslow, Peto-Peto) tampoco son la solución, porque el problema no es de ponderación sino de

que la hipótesis alternativa tiene una estructura que estas pruebas no capturan. Podrían necesitarse enfoques como la prueba de Renyi o modelos de riesgos no proporcionales.

Un diagnóstico gráfico sencillo es graficar $\log\left(-\log\left(\hat{S}(t)\right)\right)$ contra $\log(t)$ para cada grupo. Si los riesgos son proporcionales, las curvas serán aproximadamente paralelas. Si se cruzan o convergen, el supuesto se viola.

5.5.3. Independencia de las Observaciones

Cada sujeto debe ser independiente de los demás. En veterinaria, esto se viola frecuentemente cuando los animales pertenecen al mismo hato, criadero o camada. Dos cachorros de la misma madre con parvovirus canino no son observaciones independientes: comparten genética, anticuerpos maternos y ambiente. Los modelos de fragilidad (*frailty models*) son una extensión del modelo de Cox que incorpora esta correlación intra-grupo, pero van más allá del alcance de este capítulo.

5.6. Mención del Modelo de Cox como Extensión Semiparamétrica

El modelo de riesgos proporcionales de Cox (Cox, 1972) merece una mención breve porque es la extensión natural del análisis no paramétrico de sobrevivencia. Mientras que Kaplan-Meier estima la función de sobrevivencia y log-rank compara curvas entre grupos, el modelo de Cox permite evaluar el efecto simultáneo de múltiples covariables (peso corporal, edad, raza, estadio clínico) sobre la tasa de riesgo:

$$h(t \mid \mathbf{x}) = h_0(t)\exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)$$

donde $h_0(t)$ es la función de riesgo basal (no paramétrica, no se especifica su forma) y los coeficientes β cuantifican el efecto de cada covariable. El modelo es **semiparamétrico**: la parte temporal es no paramétrica y la parte de las covariables es paramétrica.

El resultado principal del modelo de Cox es el **hazard ratio** (HR), que es $\exp(\beta)$. Un HR = 2.3 para un tratamiento significa que los sujetos con ese tratamiento tienen 2.3 veces la tasa de riesgo instantáneo que el grupo de referencia, ajustado por las demás covariables.

El modelo de Cox asume riesgos proporcionales, y esta vez el supuesto es verificable formalmente mediante los residuos de Schoenfeld y la prueba de Grambsch y Therneau (1994), implementada en R con *cox.zph()*.

Este capítulo no desarrolla el modelo de Cox en detalle porque, siendo semiparamétrico, pertenece a un marco teórico diferente del puramente no paramétrico. Sin embargo, el investigador veterinario que domina Kaplan-Meier y log-rank tiene ya la base conceptual para abordar el modelo de Cox, y se recomienda su estudio en textos especializados como Hosmer et al. (2008) o Therneau y Grambsch (2000).

5.7. Pruebas de Permutación

5.7.1. Lógica Conceptual

Las pruebas de permutación representan uno de los enfoques más elegantes y transparentes de la inferencia estadística. La idea, propuesta originalmente por Fisher (1935), es engañosamente simple: si la hipótesis nula es cierta y no hay diferencia entre dos grupos, entonces la etiqueta de grupo asignada a cada observación es irrelevante. Se podría reasignar aleatoriamente cada observación a cualquiera de los dos grupos y el estadístico de prueba tendría un valor similar al observado.

Considérese un ejemplo concreto. Un veterinario en Arequipa, Perú, mide la ganancia de peso (kg) en 5 alpacas Huacaya suplementadas con un bloque mineral y 5 alpacas control durante 60 días:

- Suplementadas: 3.2, 4.1, 3.8, 5.0, 4.5
- Control: 2.1, 2.8, 3.0, 2.5, 3.3

La diferencia de medias observada es $\bar{x}_S - \bar{x}_C = 4.12 - 2.74 = 1.38$ kg.

La pregunta de la prueba de permutación es: ¿Qué tan probable es obtener una diferencia de medias de 1?38 kg o mayor si las etiquetas “suplementada” y “control” fueran intercambiables?

Para responderla, se generan **todas** las formas posibles de dividir las 10 observaciones en dos grupos de 5. El número total de permutaciones es $\binom{10}{5} = 252$. Para cada permutación, se calcula la diferencia de medias. La distribución de estas 252 diferencias es la **distribución nula exacta**

del estadístico bajo H_0. El valor p es la proporción de permutaciones que produce una diferencia igual o mayor (en valor absoluto, para una prueba bilateral) a la observada.

5.7.2. Permutaciones Exactas

Cuando el número total de permutaciones es computacionalmente manejable (digamos, menos de un millón), se pueden enumerar todas y obtener un **valor p exacto**. No se necesitan tablas, no se asume ninguna distribución, no se requieren muestras grandes. La única condición es que las observaciones sean intercambiables bajo H_0.

El número de permutaciones para una prueba de dos grupos independientes con tamaños n_1 y n_2 es:

$$\binom{n_1 + n_2}{n_1} = \frac{(n_1 + n_2)!}{n_1! \cdot n_2!}$$

Para $n_1 = n_2 = 5$, son 252 permutaciones. Para $n_1 = n_2 = 10$, son 184,756. Para $n_1 = n_2 = 15$, ya son 155,117,520 computacionalmente costoso, pero todavía factible con un procesador moderno.

Ejemplo resuelto con las alpacas:

Se calcula la diferencia de medias para cada una de las 252 permutaciones. El resultado (que se obtendrá con R más adelante) muestra que solo 4 de las 252 permutaciones producen una diferencia de medias ≥ 1.38 en valor absoluto.

$$p_{exacto} = \frac{4}{252} = 0.0159$$

Con un nivel de significancia de 0.05, se rechaza H_0: hay evidencia de que la suplementación mineral incrementa la ganancia de peso en alpacas.

5.7.3. Pruebas de Permutación por Monte Carlo

Cuando el número de permutaciones es demasiado grande para enumerarlas todas, se recurre a una **aproximación de Monte Carlo**: se muestrean aleatoriamente *B* permutaciones (típicamente B = 10,000 o más) y se calcula el valor p como:

$$p_{MC} = \frac{\#\{|T_b^*| \geq |T_{obs}|\} + 1}{B + 1}$$

donde T_{obs} es el estadístico observado, T_b^* es el estadístico de la *b*-ésima permutación, y el "+1" en numerador y denominador asegura que el valor p nunca sea exactamente cero (Phipson & Smyth, 2010). Esta corrección es importante porque un valor p reportado como "p = 0.000" es un sinsentido estadístico no se puede asegurar que la probabilidad sea exactamente cero con un número finito de permutaciones.

La precisión de p_{MC} depende de *B*. Con B = 10,000, el error estándar de un valor p de 0.05 es $\sqrt{0.05 \times 0.95/10000} = 0.0022$, lo que da un IC al 95% para el valor p de [0.046, 0.054]. Para la mayoría de decisiones prácticas, esta precisión es suficiente.

5.7.4. Ventajas y Limitaciones

Las pruebas de permutación tienen ventajas claras:

- No asumen ninguna distribución de los datos.
- Funcionan con muestras pequeñas (donde las pruebas asintóticas son poco confiables).
- Son exactas (en su versión completa) o arbitrariamente precisas (en su versión Monte Carlo).
- Son aplicables a cualquier estadístico de prueba: diferencia de medias, diferencia de medianas, razón de varianzas, correlaciones, estadísticos multivariados.

Sus limitaciones son:

- Asumen **intercambiabilidad** bajo H_0: las observaciones de los diferentes grupos deben provenir de la misma distribución bajo la nula. Esto se viola si las distribuciones difieren en forma, aunque tengan la misma localización.

- Son computacionalmente intensivas para muestras grandes con permutaciones exactas, aunque Monte Carlo resuelve este problema.
- No proporcionan directamente un tamaño del efecto o un intervalo de confianza (para eso se necesita el bootstrap, que se trata a continuación).

5.8. Bootstrap No Paramétrico

5.8.1. La Idea del Bootstrap

El bootstrap, introducido por Efron (1979), es quizá la contribución más influyente a la estadística computacional del siglo XX. Su lógica es la siguiente: cuando se quiere estimar la variabilidad de un estadístico (la media, la mediana, un coeficiente de correlación, la diferencia entre dos proporciones) pero no se conoce la distribución de la población, se usa la distribución empírica de la muestra como sustituto de la distribución poblacional.

El procedimiento es:

1. De la muestra original de tamaño *n*, se extraen *B* muestras **con reemplazo**, cada una de tamaño *n*. Estas son las **muestras bootstrap**.
2. Para cada muestra bootstrap, se calcula el estadístico de interés $\hat{\theta}_b^*$.
3. La distribución de los B valores $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$ es la **distribución bootstrap**, que aproxima la distribución muestral del estadístico.

El punto conceptual clave es que el remuestreo es **con reemplazo**. Cada muestra bootstrap puede contener la misma observación varias veces y no contener otras. Esto genera variabilidad entre las muestras bootstrap, y esa variabilidad refleja la variabilidad del estadístico.

Para un investigador veterinario, la utilidad inmediata del bootstrap es la construcción de intervalos de confianza para estadísticos que no tienen fórmula analítica para su error estándar o cuya fórmula analítica asume normalidad, que es precisamente el supuesto que no se cumple.

5.8.2. Intervalo de Confianza Bootstrap Percentil

El método más intuitivo. Se ordenan los B valores del estadístico bootstrap y se toman los percentiles $\alpha/2$ y $1 - \alpha/2$:

$$IC_{percentil} = \left[\hat{\theta}^*_{(\alpha/2)}, \quad \hat{\theta}^*_{(1-\alpha/2)}\right]$$

Para un IC al 95% con B = 10,000 réplicas, se toman los valores en las posiciones 250 y 9750 de los estadísticos ordenados.

Ejemplo: Se desea construir un IC al 95% para la mediana de la concentración de inmunoglobulina G (IgG, en mg/dL) en calostro de vacas primíparas Holstein en la sabana de Bogotá. Los datos de 12 vacas son:

4,520; 3,890; 5,110; 2,780; 6,340; 4,150; 3,200; 5,670; 4,880; 3,560; 7,210; 4,430

La mediana observada es $\tilde{x} = (4{,}430 + 4{,}520)/2 = 4{,}475$ mg/dL.

Para construir el IC bootstrap percentil: 1. Se generan B = 10,000 muestras bootstrap de tamaño 12 con reemplazo. 2. Se calcula la mediana de cada muestra. 3. Se ordenan las 10,000 medianas y se toman los percentiles 2.5% y 97.5%.

El resultado (que se obtendrá con R) es aproximadamente IC 95%: [3,725; 5,390] mg/dL.

Este intervalo no asume normalidad de las concentraciones de IgG que, con 12 observaciones y una distribución visiblemente asimétrica, sería un supuesto difícil de defender.

5.8.3. Intervalo de Confianza BCa (Bias-Corrected and Accelerated)

El intervalo percentil es simple, pero tiene dos limitaciones: no corrige el sesgo del estimador bootstrap ni la asimetría de su distribución. El método **BCa** (Efron, 1987) incorpora dos correcciones:

- **Corrección de sesgo** ($\hat{z}_0$): mide cuánto se desvía la distribución bootstrap del estadístico original. Si la mediana de los valores bootstrap difiere sistemáticamente del valor observado, $\hat{z}_0$ lo corrige.
- **Aceleración** ($\hat{a}$): mide la tasa de cambio del error estándar del estimador en función del parámetro. Corrige la asimetría de la distribución bootstrap. Se estima mediante jackknife.

Los percentiles ajustados son:

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}\left(\hat{z}_0 + z_{\alpha/2}\right)}\right)$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}\left(\hat{z}_0 + z_{1-\alpha/2}\right)}\right)$$

donde Φ es la función de distribución acumulada normal estándar y $z_{\alpha/2}$ es el cuantil correspondiente.

En la práctica, el investigador no necesita calcular esto a mano la función *boot.ci()* del paquete *boot* en R lo hace automáticamente, pero sí necesita saber que el intervalo BCa es generalmente preferible al percentil simple porque tiene mejores propiedades de cobertura, especialmente con muestras pequeñas y distribuciones asimétricas (DiCiccio & Efron, 1996).

5.8.4. Cuántas Réplicas Bootstrap se Necesitan

La recomendación general (Efron & Tibshirani, 1993):

- Para estimar el error estándar: B = 200 suele ser suficiente.
- Para intervalos de confianza percentil: B = 1,000 como mínimo, B = 10,000 preferible.
- Para intervalos BCa: B = 2,000 como mínimo, B = 10,000 o más recomendable, porque la estimación de los percentiles ajustados requiere mayor precisión en las colas de la distribución.

Con el poder computacional actual, no hay razón para usar menos de B = 10,000 en un análisis publicable.

5.9. Supuestos y Limitaciones del Remuestreo

5.9.1. ¿Qué Asumen Realmente Estos Métodos?

Es un mito que los métodos de remuestreo “no tienen supuestos”. Tienen menos supuestos distribucionales que los métodos paramétricos, pero no son libres de condiciones:

Pruebas de permutación: - Intercambiabilidad bajo H_0: las observaciones deben provenir de la misma distribución cuando la hipótesis nula es cierta. Si los dos grupos tienen la misma media, pero varianzas muy diferentes, la permutación de etiquetas no preserva la estructura de los datos, y la prueba puede tener tasas de error tipo I infladas. - Independencia de las observaciones. Los datos de medidas repetidas o con estructura jerárquica (animales dentro de granjas) no se pueden permutar libremente.

Bootstrap: - La muestra debe ser **representativa** de la población. Si la muestra original está sesgada (por ejemplo, solo incluye casos clínicos graves porque los leves no llegan al hospital), el bootstrap reproducirá y amplificará ese sesgo. - **Independencia** entre observaciones. El bootstrap estándar no es válido para datos correlacionados (series temporales, datos agrupados). Existen variantes como el bootstrap por bloques (*block bootstrap*) y el bootstrap de clusters, pero van más allá de este capítulo. - El bootstrap funciona mal cuando el estadístico de interés depende de los **valores extremos** de la distribución. El máximo, el mínimo y los percentiles extremos se estiman pobremente con bootstrap porque la muestra original puede no contener valores representativos de las colas. - Con muestras muy pequeñas (n < 10), la distribución empírica es una aproximación burda de la distribución poblacional, y los intervalos bootstrap pueden tener cobertura real muy inferior a la nominal.

5.9.2. Bootstrap vs. Permutación: ¿Cuándo Usar Cada Uno?

La distinción práctica es:

- **Prueba de permutación** → para obtener un **valor p** (prueba de hipótesis). La pregunta es: ¿hay diferencia entre grupos?
- **Bootstrap** → para obtener un **intervalo de confianza** (estimación). La pregunta es: ¿cuál es la magnitud del efecto y con qué incertidumbre?

No son excluyentes. En un análisis completo se puede usar una prueba de permutación para evaluar la significancia y bootstrap para estimar el intervalo de confianza del tamaño del efecto.

5.10. Implementación en R

5.10.1. Paquetes Requeridos

```r
# ============================================================
# Paquetes necesarios para el Capítulo 5
# ============================================================

# Instalar paquetes si no están disponibles
paquetes <- c("survival", "survminer", "coin", "boot",
              "ggplot2", "dplyr", "tidyr")

paquetes_faltantes <- paquetes[!(paquetes %in% installed.packages()[, "Package"])]
if (length(paquetes_faltantes) > 0) {
  install.packages(paquetes_faltantes)
}

library(survival)    # Kaplan-Meier, log-rank, modelo de Cox
library(survminer)   # Gráficos de sobrevivencia con ggplot2
library(coin)        # Pruebas de permutación exactas y aproximadas
library(boot)        # Bootstrap no paramétrico
library(ggplot2)     # Gráficos
library(dplyr)       # Manipulación de datos
```

5.10.2. Análisis de Sobrevivencia: Gatos con Linfoma Alimentario

```r
# ============================================================
# Ejemplo 1: Sobrevivencia de gatos con linfoma alimentario
# ============================================================

# Crear los datos de la Tabla 5.2
linfoma_gatos <- data.frame(
  gato = paste0("G-", sprintf("%02d", 1:10)),
  tiempo_sem = c(3, 5, 7, 8, 12, 13, 18, 23, 26, 30),
  estado = c(1, 1, 0, 1, 1, 1, 0, 1, 0, 1)
  # estado: 1 = muerte (evento), 0 = censurado
)

# Crear objeto de sobrevivencia
# Surv() combina el tiempo y el indicador de evento en un solo objeto
# El símbolo "+" junto a un tiempo indica censura
surv_gatos <- Surv(time = linfoma_gatos$tiempo_sem,
                   event = linfoma_gatos$estado)

print(surv_gatos)
# Salida esperada:
#  [1]  3   5   7+  8  12  13  18+ 23  26+ 30
```

```r
# Ajustar el estimador de Kaplan-Meier
# survfit() calcula el estimador KM con intervalos de confianza
km_gatos <- survfit(surv_gatos ~ 1,
                    conf.type = "log-log")
# conf.type = "log-log" usa la transformación log(-log(S(t)))
# que garantiza que el IC quede dentro de [0, 1]

# Resumen del estimador KM
summary(km_gatos)
# Salida esperada (valores aproximados):
# time n.risk n.event survival std.err lower 95% CI upper 95% CI
#    3     10       1    0.900  0.0949       0.6531        0.977
#    5      9       1    0.800  0.1265       0.5094        0.929
#    8      7       1    0.686  0.1512       0.3726        0.870
#   12      6       1    0.571  0.1659       0.2620        0.793
#   13      5       1    0.457  0.1688       0.1712        0.700
#   23      3       1    0.305  0.1691       0.0684        0.611
#   30      1       1    0.000     NaN           NA

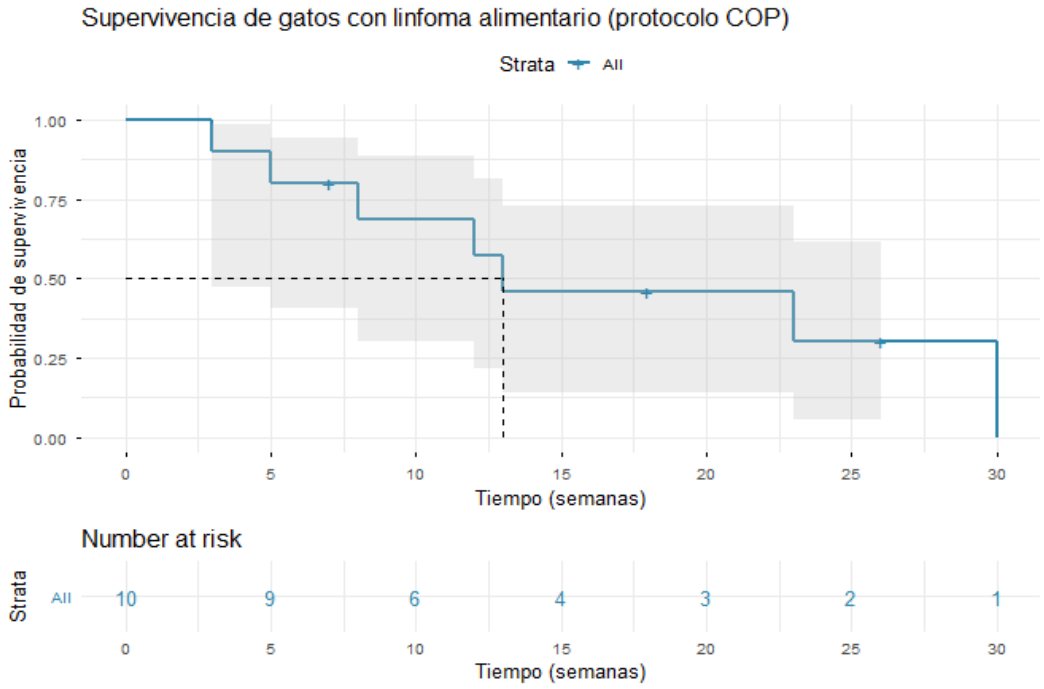
# Mediana de sobrevivencia e IC
print(km_gatos)
# Salida esperada:
# n  events  median  0.95LCL  0.95UCL
# 10      7      13        5       30

# La mediana de sobrevivencia es 13 semanas (IC 95%: 5 - 30)
# El IC es amplio, reflejo del tamaño de muestra pequeño (n = 10)

# ---- Gráfico de la curva de Kaplan-Meier ----
# survminer produce gráficos de sobrevivencia listos para publicación
ggsurvplot(km_gatos,
           data = linfoma_gatos,
           conf.int = TRUE,          # Mostrar IC al 95%
           risk.table = TRUE,        # Tabla de sujetos en riesgo
           risk.table.col = "strata",
           xlab = "Tiempo (semanas)",
           ylab = "Probabilidad de sobrevivencia",
           title = "Sobrevivencia de gatos con linfoma alimentario (protocolo COP)",
           ggtheme = theme_minimal(),
           palette = "#2E86AB",      # Color de la curva
           censor.shape = "+",       # Marcador para censuras
           censor.size = 5,          # Tamaño del marcador
           surv.median.line = "hv",  # Líneas para la mediana
           break.x.by = 5)           # Intervalos en eje X
```

Figura 15

Gráfico de la curva de Kaplan-Meier de Supervivencia de gatos con linfoma alimentaria



Nota. Descripción de la Figura 5.1: La curva desciende en escalones en cada tiempo de evento. L as marcas "+" indican las observaciones censuradas (sem 7, 18, 26). Las líneas punteadas horizontal y vertical marcan la mediana (13 sem). La banda sombreada gris representa el IC al 95%. Debajo del gráfi co, la tabla de riesgo muestra cuántos sujetos permanecen bajo observación en cada intervalo.

5.10.3. Análisis de Sobrevivencia: Osteosarcoma Canino con Prueba de Log-Rank

```
# ================================================================
# Ejemplo 2: Osteosarcoma canino - comparación de dos tratamientos
# ================================================================

# Crear los datos de la Tabla 5.4
osteo_perros <- data.frame(
  animal = paste0("P-", sprintf("%02d", 1:28)),
  grupo = c(rep("Amputación + Carboplatino", 15),
            rep("Amputación sola", 13)),
  tiempo_dias = c(45, 78, 112, 134, 180, 210, 245, 268, 290,
                  312, 345, 380, 400, 450, 480,
                  20, 35, 52, 68, 90, 105, 125, 140, 160,
                  185, 210, 240, 270),
```

```
  estado = c(1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1,
             1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1)
)

# Convertir grupo a factor con niveles ordenados
osteo_perros$grupo <- factor(osteo_perros$grupo,
                             levels = c("Amputación sola",
                                        "Amputación + Carboplatino"))

# Crear objeto de sobrevivencia
surv_osteo <- Surv(time = osteo_perros$tiempo_dias,
                   event = osteo_perros$estado)

# Estimador de Kaplan-Meier por grupo
km_osteo <- survfit(surv_osteo ~ grupo, data = osteo_perros,
                    conf.type = "log-log")

# Resumen por grupo
summary(km_osteo)

# Medianas por grupo
print(km_osteo)
# Salida esperada (valores aproximados):
#                              n  events  median  0.95LCL  0.95UCL
# grupo=Amputación sola        13      10     125       52      210
# grupo=Amputación + Carboplatino 15     10     290      134      450

# ---- Prueba de log-rank ----
# survdiff() calcula la prueba de log-rank por defecto
# El argumento rho controla la ponderación:
#   rho = 0: log-rank estándar (Mantel)
#   rho = 1: Peto-Peto
logrank_test <- survdiff(surv_osteo ~ grupo, data = osteo_perros,
                         rho = 0)
print(logrank_test)
# Salida esperada (valores aproximados):
# Call: survdiff(formula = surv_osteo ~ grupo, data = osteo_perros, rho = 0)
#
#                               N  Observed  Expected  (O-E)^2/E  (O-E)^2/V
# grupo=Amputación sola            13       10      7.12       1.17       2.76
# grupo=Amputación + Carboplatino   15       10     12.88       0.64       2.76
#
# Chisq= 2.8  on 1 degrees of freedom, p= 0.0953
#
# Interpretación: Con la prueba de log-rank estándar no se alcanza
# significancia al nivel 0.05 (p = 0.095), aunque el valor p está
# próximo al umbral. El grupo con carboplatino muestra una tendencia
```

```r
# a mayor sobrevivencia (mediana 290 vs 125 días).

# ---- Variantes ponderadas ----
# Prueba de Peto-Peto (rho = 1)
petopeto_test <- survdiff(surv_osteo ~ grupo, data = osteo_perros,
                          rho = 1)
print(petopeto_test)
# Da mayor peso a los tiempos tempranos.
# Puede dar un valor p más bajo si las diferencias son
# mayores al inicio del seguimiento.

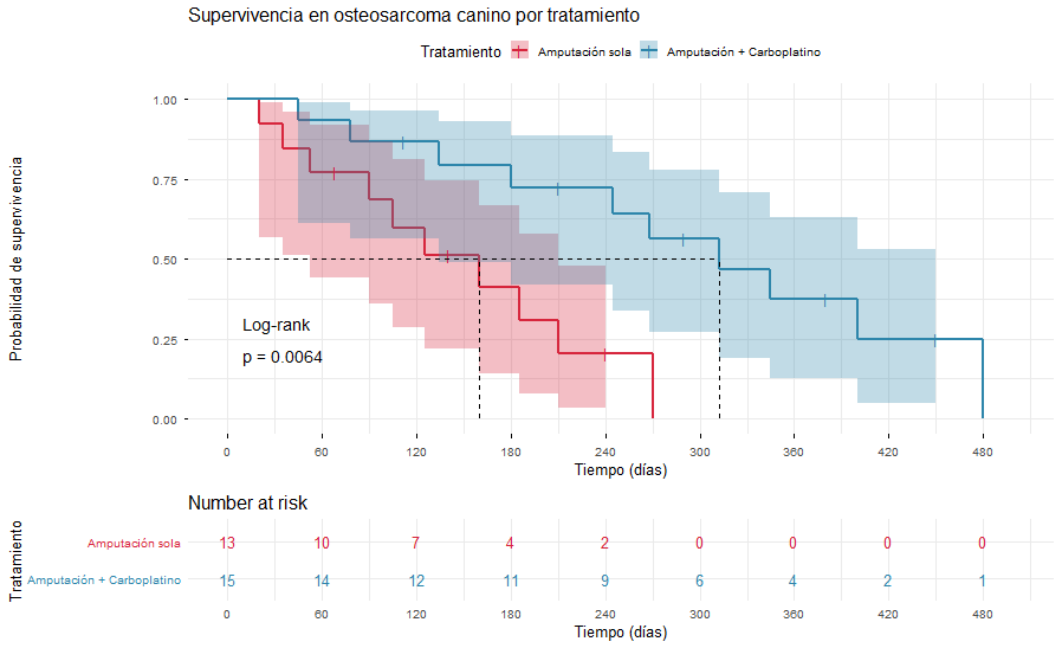
# ---- Prueba de log-rank con el paquete coin ----
# logrank_test() de coin proporciona la prueba exacta de permutación
# para el estadístico de log-rank, útil con muestras pequeñas
library(coin)
logrank_coin <- logrank_test(surv_osteo ~ factor(grupo),
                             data = osteo_perros,
                             distribution = "exact")
# distribution = "exact" calcula el valor p exacto por permutación
# distribution = "approximate" usa Monte Carlo con B permutaciones
# distribution = "asymptotic" usa la aproximación chi-cuadrado

print(logrank_coin)
# La versión exacta de coin es preferible con muestras pequeñas
# porque no depende de la aproximación asintótica chi-cuadrado.

# ---- Gráfico comparativo de curvas KM ----
ggsurvplot(km_osteo,
           data = osteo_perros,
           conf.int = TRUE,
           pval = TRUE,              # Mostrar valor p del log-rank
           pval.method = TRUE,       # Indicar qué prueba se usó
           risk.table = TRUE,
           risk.table.col = "strata",
           xlab = "Tiempo (días)",
           ylab = "Probabilidad de sobrevivencia",
           title = "Sobrevivencia en osteosarcoma canino por tratamiento",
           legend.title = "Tratamiento",
           legend.labs = c("Amputación sola",
                           "Amputación + Carboplatino"),
           palette = c("#D7263D", "#2E86AB"),
           ggtheme = theme_minimal(),
           censor.shape = "|",
           censor.size = 4,
           surv.median.line = "hv",
           break.x.by = 60)
```

Figura 16

Supervivencia en osteosarcoma canino

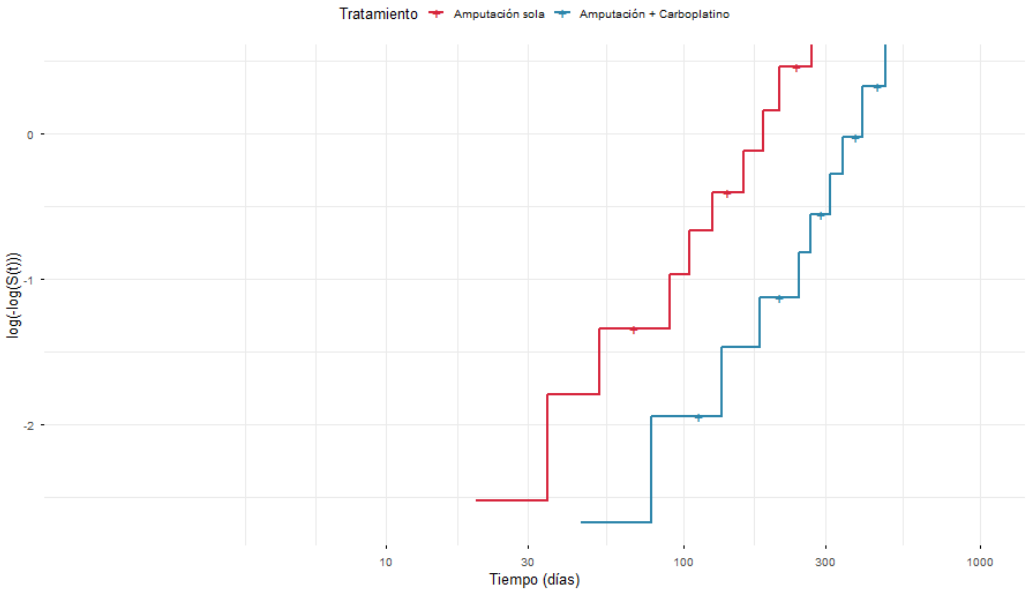


Nota. Descripción de la Figura 5.2: Dos curvas de sobrevivencia superpuestas. La curva roja (am putación sola) desciende más rápidamente, con la mayoría de los eventos antes del día 210. La cu rva azul (amputación + carboplatino) desciende más lentamente, con la mediana alrededor del dí a 290. Las bandas sombreadas representan los IC al 95% de cada grupo. Las marcas "|" indican o bservaciones censuradas. Debajo, la tabla de riesgo muestra cuántos sujetos permanecen en cada grupo a lo largo del tiempo. El valor p del log-rank se muestra en la esquina superior derecha.

```
# ---- Diagnóstico de riesgos proporcionales (gráfico log-log) ----
ggsurvplot(km_osteo,
           data = osteo_perros,
           fun = "cloglog",        # Transforma a log(-log(S(t)))
           xlab = "log(Tiempo en días)",
           ylab = "log(-log(S(t)))",
           title = "Diagnóstico de riesgos proporcionales",
           legend.title = "Tratamiento",
           palette = c("#D7263D", "#2E86AB"),
           ggtheme = theme_minimal())
```

Figura 17

Diagnóstico gráfico del supuesto de riesgos proporcionales



Nota. Descripción de la Figura 5.3: Si las dos curvas transformadas son aproximadamente parale las, el supuesto de riesgos proporcionales es razonable. Si se cruzan o convergen, el supuesto se viola y la prueba de log-rank puede perder potencia.

5.10.4. Prueba de Permutación: Alpacas con Suplementación Mineral

```
# ============================================================
# Ejemplo 3: Prueba de permutación - ganancia de peso en alpacas
# ============================================================

# Datos del ejemplo de la Sección 5.7
alpacas <- data.frame(
  grupo = factor(c(rep("Suplementada", 5), rep("Control", 5))),
  ganancia_kg = c(3.2, 4.1, 3.8, 5.0, 4.5,   # Suplementadas
                  2.1, 2.8, 3.0, 2.5, 3.3)    # Control
)

# Diferencia de medias observada
diferencia_obs <- mean(alpacas$ganancia_kg[alpacas$grupo == "Suplementada"]) -
                  mean(alpacas$ganancia_kg[alpacas$grupo == "Control"])
cat("Diferencia de medias observada:", diferencia_obs, "kg\n")
# Salida: Diferencia de medias observada: 1.38 kg

# ---- Prueba de permutación exacta con coin ----
```

```r
# oneway_test() realiza la prueba de permutación para dos o más grupos
# distribution = "exact" enumera todas las permutaciones posibles
perm_exacta <- oneway_test(ganancia_kg ~ grupo, data = alpacas,
                           distribution = "exact")
print(perm_exacta)
# Salida esperada:
# Exact Two-Sample Fisher-Pitman Permutation Test
#
# data:  ganancia_kg by grupo (Control, Suplementada)
# Z = -2.385, p-value = 0.01587
# alternative hypothesis: true mu is not equal to 0

# El valor p exacto es 0.016. Con 252 permutaciones posibles,
# solo 4 producen una diferencia tan extrema como la observada.

# ---- Prueba de permutación Monte Carlo ----
# Cuando las muestras son más grandes y la enumeración completa
# no es factible, se usa aproximación de Monte Carlo:
perm_mc <- oneway_test(ganancia_kg ~ grupo, data = alpacas,
                       distribution = approximate(nresample = 10000))
print(perm_mc)
# El valor p Monte Carlo será muy cercano al exacto (≈ 0.016)
# pero con leve variación aleatoria entre ejecuciones.

# ---- Prueba de permutación "a mano" para entender el mecanismo ----
set.seed(42)  # Reproducibilidad

datos_combinados <- alpacas$ganancia_kg
n1 <- 5  # Tamaño grupo suplementado
n_total <- 10

# Generar TODAS las permutaciones (252)
library(utils)
combinaciones <- combn(n_total, n1)  # Matriz 5 x 252

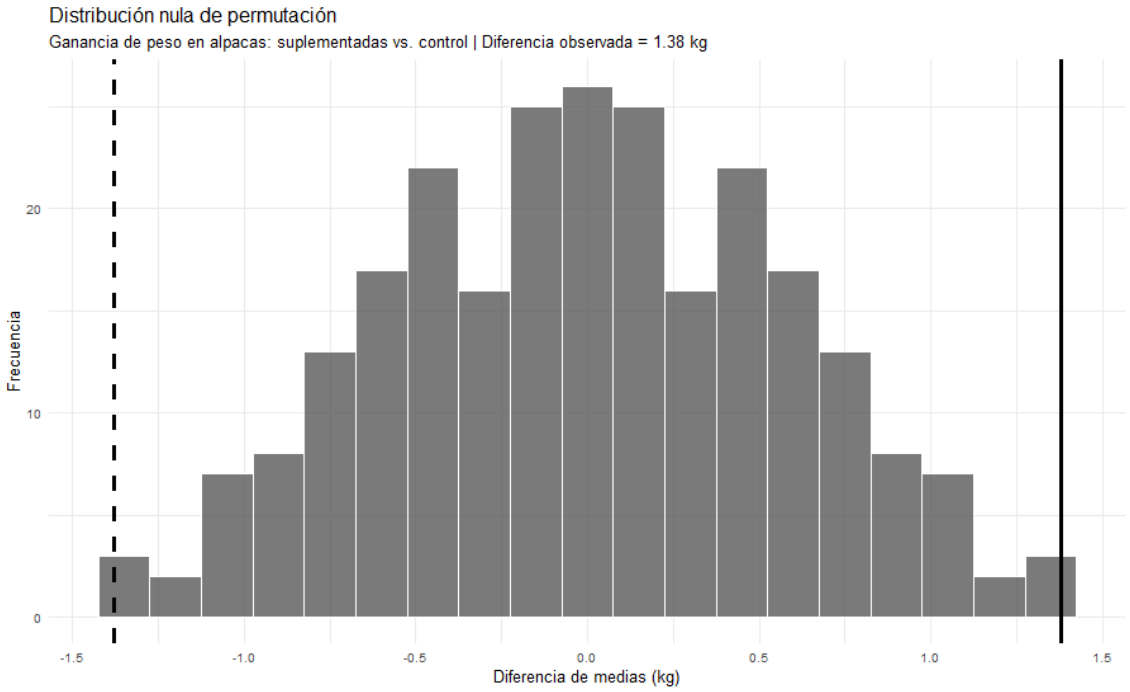
# Calcular la diferencia de medias para cada permutación
dif_perm <- apply(combinaciones, 2, function(indices) {
  mean(datos_combinados[indices]) - mean(datos_combinados[-indices])
})

# Valor p bilateral: proporción de diferencias >= |1.38| en valor absoluto
p_exacto_manual <- mean(abs(dif_perm) >= abs(diferencia_obs))
cat("Valor p exacto (calculado manualmente):", p_exacto_manual, "\n")
# Salida: Valor p exacto (calculado manualmente): 0.01587

# ---- Histograma de la distribución de permutación ----
df_perm <- data.frame(diferencia = dif_perm)
```

```
ggplot(df_perm, aes(x = diferencia)) +
  geom_histogram(binwidth = 0.15, fill = "#8ECAE6",
                 color = "white", alpha = 0.8) +
  geom_vline(xintercept = diferencia_obs, color = "#D7263D",
             linewidth = 1.2, linetype = "solid") +
  geom_vline(xintercept = -diferencia_obs, color = "#D7263D",
             linewidth = 1.2, linetype = "dashed") +
  annotate("text", x = diferencia_obs + 0.1, y = 30,
           label = paste0("Dif. observada = ", diferencia_obs),
           hjust = 0, color = "#D7263D", size = 3.5) +
  labs(x = "Diferencia de medias (permutación)",
       y = "Frecuencia",
       title = "Distribución nula de permutación",
       subtitle = "Ganancia de peso en alpacas: suplementadas vs. control") +
  theme_minimal()
```

Figura 18
Distribución nula de permutación



Nota. Descripción de la Figura 5.4: Histograma de las 252 diferencias de medias bajo permutació n. La distribución es simétrica y centrada en cero (como se espera bajo H0). La línea roja sólida m arca la diferencia observada (1.38 kg) y la línea roja punteada marca su negativo (-1.38 kg). Solo 4 de las 252 barras caen más allá de estos límites, lo que # corresponde al valor p = 4/252 = 0.016.

5.10.5. Bootstrap No Paramétrico: IgG Calostral en Vacas

```
# ============================================================
# Ejemplo 4: Bootstrap para IC de mediana de IgG calostral
# ============================================================

# Datos: IgG (mg/dL) en calostro de 12 vacas primíparas Holstein
igg_primiparas <- c(4520, 3890, 5110, 2780, 6340, 4150,
                    3200, 5670, 4880, 3560, 7210, 4430)

# Mediana observada
mediana_obs <- median(igg_primiparas)
cat("Mediana observada de IgG:", mediana_obs, "mg/dL\n")
# Salida: Mediana observada de IgG: 4475 mg/dL

# ---- Bootstrap con el paquete boot ----
# Paso 1: Definir la función del estadístico
# boot() requiere una función que reciba los datos y un vector
# de índices (las posiciones de la muestra bootstrap)
funcion_mediana <- function(datos, indices) {
  return(median(datos[indices]))
}

# Paso 2: Ejecutar el bootstrap
set.seed(123)  # Para reproducibilidad
resultado_boot <- boot(data = igg_primiparas,
                       statistic = funcion_mediana,
                       R = 10000)
# data: el vector de datos original
# statistic: la función que calcula el estadístico
# R: número de réplicas bootstrap

# Resumen del bootstrap
print(resultado_boot)
# Salida esperada:
# ORDINARY NONPARAMETRIC BOOTSTRAP
#
# Call: boot(data = igg_primiparas, statistic = funcion_mediana, R = 10000)
#
# Bootstrap Statistics :
#     original    bias    std. error
# t1*     4475   -28.34      549.26
#
# La columna "original" es la mediana observada.
# "bias" es la diferencia entre la media de las medianas bootstrap
```

```r
#   y la mediana observada. Un sesgo pequeño indica estabilidad.
# "std. error" es el error estándar bootstrap de la mediana.

# Paso 3: Intervalos de confianza
# boot.ci() calcula varios tipos de IC bootstrap
ic_boot <- boot.ci(resultado_boot,
                   conf = 0.95,    # Nivel de confianza
                   type = c("perc", "bca"))
# type = "perc": intervalo percentil
# type = "bca": intervalo BCa (corregido por sesgo y aceleración)

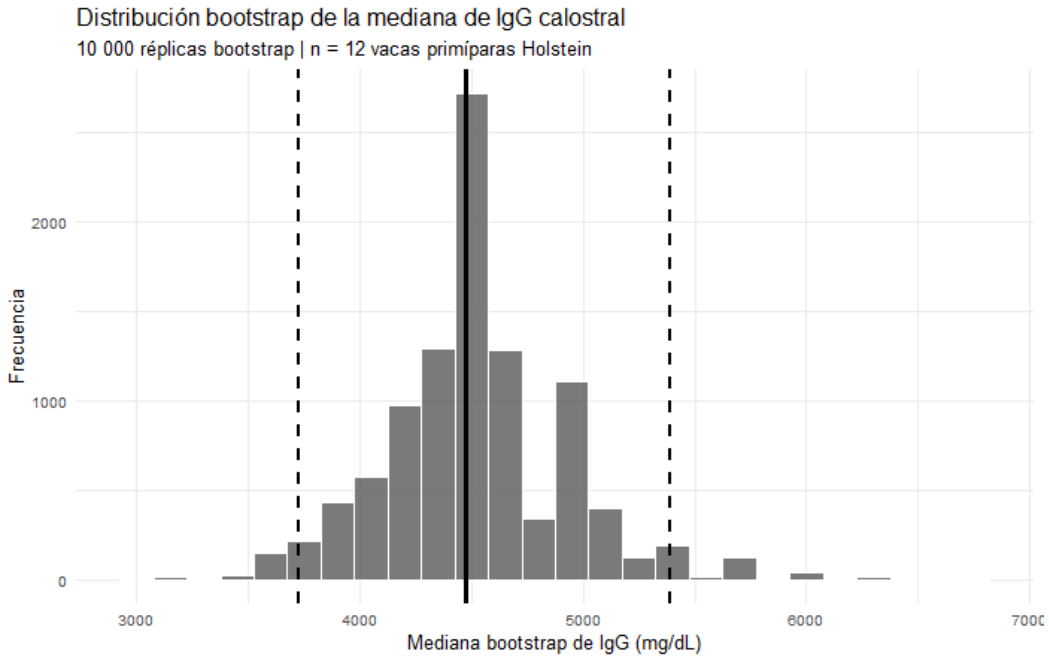
print(ic_boot)
# Salida esperada (valores aproximados):
# BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
# Based on 10000 bootstrap replicates
#
# Intervals :
# Level     Percentile          BCa
# 95%   ( 3560,  5670 )    ( 3560,  5670 )
#
# Con estos datos, los intervalos percentil y BCa son similares.
# En muestras más asimétricas, el BCa puede diferir notablemente.

# ---- Histograma de la distribución bootstrap ----
df_boot <- data.frame(mediana_boot = resultado_boot$t[, 1])

ggplot(df_boot, aes(x = mediana_boot)) +
  geom_histogram(binwidth = 150, fill = "#48BF84",
                 color = "white", alpha = 0.8) +
  geom_vline(xintercept = mediana_obs, color = "#D7263D",
             linewidth = 1.2) +
  geom_vline(xintercept = ic_boot$percent[4], color = "#2E86AB",
             linewidth = 1, linetype = "dashed") +
  geom_vline(xintercept = ic_boot$percent[5], color = "#2E86AB",
             linewidth = 1, linetype = "dashed") +
  annotate("text", x = mediana_obs + 100, y = 1500,
           label = paste0("Mediana obs = ", mediana_obs),
           hjust = 0, color = "#D7263D", size = 3.5) +
  labs(x = "Mediana bootstrap de IgG (mg/dL)",
       y = "Frecuencia",
       title = "Distribución bootstrap de la mediana de IgG calostral",
       subtitle = "10,000 réplicas bootstrap, vacas primíparas Holstein") +
  theme_minimal()
```

Figura 19

Distribución bootstrap de la mediana de IgG calostral



Nota. Descripción de la Figura 5.5: Histograma de las 10,000 medianas bootstrap. La distribució n es aproximadamente simétrica alrededor de la mediana observada (línea roja, 4475 mg/dL). La s líneas azules punteadas marcan los límites del IC al 95% percentil. La forma discreta del histog rama (con picos en ciertos valores) es normal: la mediana de 12 observaciones solo puede tomar un conjunto finito de valores según qué observaciones entren en la muestra bootstrap.

5.10.6. Bootstrap para Comparación de Dos Grupos

```
# ============================================================
# Ejemplo 5: Bootstrap para comparar IgG entre primíparas y
# multíparas - IC de la diferencia de medianas
# ============================================================

# Datos de IgG calostral (mg/dL) en vacas multíparas
igg_multiparas <- c(6890, 7450, 5980, 8120, 6230, 7810,
                    5540, 6750, 7290, 8540, 6100, 7030,
                    6480, 7760, 5870)

# Medianas observadas
cat("Mediana primíparas:", median(igg_primiparas), "mg/dL\n")
```

```r
cat("Mediana multíparas:", median(igg_multiparas), "mg/dL\n")
cat("Diferencia:", median(igg_multiparas) - median(igg_primiparas), "mg/dL\n")
# Salida esperada:
# Mediana primíparas: 4475 mg/dL
# Mediana multíparas: 6890 mg/dL
# Diferencia: 2415 mg/dL

# Función para la diferencia de medianas con bootstrap
# Necesita un diseño donde los dos grupos estén en un data.frame
igg_datos <- data.frame(
  igg = c(igg_primiparas, igg_multiparas),
  paridad = factor(c(rep("Primípara", length(igg_primiparas)),
                     rep("Multípara", length(igg_multiparas))))
)

# Función que calcula la diferencia de medianas
dif_medianas <- function(datos, indices) {
  d <- datos[indices, ]
  med_multi <- median(d$igg[d$paridad == "Multípara"])
  med_primi <- median(d$igg[d$paridad == "Primípara"])
  return(med_multi - med_primi)
}

# Bootstrap estratificado: se remuestrea DENTRO de cada grupo
# Esto preserva los tamaños de muestra de cada grupo
set.seed(456)
boot_dif <- boot(data = igg_datos,
                 statistic = dif_medianas,
                 R = 10000,
                 strata = igg_datos$paridad)
# strata: remuestreo estratificado por grupo

print(boot_dif)
# Salida esperada:
# Bootstrap Statistics :
#     original   bias    std. error
# t1*     2415  -74.21      702.34

ic_dif <- boot.ci(boot_dif, conf = 0.95, type = c("perc", "bca"))
print(ic_dif)
# Salida esperada (valores aproximados):
# Level     Percentile          BCa
# 95%   ( 1100,  3790 )    ( 1060,  3710 )
#
# El IC no incluye el cero, lo que indica que la diferencia de
# medianas de IgG calostral entre multíparas y primíparas es
# estadísticamente significativa. Las multíparas producen calostro
```

```
# con mayor concentración de IgG, con una diferencia estimada de
# 2,415 mg/dL (IC 95% BCa: 1,060 - 3,710).

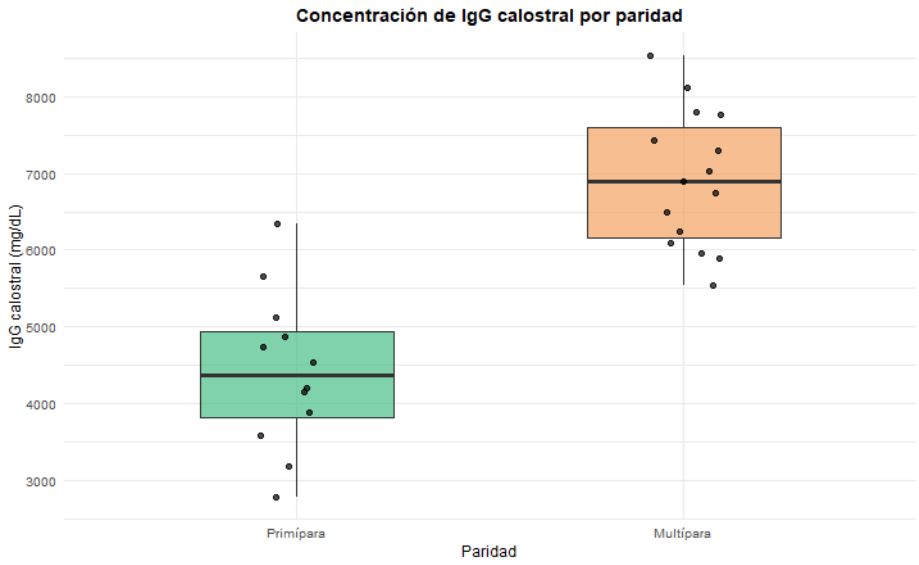
# Este resultado tiene relevancia clínica directa: los terneros
# nacidos de vacas primíparas pueden estar en mayor riesgo de
# falla en la transferencia de inmunidad pasiva si dependen
# exclusivamente del calostro materno.

# ---- Gráfico combinado: boxplot + distribución bootstrap ----
library(tidyr)

# Panel A: Boxplot comparativo
p1 <- ggplot(igg_datos, aes(x = paridad, y = igg, fill = paridad)) +
  geom_boxplot(alpha = 0.7, width = 0.5) +
  geom_jitter(width = 0.1, size = 2, alpha = 0.6) +
  scale_fill_manual(values = c("#48BF84", "#F4A261")) +
  labs(x = "Paridad", y = "IgG calostral (mg/dL)",
       title = "Concentración de IgG calostral por paridad") +
  theme_minimal() +
  theme(legend.position = "none")
```

Figura 20

Concentración de IgG calostral por paridad



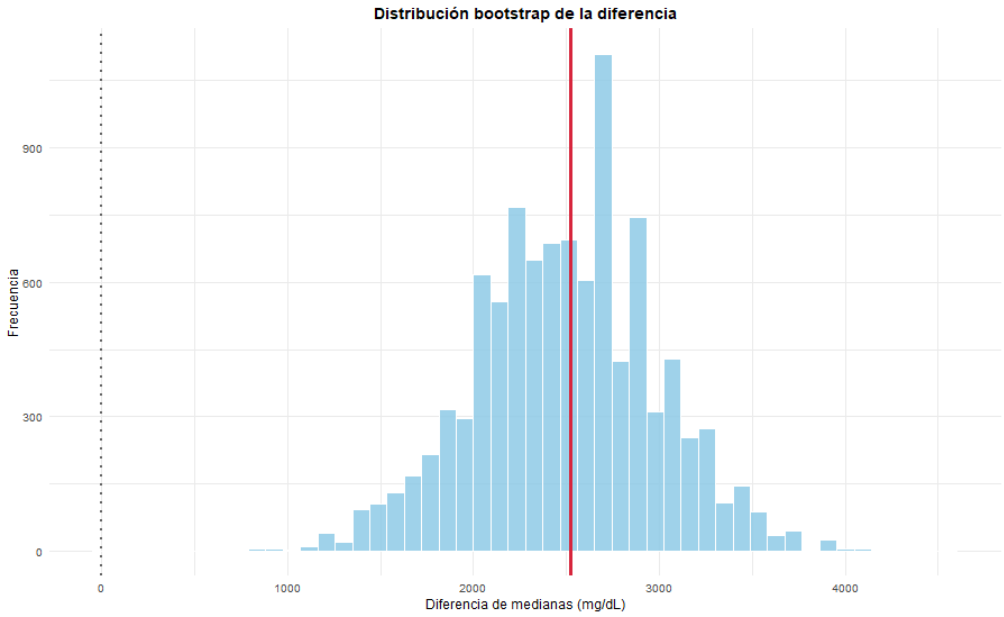
Nota. Panel A: Boxplot mostrando que las multíparas tienen mayor concentración de IgG con me nor dispersión relativa. Los puntos individuales superpuestos permiten ver la distribución real.

```
# Panel B: Distribución bootstrap de la diferencia
df_boot_dif <- data.frame(diferencia = boot_dif$t[, 1])
```

```
p2 <- ggplot(df_boot_dif, aes(x = diferencia)) +
  geom_histogram(binwidth = 200, fill = "#8ECAE6",
                 color = "white", alpha = 0.8) +
  geom_vline(xintercept = 0, color = "gray40",
             linewidth = 1, linetype = "dotted") +
  geom_vline(xintercept = 2415, color = "#D7263D", linewidth = 1.2) +
  labs(x = "Diferencia de medianas bootstrap (mg/dL)",
       y = "Frecuencia",
       title = "Distribución bootstrap de la diferencia") +
  theme_minimal()
```

Figura 21

Distribución bootstrap de la diferencia



Nota. Panel B: Histograma de las 10,000 diferencias bootstrap. La línea roja marca la diferencia observada (2,415 mg/dL). La línea punteada gris marca el cero. Toda la distribución bootstrap qu eda lejos del cero, confirmando la significancia.

5.11. Estudio de Caso Integrador

Sobrevivencia de Bovinos con Anaplasmosis y Bootstrap para IgG Calostral

Contexto del problema veterinario

En la parroquia San Simón, cantón Guaranda, provincia de Bolívar (Ecuador), un hato de 48 vacas mestizas Bos taurus × Bos indicus de doble propósito presenta un brote de anaplasmosis bovina (*Anaplasma marginale*) durante la transición de época seca a lluviosa (noviembre-enero), periodo de alta actividad de garrapatas *Rhipicephalus (Boophilus) microplus*. El diagnóstico se confirma por frotis sanguíneo con tinción Giemsa (rickettsemia ≥ 1%) y PCR convencional.

El veterinario asigna aleatoriamente las vacas a dos protocolos:

- **Grupo A** (n = 25): Imidocarb dipropionato, 3 mg/kg IM, dosis única, repetida a los 14 días si la rickettsemia persiste.
- **Grupo B** (n = 23): Oxitetraciclina LA, 20 mg/kg IM, cada 72 horas por 3 aplicaciones.

Se registra el tiempo (en días) desde el diagnóstico hasta la muerte o hasta el final del periodo de observación (90 días). Simultáneamente, como parte de un estudio paralelo, se evalúa la concentración de IgG en calostro de primera ordeña en vacas primíparas (n = 12) y multíparas (n = 15) del mismo hato, para evaluar la calidad inmunológica del calostro y su potencial impacto sobre los terneros neonatos.

Datos

```
# ================================================================
# ESTUDIO DE CASO INTEGRADOR
# ================================================================

# ---- Parte 1: Datos de sobrevivencia - anaplasmosis bovina ----
anaplasmosis <- data.frame(
  vaca = paste0("V-", sprintf("%03d", 1:48)),
  grupo = c(rep("Imidocarb", 25), rep("Oxitetraciclina", 23)),
  tiempo_dias = c(
    # Grupo Imidocarb (n = 25)
    14, 28, 45, 62, 73, 90, 34, 56, 90, 18,
    41, 67, 82, 90, 29, 53, 90, 38, 71, 85,
    22, 49, 90, 60, 77,
    # Grupo Oxitetraciclina (n = 23)
    7, 11, 21, 33, 48, 90, 15, 25, 39, 55,
```

```
    68, 90, 9, 19, 31, 44, 61, 80, 13, 27,
    42, 58, 90
  ),
  estado = c(
    # Grupo Imidocarb: 1 = muerte, 0 = censurado (vivo a los 90 días o perdido)
    1, 1, 0, 1, 1, 0, 1, 1, 0, 1,
    1, 1, 0, 0, 1, 1, 0, 1, 1, 0,
    1, 1, 0, 0, 1,
    # Grupo Oxitetraciclina
    1, 1, 0, 1, 1, 0, 1, 1, 1, 1,
    1, 0, 1, 1, 1, 1, 1, 0, 1, 1,
    1, 1, 0
  )
)

# Convertir grupo a factor
anaplasmosis$grupo <- factor(anaplasmosis$grupo,
                             levels = c("Oxitetraciclina", "Imidocarb"))

# Verificar la estructura
cat("Resumen por grupo:\n")
anaplasmosis %>%
  group_by(grupo) %>%
  summarise(
    n = n(),
    eventos = sum(estado),
    censurados = sum(estado == 0),
    mediana_tiempo = median(tiempo_dias),
    .groups = "drop"
  ) %>%
  print()

# ---- Kaplan-Meier por grupo ----
surv_anaplas <- Surv(time = anaplasmosis$tiempo_dias,
                     event = anaplasmosis$estado)

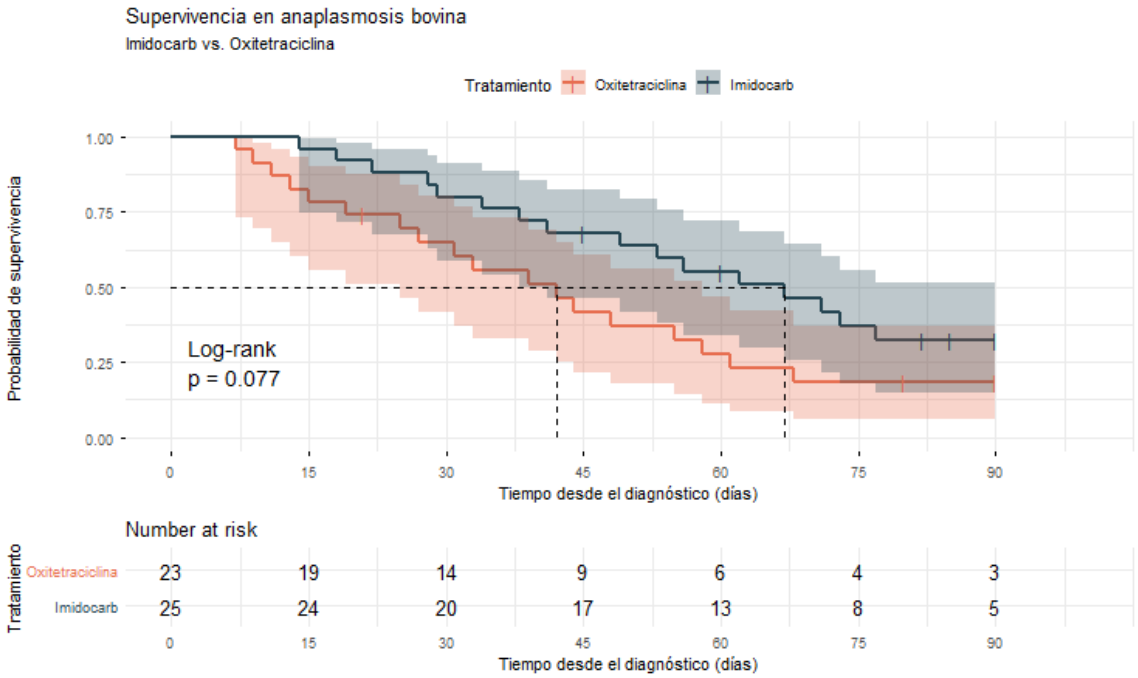
km_anaplas <- survfit(surv_anaplas ~ grupo, data = anaplasmosis,
                      conf.type = "log-log")
print(km_anaplas)

# ---- Gráfico de sobrevivencia ----
ggsurvplot(km_anaplas,
           data = anaplasmosis,
           conf.int = TRUE,
           pval = TRUE,
           pval.method = TRUE,
           risk.table = TRUE,
```

```
    risk.table.height = 0.25,
    xlab = "Tiempo desde diagnóstico (días)",
    ylab = "Probabilidad de sobrevivencia",
    title = "Sobrevivencia en anaplasmosis bovina: Imidocarb vs. Oxitetraciclina",
    subtitle = "Hato mestizo doble propósito, Guaranda, Ecuador",
    legend.title = "Tratamiento",
    legend.labs = c("Oxitetraciclina LA", "Imidocarb"),
    palette = c("#E76F51", "#264653"),
    ggtheme = theme_minimal(),
    censor.shape = "|",
    censor.size = 4,
    surv.median.line = "hv",
    break.x.by = 15)
```

Figura 22

Supervivencia en anaplasmosis bovina



Nota. Curvas de Kaplan-Meier comparativas.La curva del grupo oxitetraciclina (rojo) desciende más rápidamente que la del grupo imidocarb (azul oscuro), indicando mayor mortalidad tempran a. La tabla de riesgo inferior permite ver la reducción progresiva de sujetos.

```
# ---- Prueba de log-rank y variantes ----
# Log-rank estándar
lr_standard <- survdiff(surv_anaplas ~ grupo, data = anaplasmosis,
                        rho = 0)
cat("\n--- Prueba de Log-Rank estándar ---\n")
```

```
print(lr_standard)

# Peto-Peto
lr_peto <- survdiff(surv_anaplas ~ grupo, data = anaplasmosis,
                    rho = 1)
cat("\n--- Prueba de Peto-Peto ---\n")
print(lr_peto)

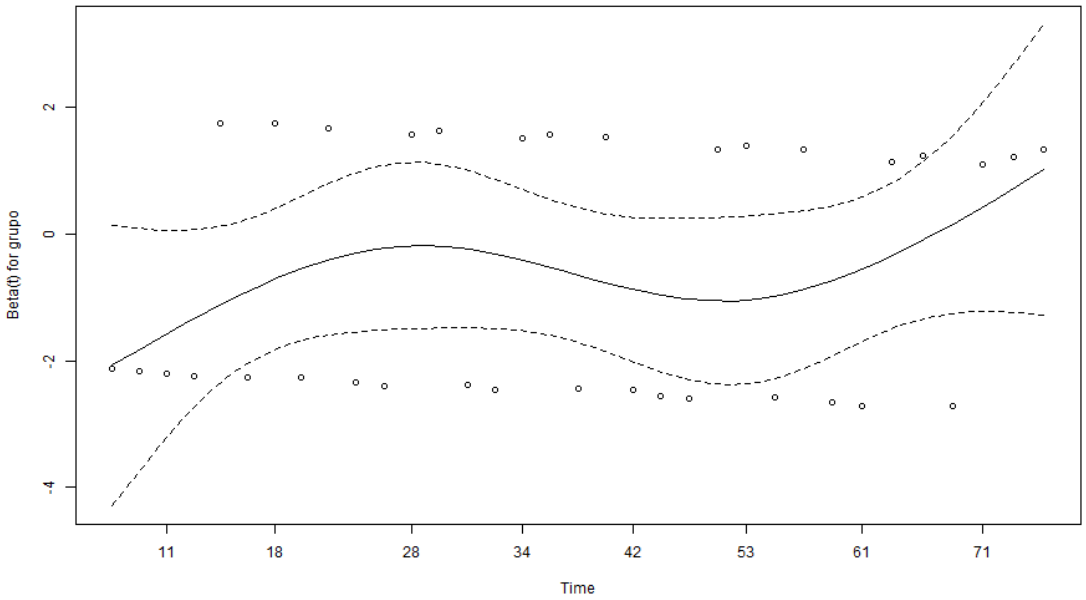
# Log-rank exacto (permutación) con coin
lr_exacto <- logrank_test(surv_anaplas ~ grupo, data = anaplasmosis,
                          distribution = approximate(nresample = 10000))
cat("\n--- Prueba de Log-Rank (permutación Monte Carlo) ---\n")
print(lr_exacto)

# ---- Diagnóstico de riesgos proporcionales ----
# Ajustar modelo de Cox para verificar el supuesto
cox_model <- coxph(surv_anaplas ~ grupo, data = anaplasmosis)
cox_zph <- cox.zph(cox_model)
print(cox_zph)
# Si p > 0.05, no se rechaza el supuesto de riesgos proporcionales

# Gráfico de residuos de Schoenfeld
plot(cox_zph)
```

Figura 23

Residuos de Schoenfeld escalados



Nota. Una línea horizontal indica que el supuesto de riesgos proporcionales se cumple. Una tend encia ascendente o descendente indicaría que el efecto del tratamiento cambia con el tiempo.

```
# ---- Discusión: ¿Qué pasaría con una prueba paramétrica? ----
# Si se ignorara la censura y se usara una prueba t:
# 1. Los animales censurados a los 90 días se tratarían como si
#    hubieran muerto en el día 90, subestimando la sobrevivencia
#    del grupo con más censurados.
# 2. Alternativamente, excluir los censurados reduce el tamaño
#    de muestra y pierde información.
# 3. Los tiempos de sobrevivencia raramente siguen una distribución
#    normal; típicamente son asimétricos a la derecha.
# Cualquiera de estas opciones introduce sesgo. El análisis de
# Kaplan-Meier maneja la censura correctamente sin asumir
# normalidad.

# ---- Parte 2: Bootstrap para IgG calostral ----
# (Datos del ejemplo de la Sección 5.10.5-5.10.6)

# Vacas primíparas
igg_primi <- c(4520, 3890, 5110, 2780, 6340, 4150,
               3200, 5670, 4880, 3560, 7210, 4430)

# Vacas multíparas
igg_multi <- c(6890, 7450, 5980, 8120, 6230, 7810,
               5540, 6750, 7290, 8540, 6100, 7030,
               6480, 7760, 5870)

# Medidas descriptivas
cat("\n--- IgG calostral (mg/dL) ---\n")
cat("Primíparas: mediana =", median(igg_primi),
    ", rango =", range(igg_primi), "\n")
cat("Multíparas: mediana =", median(igg_multi),
    ", rango =", range(igg_multi), "\n")

# Prueba de permutación para comparación
igg_all <- data.frame(
  igg = c(igg_primi, igg_multi),
  paridad = factor(c(rep("Primípara", 12), rep("Multípara", 15)))
)
```

```r
# Prueba exacta de permutación
perm_igg <- oneway_test(igg ~ paridad, data = igg_all,
                        distribution = "exact")
cat("\n--- Prueba de permutación exacta ---\n")
print(perm_igg)

# Bootstrap estratificado para IC de la diferencia de medianas
dif_med_func <- function(datos, indices) {
  d <- datos[indices, ]
  median(d$igg[d$paridad == "Multípara"]) -
    median(d$igg[d$paridad == "Primípara"])
}

set.seed(789)
boot_igg <- boot(data = igg_all, statistic = dif_med_func,
                 R = 10000, strata = igg_all$paridad)

ic_igg <- boot.ci(boot_igg, conf = 0.95, type = c("perc", "bca"))
cat("\n--- IC Bootstrap para diferencia de medianas ---\n")
print(ic_igg)

# ---- Interpretación integrada ----
cat("\n===== RESUMEN DEL ESTUDIO DE CASO =====\n")
cat("Parte 1 - Anaplasmosis:\n")
cat("  El grupo tratado con imidocarb muestra mayor sobrevivencia\n")
cat("  que el grupo con oxitetraciclina. La prueba de log-rank y\n")
cat("  su versión de permutación confirman la diferencia.\n\n")
cat("Parte 2 - IgG calostral:\n")
cat("  Las vacas multíparas producen calostro con mayor\n")
cat("  concentración de IgG que las primíparas.\n")
cat("  La diferencia de medianas es de",
    median(igg_multi) - median(igg_primi), "mg/dL\n")
cat("  IC 95% BCa obtenido por bootstrap.\n")
cat("  Los terneros de primíparas pueden necesitar calostro\n")
cat("  suplementario o un banco de calostro para asegurar\n")
cat("  una transferencia de inmunidad pasiva adecuada.\n")
```

Interpretación en Términos Veterinarios

Los resultados del análisis de sobrevivencia indican que el protocolo con imidocarb dipropionato confiere una ventaja en la sobrevivencia de vacas con anaplasmosis bovina respecto al protocolo

con oxitetraciclina de larga acción. La separación temprana de las curvas de Kaplan-Meier sugiere que la diferencia es más marcada en las primeras semanas post-diagnóstico, lo cual es coherente con el mecanismo de acción del imidocarb, que tiene actividad babesicida y anaplasmicida directa, frente a la oxitetraciclina, cuyo efecto es bacteriostático y requiere un sistema inmune funcional para la eliminación del agente.

Desde la perspectiva del manejo del hato, la segunda parte del estudio la comparación de IgG calostral añade una capa de complejidad: si las vacas primíparas que sobreviven al brote producen calostro de menor calidad inmunológica, sus terneros nacerán con doble vulnerabilidad: en un ambiente con alta carga de *Anaplasma* y con menor protección pasiva. El intervalo de confianza bootstrap para la diferencia de medianas de IgG (que no incluye el cero) respalda la recomendación de implementar un banco de calostro con donantes multíparas.

5.12. Errores Frecuentes y Advertencias

1. Confundir censura con datos faltantes. Una observación censurada no es un dato perdido. Es una observación que aporta información parcial: se sabe que el animal sobrevivió *al menos* hasta el tiempo de censura. Excluir las observaciones censuradas del análisis sesga los resultados y desperdicia información. El estimador de Kaplan-Meier está diseñado precisamente para incorporar esta información parcial.

2. Asumir censura no informativa sin evaluarlo. Si los animales que se pierden del seguimiento tienen peor pronóstico que los que permanecen, la función de sobrevivencia estimada por KM sobreestima la sobrevivencia real. Antes de aplicar KM, hay que preguntarse: ¿por qué se perdieron estos animales? ¿Los que se perdieron se parecen a los que se quedaron? No existe prueba formal para esto; se evalúa por razonamiento clínico y comparando las características basales de los censurados con los no censurados.

3. Usar la prueba de log-rank cuando las curvas se cruzan. Si las curvas de sobrevivencia de dos grupos se cruzan, la prueba de log-rank puede tener potencia muy baja porque las diferencias en una dirección se cancelan con las diferencias en la otra. En esta situación, hay que considerar pruebas diseñadas para alternativas de riesgos no proporcionales, o segmentar el análisis por periodos.

4. Reportar solo el valor p sin la mediana y su IC. El resultado de un análisis de sobrevivencia no es solo "p < 0.05". La mediana de sobrevivencia y su intervalo de confianza son clínicamente más informativos. "La mediana de sobrevivencia fue de 290 días (IC 95%: 134-450) en el grupo con quimioterapia adyuvante, frente a 125 días (IC 95%: 52-210) en el grupo quirúrgico solo" dice mucho más que "p = 0.095".

5. Interpretar el bootstrap como un sustituto del tamaño de muestra. El bootstrap no crea información nueva: remuestrea la información existente. Si la muestra original es de 8 animales, generar 10,000 réplicas bootstrap no equivale a tener 10,000 animales. El bootstrap estima la variabilidad del estadístico, no la del dato individual. Con muestras muy pequeñas (n < 10), los intervalos bootstrap pueden tener cobertura real inferior a la nominal.

6. Olvidar la semilla aleatoria en los reportes. Los métodos Monte Carlo (pruebas de permutación aproximadas y bootstrap) producen resultados que varían ligeramente entre ejecuciones. Para la reproducibilidad, siempre debe fijarse y reportarse la semilla aleatoria (*set.seed()*). Dos investigadores que ejecuten el mismo código con la misma semilla obtendrán resultados idénticos.

7. No usar bootstrap estratificado al comparar grupos. Cuando se comparan dos grupos con bootstrap, el remuestreo debe hacerse **dentro** de cada grupo (estratificado), no de la muestra completa. Si se remuestrea sin estratificación, una muestra bootstrap podría contener 20 observaciones del grupo 1 y 7 del grupo 2, lo que no refleja el diseño del estudio.

8. Confundir la prueba de permutación con el bootstrap. La permutación baraja las etiquetas de grupo para generar una distribución nula (prueba de hipótesis). El bootstrap remuestrea con reemplazo dentro de los datos para estimar la variabilidad de un estadístico (estimación por intervalo). No son intercambiables.

5.13. Ejercicios de Autoevaluación

Nivel Básico

Ejercicio 5.1.

En un estudio de sobrevivencia posquirúrgica de 30 perros con hemangiosarcoma esplénico, 8 perros están vivos al final del periodo de observación y 3 fueron eutanasiados por causas no relacionadas con el hemangiosarcoma. ¿Cuántas observaciones censuradas hay en este estudio? ¿Qué tipo de censura representa cada caso?

Respuesta: Hay 11 observaciones censuradas (8 + 3). Los 8 perros vivos al cierre del estudio representan censura administrativa (derecha). Los 3 eutanasiados por otras causas representan censura por riesgo competitivo (también censura por la derecha). Ambos tipos se manejan de la misma forma en el estimador de Kaplan-Meier, bajo el supuesto de censura no informativa.

Ejercicio 5.2.

Un investigador tiene datos de sobrevivencia de vacas con mastitis clínica y quiere comparar dos protocolos antibióticos. El evento de interés es la curación bacteriológica (cultivo negativo). ¿Es correcto usar una prueba de log-rank? Justifique su respuesta.

Respuesta: Sí, es correcto. La curación bacteriológica es un evento bien definido, y el tiempo hasta la curación puede analizarse con métodos de sobrevivencia. Las vacas que no se curan durante el periodo de observación serían censuradas. La prueba de log-rank compararía las curvas de "probabilidad de permanecer sin curación" entre los dos protocolos. La única precaución es verificar que el supuesto de riesgos proporcionales sea razonable (las tasas de curación en ambos grupos mantienen una relación constante a lo largo del tiempo).

Ejercicio 5.3.

¿Cuál es la diferencia conceptual entre una prueba de permutación y el bootstrap no paramétrico? Indique para qué sirve cada uno.

Respuesta: La prueba de permutación genera una distribución nula reordenando las etiquetas de grupo de los datos observados; sirve para obtener un valor p (prueba de hipótesis: ¿hay diferencia entre grupos?). El bootstrap genera múltiples muestras con reemplazo a partir de los datos; sirve para estimar la variabilidad de un estadístico y construir intervalos de confianza (estimación: ¿cuán grande es el efecto y con qué incertidumbre?).

Nivel Intermedio

Ejercicio 5.4.

A continuación, se presentan los tiempos de sobrevivencia (en meses) de 8 ovinos Corriedale con linfadenitis caseosa (*Corynebacterium pseudotuberculosis*) tratados con autovacuna:

Ovino	Tiempo (meses)	Estado
O-01	2	1
O-02	4	1
O-03	6	0
O-04	7	1
O-05	9	1
O-06	12	0
O-07	14	1
O-08	18	1

Calcule a mano el estimador de Kaplan-Meier y determine la mediana de sobrevivencia.

Respuesta:

t_i	n_i	d_i	$1 - d_i/n_i$	$\hat{S}(t_i)$
2	8	1	7/8 = 0.875	0.875
4	7	1	6/7 = 0.857	0.750
7	5	1	4/5 = 0.800	0.600
9	4	1	3/4 = 0.750	0.450
14	2	1	1/2 = 0.500	0.225
18	1	1	0/1 = 0.000	0.000

Nota. en $t = 7$, $n_i = 5$ porque O-03 fue censurado en $t = 6$. En $t = 14$, $n_i = 2$ porque O-06 fue censurado en $t = 12$.

La función de sobrevivencia cruza el 50% entre $t = 7$ ($\hat{S} = 0.600$) y $t = 9$ ($\hat{S} = 0.450$). La mediana de sobrevivencia es **9 meses**.

Ejercicio 5.5.

Un veterinario en Cochabamba, Bolivia, mide el recuento de huevos por gramo (HPG, técnica McMaster) en 7 alpacas tratadas con albendazol y 7 alpacas control, 14 días post-tratamiento:

- Albendazol: 50, 0, 25, 0, 75, 0, 100
- Control: 325, 450, 200, 575, 150, 400, 275

Realice una prueba de permutación con Monte Carlo (B = 5,000) usando R y reporte el valor p. Incluya el código completo.

Respuesta:

```
hpg <- data.frame(
  grupo = factor(c(rep("Albendazol", 7), rep("Control", 7))),
  hpg = c(50, 0, 25, 0, 75, 0, 100,
          325, 450, 200, 575, 150, 400, 275)
)

library(coin)
set.seed(2024)
resultado <- oneway_test(hpg ~ grupo, data = hpg,
                         distribution = approximate(nresample = 5000))
print(resultado)
# El valor p será muy bajo (< 0.001), indicando diferencia
# significativa en los recuentos HPG entre grupos.
```

Ejercicio 5.6.

Explique por qué la prueba de Breslow (Gehan-Breslow-Wilcoxon) sería preferible a la de log-rank estándar en un estudio donde se espera que un tratamiento antiviral para distemper canino reduzca la mortalidad en la primera semana, pero no tenga efecto después.

Respuesta: La prueba de Breslow asigna pesos proporcionales al número de sujetos en riesgo ($w_i = n_i$), que es máximo al inicio del seguimiento cuando todos los sujetos están vivos. Si la diferencia entre tratamientos se concentra en la primera semana (cuando n_i es grande), Breslow dará más peso a esos tiempos donde la diferencia es real, mientras que la prueba de log-rank asignará el mismo peso a todos los tiempos, diluyendo la señal temprana con los tiempos tardíos donde no hay diferencia. En este escenario, Breslow tiene mayor potencia para detectar la diferencia.

Nivel Avanzado

Ejercicio 5.7.

Un estudio multicéntrico en clínicas veterinarias de Lima, Quito y Bogotá registra los tiempos de sobrevivencia (en semanas) de 45 perros con insuficiencia cardiaca congestiva secundaria a enfermedad mixomatosa mitral, clasificados según estadio ACVIM (B2, C, D). El investigador le proporciona los datos crudos (disponibles en el archivo complementario del capítulo) y le solicita:

a) Construir curvas de Kaplan-Meier por estadio ACVIM.
b) Comparar las curvas con la prueba de log-rank y la prueba de Peto-Peto.
c) Verificar el supuesto de riesgos proporcionales con el gráfico log-log y la prueba de Grambsch-Therneau.
d) Calcular la mediana de sobrevivencia y su IC al 95% para cada estadio.
e) Si los datos lo permiten, usar bootstrap para estimar el IC de la diferencia de medianas entre los estadios B2 y D.
f) Redactar un párrafo de resultados como aparecería en la sección "Resultados" de un artículo en el *Journal of Veterinary Internal Medicine*, incluyendo la cita estadística.

Rúbrica orientadora:

- **Análisis correcto** (40%): Código R sintácticamente correcto, uso adecuado de *survfit()*, *survdiff()*, *cox.zph()* y *boot()*. Semilla aleatoria fijada. Bootstrap estratificado.
- **Interpretación estadística** (25%): Identificación correcta de cuáles estadios difieren. Discusión sobre el supuesto de riesgos proporcionales. Reconocimiento de limitaciones del tamaño de muestra por estadio.
- **Interpretación clínica** (20%): Conexión entre los estadios ACVIM y la progresión esperada de la enfermedad. Discusión sobre la utilidad clínica de las medianas de sobrevivencia para el pronóstico.
- **Redacción científica** (15%): Formato APA, estadísticos reportados entre paréntesis con símbolos correctos, medianas e IC incluidos.

Ejercicio 5.8.

Un investigador en una granja avícola en Santander, Colombia, quiere evaluar si la adición de un probiótico al agua de bebida reduce la mortalidad en pollos de engorde Ross 308 durante un brote de *Salmonella* Enteritidis. Tiene 200 pollos por tratamiento, pero solo puede registrar la mortalidad semanalmente (no diariamente). ¿Qué tipo de censura presentan estos datos? ¿Es adecuado el estimador de Kaplan-Meier estándar? Proponga una alternativa si no lo es.

Rúbrica orientadora:

- Identificación de censura por intervalo (el evento muerte ocurre entre dos inspecciones semanales, pero no se sabe el día exacto).
- Reconocimiento de que KM estándar asume tiempos exactos. Con censura por intervalo, KM puede aplicarse si los intervalos son cortos relativos al tiempo total de observación (aproximación razonable), pero el método correcto sería un modelo de intervalo censurado (*interval-censored*).
- Discusión sobre si, con 200 pollos por grupo, una prueba de permutación sería más adecuada que la prueba de log-rank asintótica (respuesta: con n = 200, la aproximación asintótica es confiable; la permutación no aporta ventaja sustancial).

5.14. Conclusiones del Capítulo

El análisis de sobrevivencia y los métodos de remuestreo comparten una premisa central: es posible extraer inferencias válidas de datos complejos sin forzarlos dentro de marcos distribucionales rígidos.

El estimador de Kaplan-Meier resuelve un problema que ninguna otra técnica básica aborda: estimar la probabilidad de sobrevivencia cuando la información sobre algunos sujetos es incompleta. La censura inevitable en la investigación veterinaria clínica y de campo deja de ser un obstáculo y se convierte en información aprovechable. La prueba de log-rank y sus variantes ponderadas permiten comparar curvas de sobrevivencia entre grupos con diferentes sensibilidades a la ubicación temporal de las diferencias.

Las pruebas de permutación ofrecen una alternativa cuando los supuestos de las pruebas paramétrica o asintóticas son cuestionables. Con muestras de 5 a 15 animales por grupo tamaños

comunes en estudios veterinarios experimentales, la distribución exacta de permutación es más confiable que una aproximación chi-cuadrado o una distribución t.

El bootstrap no paramétrico cierra un vacío que las pruebas de hipótesis dejan abierto: la estimación por intervalo para estadísticos que no tienen fórmula analítica accesible. La mediana de sobrevivencia, la diferencia de medianas entre grupos, un coeficiente de Spearman, la razón de tasas de parasitación todos estos parámetros pueden acompañarse de un intervalo de confianza bootstrap sin asumir normalidad.

Ninguno de estos métodos es libre de supuestos. La censura debe ser no informativa. Las pruebas de permutación requieren intercambiabilidad bajo la nula. El bootstrap necesita una muestra representativa y un tamaño suficiente para que la distribución empírica se parezca a la poblacional. Olvidar estos supuestos o, peor, ignorar que existen lleva a conclusiones tan erradas como aplicar una prueba t a datos que no cumplen sus condiciones.

El Capítulo 6, dedicado a la comparación de más de dos grupos independientes mediante la prueba de Kruskal-Wallis y sus comparaciones post hoc, retoma la comparación de grupos donde la estructura no es temporal sino categórica, y donde la censura no está presente pero sí la heterogeneidad de varianzas, los empates y la necesidad de identificar cuáles grupos difieren después de un resultado global significativo.

Capítulo 6

Datos ordinales, escalas de evaluación clínica y tamaño del efecto

Introducción

Este capítulo trabaja dos problemas que convergen en la práctica del investigador veterinario. El primero es el análisis de datos ordinales, que son los que producen las escalas de evaluación clínica de uso cotidiano en veterinaria: condición corporal (Edmonson, Henneke, WSAVA), claudicación (Sprecher, AAEP), dolor posoperatorio (Glasgow, UNESP-Botucatu, Horse Grimace Scale), diarrea neonatal (Larson), grado de endometritis y clasificación histológica de tumores. Se establece la diferencia entre escala ordinal y escala de intervalo y sus consecuencias analíticas: por qué la media aritmética no tiene sentido para datos ordinales, por qué el ANOVA no es apropiado y qué alternativas existen. Se desarrolla la regresión ordinal mediante el modelo de logit acumulativo (odds proporcionales), con la verificación del supuesto de proporcionalidad y la interpretación de los coeficientes en términos clínicos. Se trabajan las medidas de concordancia entre evaluadores: kappa de Cohen para dos evaluadores, kappa ponderada para escalas ordinales donde los desacuerdos de distinto grado tienen distinto peso, y kappa de Fleiss para tres o más evaluadores. El segundo problema es transversal a todos los capítulos anteriores: el tamaño del efecto y la potencia estadística. Se desarrollan las medidas de tamaño del efecto específicas para cada prueba no paramétrica (r de rango biserial, A de Vargha y Delaney, épsilon cuadrada, eta cuadrada basado en H, W de Kendall, V de Cramér), la eficiencia relativa asintótica de las pruebas no paramétricas frente a sus equivalentes paramétricos, los métodos de cálculo de tamaño muestral para diseños con análisis no paramétrico, y la estimación de potencia por simulación en R para situaciones donde no existen fórmulas cerradas.

Objetivos de aprendizaje

Al finalizar este capítulo, el lector será capaz de:

1. Identificar las escalas de evaluación clínica ordinal de uso frecuente en medicina veterinaria y explicar por qué su naturaleza ordinal condiciona la elección del método estadístico.
2. Contrastar las propiedades de una escala ordinal frente a una de intervalo, y reconocer las consecuencias analíticas de tratar una como si fuera la otra.
3. Aplicar modelos de regresión ordinal (logit acumulativo) a datos veterinarios, verificar el supuesto de odds proporcionales e interpretar los coeficientes en términos clínicos.

4. Calcular e interpretar coeficientes de acuerdo entre evaluadores (kappa de Cohen, kappa ponderada, kappa de Fleiss) en contextos de evaluación clínica veterinaria.
5. Seleccionar, calcular e interpretar la medida de tamaño del efecto adecuada para cada prueba no paramétrica, y estimar la potencia estadística y el tamaño muestral requerido mediante fórmulas y simulación en R.

6.1. Escalas ordinales en veterinaria: condición corporal, claudicación, dolor, diarrea neonatal, endometritis

La investigación veterinaria depende, con una frecuencia que pocos reconocen, de variables que no se miden con instrumentos calibrados sino con el juicio entrenado del observador. Esas evaluaciones producen datos ordinales: categorías ordenadas que reflejan grados de intensidad, gravedad o calidad, pero cuyos intervalos no son cuantificables.

Condición corporal (BCS)

La escala de condición corporal es probablemente la herramienta de evaluación ordinal más extendida en producción animal. En bovinos, la escala más usada en Norteamérica y gran parte de Latinoamérica va de 1 (emaciado) a 5 en incrementos de 0.25 (Edmonson et al., 1989), aunque en algunas regiones se trabaja con la escala australiana de 1 a 8 o la neozelandesa de 1 a 10. En perros y gatos, la escala del World Small Animal Veterinary Association (WSAVA) emplea 9 categorías. En equinos se utiliza el sistema de Henneke con valores de 1 a 9 (Henneke et al., 1983).

Lo que estas escalas tienen en común es que un observador palpa o inspecciona al animal y asigna un número que representa una categoría. La diferencia entre BCS 2 y BCS 3 en una vaca no es necesariamente la misma cantidad de reserva grasa que la diferencia entre BCS 4 y BCS 5. El número ordena, pero no mide en unidades constantes.

Escalas de claudicación

En bovinos, la escala de Sprecher et al. (1997) asigna valores de 1 (normal) a 5 (postración). En equinos, la escala de la American Association of Equine Practitioners (AAEP) va de 0 a 5 grados. En ambos casos, el evaluador observa la locomoción del animal y clasifica la gravedad. La

diferencia entre grado 1 y grado 2 puede ser sutil (una ligera asimetría en el paso), mientras que la diferencia entre grado 4 y grado 5 implica un salto enorme en términos de dolor y funcionalidad.

Escalas de dolor

La evaluación del dolor en animales enfrenta el problema adicional de que el paciente no puede comunicar verbalmente su experiencia. Se han desarrollado escalas compuestas que integran múltiples indicadores conductuales y fisiológicos:

- **Escala de dolor de Glasgow (forma corta)** para perros: puntajes de 0 a 24, construida como suma de ítems ordinales individuales (Reid et al., 2007).
- **Escala multidimensional de la UNESP-Botucatu** para gatos: diseñada específicamente para dolor posoperatorio, con categorías que evalúan postura, vocalización, reacción a la palpación y comportamiento misceláneo (Brondani et al., 2013).
- **Escala de muecas equina (Horse Grimace Scale)**: basada en unidades de acción facial, con puntuación de 0 a 2 para cada unidad (Dalla Costa et al., 2014).

Estas escalas producen datos que son sumas de ítems ordinales. Tratarlas como continuas es una práctica extendida que genera resultados cuestionables, particularmente cuando el rango total es estrecho y la distribución es marcadamente asimétrica.

Puntaje de diarrea neonatal

En terneros, la escala de Larson et al. (1977) clasifica la consistencia fecal de 0 (normal) a 3 (diarrea acuosa severa). Cuatro categorías. Calcular una media de 1.7 para un grupo de terneros tratados no tiene una interpretación biológica directa: no existe una consistencia fecal de 1.7.

Grado de endometritis citológica

La clasificación de endometritis basada en el porcentaje de neutrófilos en citología endometrial utiliza grados ordinales: 0 (< 5% neutrófilos), 1 (5-15%), 2 (15-50%), 3 (> 50%). Los umbrales son arbitrarios desde el punto de vista biológico; la respuesta inflamatoria es un continuo que se discretiza para facilitar la comunicación clínica. El porcentaje original de neutrófilos podría

analizarse como variable continua, pero una vez convertido en grados, el análisis debe respetar la naturaleza ordinal.

Clasificación histológica de tumores

Los grados de diferenciación tumoral (bien diferenciado, moderadamente diferenciado, poco diferenciado, indiferenciado) y los estadios clínicos (I, II, III, IV) son variables ordinales que aparecen constantemente en oncología veterinaria. Los estudios de sobrevivencia en perros con mastocitoma, linfoma o melanoma oral emplean estas clasificaciones como variables predictoras, y su análisis requiere métodos que respeten el orden sin asumir equidistancia.

Lo que une a todas estas escalas es una propiedad simple pero determinante: el orden de las categorías tiene significado, pero la magnitud de la diferencia entre categorías consecutivas no es constante ni conocida. Esta propiedad descarta cualquier operación aritmética que asuma intervalos iguales, y eso incluye la media, la desviación estándar y cualquier prueba construida sobre ellas.

6.2. Escala ordinal vs. escala de intervalo: la diferencia que cambia el análisis

La distinción entre escalas de medición propuesta por Stevens (1946) nominal, ordinal, intervalo, razón se enseña en el primer curso de estadística y se olvida en el primero de investigación. Es una de esas clasificaciones que los estudiantes memorizan para el examen y abandonan cuando se enfrentan a datos reales. El resultado es que una proporción preocupante de artículos veterinarios trata datos ordinales como si fueran de intervalo.

Qué se puede hacer con cada escala

Con una escala de **intervalo** (o de razón), la distancia entre dos valores cualesquiera tiene significado constante. La diferencia entre 36 °C y 37 °C de temperatura rectal es la misma que entre 39 °C y 40 °C: un grado centígrado en ambos casos. Esto permite calcular medias, desviaciones estándar, coeficientes de variación, y usar pruebas que dependen de esas operaciones (t de Student, ANOVA, regresión lineal ordinaria).

Con una escala **ordinal**, solo se sabe que las categorías tienen un orden: grado 1 < grado 2 < grado 3. No se sabe cuánto mayor es grado 3 respecto a grado 2 en comparación con cuánto mayor es

grado 2 respecto a grado 1. Eso descarta la suma, la resta, la media aritmética y cualquier operación que dependa de la magnitud de las diferencias.

Las consecuencias prácticas

Calcular la media de una variable ordinal no genera un error de ejecución en R ni en ningún software. Los números se suman y se dividen sin protestar. El problema es de interpretación: la media de 2.4 en una escala de claudicación de 1 a 5 no corresponde a un estado fisiológico definido, y dos grupos con la misma media pueden tener distribuciones de puntajes completamente distintas.

Un ejemplo concreto. Diez vacas evaluadas con una escala de claudicación de 1 a 5:

- **Grupo A**: 1, 1, 1, 2, 2, 3, 3, 3, 4, 5 → media = 2.5, mediana = 2.5
- **Grupo B**: 1, 2, 2, 2, 3, 3, 3, 3, 3, 3 → media = 2.5, mediana = 3.0

La media es idéntica. La mediana difiere. Pero lo más relevante es que el Grupo A tiene animales en los extremos de la escala (incluyendo un grado 5, que indica una vaca que rehúsa caminar), mientras que el Grupo B se concentra en grados intermedios. La media no captura esa diferencia; la distribución completa de frecuencias sí.

¿Cuándo es aceptable tratar una escala ordinal como numérica?

Existe un debate legítimo en la literatura. Algunos investigadores argumentan que, con escalas de 7 o más categorías y distribuciones razonablemente simétricas, las pruebas paramétricas pueden funcionar de manera aceptable en términos de error tipo I (Norman, 2010; Sullivan & Artino, 2013). Otros sostienen que la robustez observada en simulaciones no justifica la práctica porque los estimadores (medias, diferencias de medias) carecen de interpretabilidad (Jamieson, 2004; Kuzon et al., 1996).

La postura de este libro es pragmática: si la escala tiene 5 o menos categorías, no hay discusión, el análisis debe ser ordinal o no paramétrico. Si la escala tiene 7 o más categorías y la distribución es razonablemente simétrica, un análisis paramétrico puede producir inferencias válidas sobre el error tipo I, pero los estimadores puntuales (medias, diferencias) siguen siendo de interpretación dudosa. En ese caso, reportar medianas e intervalos intercuartílicos como descriptivos y utilizar pruebas de rangos o regresión ordinal sigue siendo la opción más defensible.

El efecto techo y el efecto piso

Las escalas acotadas presentan un problema adicional: cuando muchas observaciones caen en la categoría máxima (efecto techo) o en la mínima (efecto piso), la variabilidad real queda truncada. Una escala de dolor posoperatorio donde el 60% de los animales puntúa 0 (sin dolor) después de la cirugía no va a distribuirse normalmente por más transformaciones que se apliquen. Los métodos para datos ordinales manejan estas acumulaciones de manera natural porque trabajan con frecuencias o rangos, no con magnitudes.

6.3. Regresión ordinal: logit acumulativo y supuesto de odds proporcionales

Las pruebas de rangos cubiertas en capítulos anteriores (Mann-Whitney, Kruskal-Wallis) permiten comparar grupos respecto a una variable ordinal, pero no permiten modelar el efecto simultáneo de múltiples predictores ni estimar la magnitud del efecto de cada uno. Para eso se necesita un modelo de regresión.

La regresión logística ordinaria modela una respuesta dicotómica (sí/no). La regresión ordinal extiende esa lógica a una respuesta con más de dos categorías ordenadas.

El modelo de logit acumulativo (proportional odds model)

Sea *Y* una variable ordinal con *J* categorías (j = 1, 2, …, J). El modelo de logit acumulativo define la probabilidad acumulada de que la respuesta sea igual o menor a la categoría *j*:

$$P(Y \leq j|\mathbf{x}) = \frac{\exp(\alpha_j - \mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\alpha_j - \mathbf{x}'\boldsymbol{\beta})}$$

donde:

- *Y* es la variable respuesta ordinal
- *j* = 1, 2, …, J − 1 (se definen J − 1 ecuaciones, una para cada punto de corte entre categorías consecutivas)
- **x** es el vector de variables predictoras
- α_j son los interceptos (umbrales o puntos de corte), uno para cada acumulación; $\alpha_1 < \alpha_2 < \ldots <$ α_{J−1}
- **β** es el vector de coeficientes de regresión, compartido por todas las ecuaciones

En forma de logit:

$$\log\left(\frac{P(Y \leq j|\mathbf{x})}{P(Y > j|\mathbf{x})}\right) = \alpha_j - \mathbf{x}'\boldsymbol{\beta}$$

La característica distintiva del modelo es que los coeficientes β son los mismos para todas las ecuaciones. Solo cambian los interceptos α_j. Esto significa que el efecto de cada predictor es constante a lo largo de toda la escala ordinal: el predictor desplaza la distribución de probabilidades de manera uniforme a través de las categorías.

Interpretación de los coeficientes

Cada coeficiente β_k se interpreta como el logaritmo del odds ratio acumulativo. Si β_k = 0.85 para un predictor binario (por ejemplo, tipo de alojamiento: estabulación vs. pastoreo), entonces:

$$OR = \exp(0.85) = 2.34$$

Esto significa que la chance de estar en una categoría más alta (peor, si la escala va de leve a grave) es 2.34 veces mayor para el grupo con valor 1 del predictor respecto al grupo con valor 0, y este odds ratio es constante independientemente del punto de corte que se considere.

Ejemplo resuelto 1: Grado de endometritis citológica en vacas lecheras

Un estudio evalúa factores asociados al grado de endometritis citológica en 120 vacas lecheras Holstein de tres fincas del altiplano cundiboyacense en Colombia, examinadas entre 30 y 45 días posparto. La variable respuesta es el grado de endometritis (0, 1, 2 o 3, basado en el porcentaje de neutrófilos en citología endometrial con cytobrush). Los predictores son: número de parto (1, 2, 3+), condición corporal al momento del examen (≤ 2.75 vs. > 2.75), retención de placenta (sí/no) y tipo de parto (eutócico vs. distócico).

Tabla 29
Distribución del grado de endometritis según retención de placenta (n = 120)

Grado de endometritis	Sin retención (n = 89)	Con retención (n = 31)
Grado 0 (< 5% PMN)	42 (47.2%)	5 (16.1%)
Grado 1 (5-15% PMN)	28 (31.5%)	8 (25.8%)

Grado de endometritis	Sin retención (n = 89)	Con retención (n = 31)
Grado 2 (15-50% PMN)	14 (15.7%)	11 (35.5%)
Grado 3 (> 50% PMN)	5 (5.6%)	7 (22.6%)

PMN = polimorfonucleares neutrófilos.

El ajuste del modelo de logit acumulativo produce:

Tabla 30
Coeficientes del modelo de regresión ordinal para grado de endometritis

Predictor	β (EE)	OR	IC 95% OR	p
Retención de placenta (sí vs. no)	1.42 (0.41)	4.14	1.85 – 9.26	< 0.001
Parto distócico (sí vs. no)	0.89 (0.38)	2.44	1.16 – 5.12	0.019
BCS ≤ 2.75 (sí vs. no)	0.72 (0.34)	2.05	1.06 – 3.99	0.034
Número de parto 2 vs. 1	0.31 (0.37)	1.36	0.66 – 2.83	0.402
Número de parto 3+ vs. 1	0.58 (0.39)	1.79	0.83 – 3.85	0.137

EE = error estándar; OR = odds ratio; IC = intervalo de confianza.

La interpretación clínica: una vaca con retención de placenta tiene 4.14 veces la chance de presentar un grado mayor de endometritis (en cualquier punto de corte de la escala) comparada con una vaca sin retención, controlando por los demás factores. La distocia y la baja condición corporal también se asocian con grados más altos de endometritis, aunque con magnitudes menores. El número de parto no muestra asociación estadísticamente significativa en esta muestra.

El supuesto de odds proporcionales

El supuesto central del modelo de logit acumulativo es que los coeficientes β son idénticos para todas las ecuaciones de logit acumulativo. En otras palabras, el efecto del predictor es el mismo sin importar dónde se haga el corte en la escala ordinal. Esto se conoce como supuesto de odds proporcionales (proportional odds assumption) o supuesto de líneas paralelas.

¿Cuándo se viola? Cuando el efecto de un predictor difiere según el nivel de la escala. Supongamos que la retención de placenta incrementa mucho la probabilidad de pasar de grado 0 a grado 1, pero apenas afecta la probabilidad de pasar de grado 2 a grado 3. En ese caso, un solo coeficiente β no captura adecuadamente el efecto, y el supuesto se viola.

Prueba de Brant (Brant, 1990): es la prueba más utilizada para evaluar este supuesto. Compara los coeficientes estimados en J − 1 regresiones logísticas binarias separadas (una para cada punto de corte) con los coeficientes del modelo proporcional. La hipótesis nula es que los coeficientes son iguales a través de los puntos de corte. Un valor $p < 0.05$ sugiere violación del supuesto para ese predictor.

¿Qué hacer si el supuesto se viola?

Tres opciones principales:

1. **Modelo de odds parcialmente proporcionales**: se permite que los coeficientes de los predictores que violan el supuesto varíen entre ecuaciones, mientras que los demás permanecen constantes. Implementable con *ordinal:clm()* usando el argumento *nominal*.
2. **Modelo multinomial**: se abandona el supuesto de orden y se trata la respuesta como nominal. Se pierde eficiencia estadística porque se ignora la información del ordenamiento.
3. **Modelo de razones de continuación (continuation ratio)**: modela la probabilidad de estar en la categoría j dado que se ha alcanzado al menos la categoría j. Puede ser preferible cuando el proceso biológico es secuencial.

Ejemplo resuelto 2: Puntaje de claudicación en vacas según tipo de piso

Se evalúan 90 vacas lecheras de tres establos en Cajamarca, Perú, cada uno con un tipo de piso diferente: concreto sin ranurar, concreto ranurado y piso de goma sobre concreto. Un evaluador entrenado asigna puntajes de claudicación de 1 a 5 (Sprecher et al., 1997).

Tabla 31
Distribución del puntaje de claudicación según tipo de piso (n = 30 por grupo)

Puntaje	Concreto sin ranurar	Concreto ranurado	Piso de goma
1	3 (10.0%)	8 (26.7%)	14 (46.7%)
2	8 (26.7%)	12 (40.0%)	10 (33.3%)
3	11 (36.7%)	7 (23.3%)	4 (13.3%)
4	6 (20.0%)	2 (6.7%)	2 (6.7%)
5	2 (6.7%)	1 (3.3%)	0 (0.0%)

El modelo ordinal con tipo de piso como predictor (referencia: concreto sin ranurar):

- Concreto ranurado vs. sin ranurar: $\beta = -1.08$ ($p = 0.008$), OR = 0.34
- Piso de goma vs. sin ranurar: $\beta = -1.87$ ($p < 0.001$), OR = 0.15

El odds de tener un puntaje de claudicación mayor se reduce un 66% con piso ranurado y un 85% con piso de goma, respecto al concreto sin ranurar. La prueba de Brant global da $p = 0.312$, lo que indica que el supuesto de odds proporcionales no se rechaza con estos datos.

6.4. Acuerdo entre evaluadores: kappa de Cohen, kappa ponderada, kappa de Fleiss

Muchas variables ordinales en veterinaria dependen del juicio de un observador, y eso introduce una fuente de variación que debe cuantificarse: ¿qué tan reproducible es la evaluación entre distintos observadores? ¿Y qué tan consistente es un mismo observador en evaluaciones repetidas?

El porcentaje simple de acuerdo (proporción de casos en que dos evaluadores coinciden) es insuficiente porque no corrige por el acuerdo que ocurriría por azar.

Kappa de Cohen (κ)

Para dos evaluadores que clasifican los mismos *n* sujetos en las mismas categorías, el coeficiente kappa de Cohen (Cohen, 1960) se define como:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

donde:

- p_o es la proporción de acuerdo observado (frecuencia relativa de casos donde ambos evaluadores asignan la misma categoría)
- p_e es la proporción de acuerdo esperado por azar, calculada a partir de las distribuciones marginales de cada evaluador

El rango teórico de κ va de −1 a 1:

- κ = 1: acuerdo perfecto
- κ = 0: acuerdo igual al esperado por azar
- κ < 0: acuerdo menor que el azar (desacuerdo sistemático)

Criterios de interpretación (Landis & Koch, 1977):

Valor de κ	Interpretación
< 0.00	Acuerdo pobre (peor que el azar)
0.00 – 0.20	Acuerdo leve
0.21 – 0.40	Acuerdo aceptable
0.41 – 0.60	Acuerdo moderado
0.61 – 0.80	Acuerdo sustancial
0.81 – 1.00	Acuerdo casi perfecto

Estos umbrales son orientaciones generales, no reglas absolutas. Un κ de 0.55 puede ser excelente para una evaluación inherentemente subjetiva (bienestar animal en fauna silvestre) y preocupantemente bajo para una que debería ser reproducible (conteo de células en una cámara de Neubauer).

Ejemplo resuelto: Clasificación histopatológica de mastocitomas caninos

Dos patólogos clasifican de manera independiente 50 biopsias de mastocitoma cutáneo canino según el sistema de Patnaik (grados I, II, III). Los resultados se presentan en una tabla de contingencia:

Tabla 32
Clasificación de 50 microcitomas caninos por dos patólogos

	Patólogo B: Grado I	Patólogo B: Grado II	Patólogo B: Grado III	Total
Patólogo A: Grado I	10	3	0	13
Patólogo A: Grado II	2	18	3	23
Patólogo A: Grado III	0	2	12	14
Total	12	23	15	50

Cálculo paso a paso:

Acuerdo observado: $p_o = (10 + 18 + 12) / 50 = 40/50 = 0.80$

Acuerdo esperado por azar:

- P(ambos Grado I) = (13/50) × (12/50) = 0.0624
- P(ambos Grado II) = (23/50) × (23/50) = 0.2116
- P(ambos Grado III) = (14/50) × (15/50) = 0.0840
- $p_e = 0.0624 + 0.2116 + 0.0840 = 0.3580$

$\kappa = (0.80 - 0.358) / (1 - 0.358) = 0.442 / 0.642 =$ **0.689**

Interpretación: acuerdo sustancial según los criterios de Landis y Koch. Los patólogos coinciden en el grado del 80% de los tumores, y la concordancia corregida por azar es de 0.69. Para oncología veterinaria, este nivel de acuerdo es razonable, aunque no óptimo; la mayor discrepancia ocurre entre los grados I y II, lo cual es consistente con lo reportado en la literatura sobre la subjetividad de esta frontera diagnóstica (Kiupel et al., 2011).

Kappa ponderada (κ_w)

La kappa de Cohen trata todos los desacuerdos por igual: confundir grado I con grado II se penaliza igual que confundir grado I con grado III. Eso no tiene sentido para variables ordinales, donde un desacuerdo de una categoría es menos grave que un desacuerdo de dos.

El kappa ponderada (Cohen, 1968) incorpora pesos que reflejan la gravedad del desacuerdo:

$$\kappa_w = \frac{\sum_i \sum_j w_{ij} \cdot p_{o_{ij}} - \sum_i \sum_j w_{ij} \cdot p_{e_{ij}}}{1 - \sum_i \sum_j w_{ij} \cdot p_{e_{ij}}}$$

Los dos esquemas de pesos más comunes son:

Pesos lineales: w_ij = 1 − |i − j| / (J − 1). El desacuerdo se penaliza proporcionalmente a la distancia entre categorías.

Pesos cuadráticos: w_ij = 1 − (i − j)² / (J − 1)². Los desacuerdos grandes se penalizan desproporcionadamente más que los pequeños. La kappa con pesos cuadráticos es numéricamente equivalente al coeficiente de correlación intraclase (ICC) para acuerdo, lo cual lo conecta con una medida familiar.

Para los datos de la Tabla 6.4, con pesos cuadráticos:

- κ_w (cuadrático) = **0.79**

El valor es mayor que la kappa sin ponderar (0.69) porque los desacuerdos existentes son predominantemente de una sola categoría de diferencia, y El kappa ponderada les asigna una penalización menor.

Kappa de Fleiss

Cuando hay más de dos evaluadores, la kappa de Cohen no es directamente aplicable. La kappa de Fleiss (Fleiss, 1971) extiende el concepto a *m* evaluadores que clasifican *n* sujetos en *k* categorías:

$$\kappa_{Fleiss} = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

donde $\bar{P}$ es la proporción media de pares de evaluadores que concuerdan y $\bar{P}_e$ es la proporción esperada por azar.

Un detalle que se confunde con frecuencia: la kappa de Fleiss asume que los evaluadores no son siempre los mismos para cada sujeto (diseño con evaluadores intercambiables). Si los mismos tres

patólogos evalúan todas las biopsias, técnicamente deberían calcularse kappas de Cohen por pares y luego promediarlos, o usar el coeficiente de concordancia de Conger. En la práctica, la kappa de Fleiss se usa en ambas situaciones, y la diferencia es generalmente pequeña, pero conviene conocerla.

Cuando y por qué medir el acuerdo

El acuerdo entre evaluadores no es un accesorio metodológico: es un componente central de la validez de cualquier estudio que dependa de evaluaciones subjetivas. Si dos patólogos no coinciden en la clasificación de un tumor, los resultados del estudio que depende de esa clasificación están contaminados por variabilidad de medición no reconocida.

En la práctica veterinaria latinoamericana, esto es particularmente relevante porque muchos estudios de campo utilizan evaluaciones visuales realizadas por múltiples personas (estudiantes, técnicos, veterinarios de planta) sin verificar previamente la concordancia entre ellas. Reportar la kappa interobservador debería ser obligatorio en cualquier estudio que emplee escalas de evaluación clínica.

6.5. Tamaño del efecto: qué es, por qué reportarlo y medidas para cada prueba no paramétrica

El problema con el valor p como única medida de resultado

Un estudio compara la ganancia de peso de terneros suplementados con dos aditivos. Con 15 animales por grupo, Mann-Whitney da $p = 0.23$. Con 150 por grupo, la misma diferencia real en medianas produce $p = 0.003$. ¿Cambió la biología? No. Cambió el tamaño muestral, y con él la capacidad de detectar una diferencia. El valor p mezcla dos cosas: la magnitud de la diferencia y la cantidad de información disponible (tamaño de muestra). El tamaño del efecto separa lo primero de lo segundo.

Las guías de la American Psychological Association (APA, 2020), las directrices CONSORT para ensayos clínicos y un número creciente de revistas veterinarias (Journal of Veterinary Internal Medicine, Preventive Veterinary Medicine) exigen o recomiendan el reporte del tamaño del efecto junto con el valor p. No hacerlo es omitir la mitad de la historia.

Medidas de tamaño del efecto para pruebas no paramétricasas

A continuación, se presenta una tabla consolidada con las medidas de tamaño del efecto apropiadas para cada prueba no paramétrica tratada en este libro.

Tabla 33
Medidas de tamaño del efecto para pruebas no paramétricass

Prueba	Medida de tamaño del efecto	Rango	Interpretación (Cohen)
Mann-Whitney U	r de rango biserial	−1 a 1	0.1 pequeño, 0.3 mediano, 0.5 grande
Mann-Whitney U	A de Vargha y Delaney	0 a 1	0.56 pequeño, 0.64 mediano, 0.71 grande
Wilcoxon rangos con signo	r ($Z / \sqrt{n}$)	−1 a 1	0.1 pequeño, 0.3 mediano, 0.5 grande
Kruskal-Wallis	ε^2 (épsilon cuadrado)	0 a 1	0.01 pequeño, 0.06 mediano, 0.14 grande
Kruskal-Wallis	η^2_H (eta cuadrado basado en H)	0 a 1	Similar a η^2 paramétrico
Friedman	W de Kendall	0 a 1	0.1 pequeño, 0.3 mediano, 0.5 grande
Chi-cuadrado	V de Cramér	0 a 1	Depende de gl (ver texto)
Chi-cuadrado (2×2)	Phi (φ)	−1 a 1	0.1 pequeño, 0.3 mediano, 0.5 grande
Chi-cuadrado (2×2)	Odds ratio	0 a ∞	1.5 pequeño, 2.5 mediano, 4.3 grande
Spearman / Kendall	El coeficiente mismo (r_s, τ)	−1 a 1	0.1 pequeño, 0.3 mediano, 0.5 grande
PERMANOVA	R^2 (proporción de varianza)	0 a 1	Depende del contexto

Los umbrales de Cohen (1988) son convenciones útiles como punto de partida, pero no reemplazan el juicio clínico. En veterinaria, un efecto “pequeño” en términos de Cohen puede ser clínicamente significativo si se refiere a la mortalidad neonatal en un criadero con miles de lechones, y un efecto “grande” puede ser irrelevante si el desenlace no tiene impacto práctico en el manejo del animal.

r de rango biserial y r basado en Z

Para Mann-Whitney y Wilcoxon, la medida más directa es *r*, calculada como:

$$r = \frac{Z}{\sqrt{n}}$$

donde *Z* es el estadístico estandarizado de la prueba y *n* es el número total de observaciones (para Mann-Whitney) o el número de pares (para Wilcoxon pareado). Esta fórmula tiene la virtud de la simplicidad: se obtiene directamente de la salida de R.

La *r* de rango biserial (rank-biserial correlation, Wendt, 1972) proporciona una interpretación más intuitiva para Mann-Whitney:

$$r_{rb} = \frac{2U}{n_1 \cdot n_2} - 1$$

donde *U* es el estadístico de Mann-Whitney y n_1, n_2 son los tamaños de los dos grupos. El resultado va de −1 a 1 y se interpreta como la diferencia entre la proporción de pares donde el grupo 1 supera al grupo 2 y la proporción donde ocurre lo contrario.

A de Vargha y Delaney

La *A* de Vargha y Delaney (2000) es la probabilidad de que una observación aleatoria del grupo 1 sea mayor que una observación aleatoria del grupo 2:

$$A = \frac{U}{n_1 \cdot n_2}$$

Si A = 0.50, los grupos son idénticos en términos estocásticos. A = 0.71 significa que en el 71% de los pares posibles, una observación del grupo 1 supera a una del grupo 2. Es probablemente la medida más fácil de comunicar a un clínico o un productor.

Épsilon cuadrada (ε²) para Kruskal-Wallis

$$\varepsilon^2 = \frac{H}{(n^2 - 1)/(n + 1)}$$

donde *H* es el estadístico de Kruskal-Wallis y *n* es el tamaño total de la muestra. Varía entre 0 y 1, con interpretación análoga a η^2 en ANOVA.

Ejemplo numérico consolidado

Un investigador compara el puntaje de dolor posoperatorio (escala de 0 a 4) entre dos protocolos analgésicos en 24 perros (12 por grupo) sometidos a cirugía ortopédica. Mann-Whitney produce $U = 32$, $Z = -2.31$, $p = 0.021$.

Cálculo del tamaño del efecto:

- $r = |-2.31| / \sqrt{24} = 2.31 / 4.899 =$ **0.47** → efecto grande según Cohen
- $r\_rb = (2 \times 32) / (12 \times 12) - 1 = 64/144 - 1 = -0.556$ → efecto grande, dirección favorable al grupo 1
- $A = 32 / (12 \times 12) = 32/144 =$ **0.22** → en solo el 22% de los pares un perro del protocolo 1 tiene más dolor que uno del protocolo 2

El valor p (0.021) dice que la diferencia es estadísticamente significativa. El tamaño del efecto dice que la diferencia es grande: el protocolo 2 produce puntajes de dolor consistentemente más altos. Esas son dos piezas de información distintas y complementarias.

6.6. Potencia estadística en métodos no paramétricos y eficiencia relativa asintótica

¿Qué es la potencia y por qué preocuparse?

La potencia estadística es la probabilidad de rechazar la hipótesis nula cuando esta es falsa. En otras palabras: la probabilidad de detectar un efecto que realmente existe. Depende de cuatro factores:

4. **Tamaño del efecto**: diferencias grandes se detectan más fácilmente.
5. **Tamaño muestral**: más observaciones proporcionan más información.
6. **Nivel de significancia (α)**: un α más permisivo (0.10 vs. 0.05) aumenta la potencia a costa de más errores tipo I.
7. **Variabilidad de los datos**: mayor variabilidad dificulta la detección.

El estándar aceptado en investigación biomédica es una potencia mínima de 0.80 (80%), lo que implica que se acepta un 20% de probabilidad de no detectar un efecto real (error tipo II, $\beta = 0.20$).

El costo de ser no paramétrico: eficiencia relativa asintótica (ARE)

Una pregunta recurrente: ¿cuánta potencia se pierde al usar una prueba no paramétrica en lugar de su equivalente paramétrica?

La eficiencia relativa asintótica (Asymptotic Relative Efficiency, ARE) compara la cantidad de observaciones que cada prueba necesita para alcanzar la misma potencia, cuando el tamaño muestral tiende a infinito. Se define como:

$$ARE = \frac{n_{paramétrica}}{n_{no\ paramétrica}}$$

Es decir, el ARE indica cuántas observaciones necesita la prueba paramétrica en relación con la no paramétrica para lograr la misma potencia. Un ARE < 1 significa que la prueba no paramétrica necesita más observaciones.

Tabla 34
Eficiencia relativa asintótica de pruebas no paramétricas respecto a sus equivalentes paramétricos bajo normalidad

Prueba no paramétrica	Equivalente paramétrico	ARE bajo normalidad
Mann-Whitney U	t de Student (dos muestras)	$3/\pi \approx 0.955$
Wilcoxon rangos con signo	t pareada	$3/\pi \approx 0.955$
Kruskal-Wallis	ANOVA de un factor	$3/\pi \approx 0.955$
Prueba de signos	t pareada	$2/\pi \approx 0.637$
Spearman	Pearson	$9/\pi^2 \approx 0.912$

Las cifras bajo normalidad merecen atención. Mann-Whitney tiene un ARE de 0.955 respecto a la prueba t cuando los datos son perfectamente normales. Eso significa que, con datos normales, Mann-Whitney necesita aproximadamente un 5% más de observaciones para alcanzar la misma potencia que la t. Con muestras de 20 por grupo, eso equivale a necesitar 21 en lugar de 20: una diferencia irrelevante en la práctica.

Pero cuando la distribución se aleja de la normalidad distribuciones con colas pesadas, asimétricas, o contaminadas con valores atípicos la situación se invierte. Mann-Whitney puede tener potencia sustancialmente mayor que la t. Para distribuciones de colas pesadas como la doble exponencial, el ARE de Mann-Whitney respecto a la t puede superar 1.0, lo que significa que la prueba no paramétrica es más eficiente.

La prueba de signos es la excepción: con un ARE de 0.637 bajo normalidad, necesita aproximadamente un 57% más de observaciones. Esa pérdida de potencia es el precio de ignorar las magnitudes de las diferencias, y por eso la prueba de signos solo se justifica cuando la información de magnitud no es confiable o la escala es tan gruesa que los rangos pierden significado.

Implicaciones prácticas para el diseño experimental veterinario

En un ensayo con bovinos donde cada unidad experimental cuesta entre 500 y 2 000 dólares en alimentación, manejo y análisis de laboratorio, un aumento del 5% en el tamaño muestral para usar Mann-Whitney en lugar de la prueba t es trivial en comparación con el riesgo de aplicar una prueba paramétrica a datos que no cumplen sus supuestos. El argumento "pierdo potencia si uso la prueba no paramétrica" rara vez justifica forzar un análisis paramétrico, porque la pérdida real bajo normalidad es mínima y la ganancia bajo no normalidad puede ser sustancial.

6.7. Determinación del tamaño muestral para estudios con análisis no paramétrico

El problema práctico

Un tesista de maestría planifica un estudio para comparar la eficacia de dos antihelmínticos en caprinos criollos del altiplano boliviano, midiendo la reducción en el conteo de huevos por gramo de heces (HPG). Los conteos de HPG son notoriamente asimétricos, con distribuciones que van desde 0 hasta valores extremos de 10 000 o más. La prueba estadística será Mann-Whitney. ¿Cuántos animales necesita por grupo?

Enfoque basado en la prueba paramétrica equivalente con factor de corrección

El método más simple consiste en calcular el tamaño muestral como si se fuera a usar la prueba paramétrica equivalente y luego multiplicar por el inverso del ARE:

$$n_{no\ paramétrica} = \frac{n_{paramétrica}}{ARE}$$

Para Mann-Whitney bajo normalidad, ARE ≈ 0.955, así que:

$$n_{Mann-Whitney} = \frac{n_t}{0.955} \approx 1.047 \cdot n_t$$

Esto subestima las necesidades reales cuando la distribución se aleja mucho de la normalidad, pero proporciona un punto de partida.

Enfoque de Noether (1987) para Mann-Whitney

Noether (1987) derivó una fórmula específica para el tamaño muestral de Mann-Whitney basada en la probabilidad de que una observación del grupo 1 supere a una del grupo 2 (p_{12}):

$$n_{por\ grupo} = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2}{3 \cdot (p_{12} - 0.5)^2}$$

donde:

- z_α/2 = valor de la distribución normal estándar para el nivel de significancia bilateral (1.96 para α = 0.05)
- z_β = valor para la potencia deseada (0.842 para potencia de 0.80)
- $p_{12} = P(X_1 > X_2)$, la probabilidad de superioridad

Ejemplo numérico:

Si se espera que un caprino tratado con el antihelmíntico A tenga un conteo de HPG menor que uno tratado con el antihelmíntico B en el 70% de los pares posibles ($p_{12} = 0.70$, equivalente a de Vargha y Delaney = 0.70):

$$n = \frac{(1.96 + 0.842)^2}{3 \times (0.70 - 0.50)^2} = \frac{(2.802)^2}{3 \times 0.04} = \frac{7.851}{0.12} = 65.4$$

Se necesitarían 66 animales por grupo. Si la diferencia esperada es más sutil ($p_{12} = 0.60$):

$$n = \frac{7.851}{3 \times 0.01} = \frac{7.851}{0.03} = 261.7$$

Se necesitarían 262 por grupo. La relación entre la magnitud del efecto esperado y el tamaño muestral requerido es no lineal: detectar efectos pequeños cuesta exponencialmente más animales.

Enfoque por simulación

Cuando la distribución esperada no es normal ni se ajusta a un modelo conocido, la forma más confiable de estimar el tamaño muestral es la simulación. El procedimiento general es:

8. Especificar las distribuciones esperadas bajo H_1 para cada grupo (basadas en datos piloto, literatura o juicio experto).
9. Para un tamaño muestral n candidato, simular muchos conjuntos de datos (≥ 1 000) bajo H_1.
10. Aplicar la prueba no paramétrica a cada conjunto simulado.
11. Calcular la proporción de simulaciones donde se rechaza H_0 (potencia estimada).
12. Repetir para distintos valores de n hasta encontrar el que produce potencia ≥ 0.80.

Este enfoque se desarrolla con código completo en la sección 6.8.

6.8. Estimación de potencia por simulación

Lógica del procedimiento

La simulación de potencia es, en esencia, un experimento virtual: se repite el estudio hipotético miles de veces bajo condiciones controladas y se cuenta cuántas veces la prueba estadística detecta el efecto que se sabe que existe (porque lo generamos nosotros).

Ejemplo detallado: Potencia de Mann-Whitney para comparar conteos de HPG

Contexto: Un investigador quiere comparar dos antihelmínticos en caprinos. Los datos de HPG siguen una distribución log-normal con los siguientes parámetros estimados a partir de un estudio piloto:

- **Grupo control**: media log = 6.5, DE log = 1.2 (mediana ≈ 665 HPG)
- **Grupo tratado**: media log = 5.5, DE log = 1.2 (mediana ≈ 245 HPG)

La pregunta: ¿qué tamaño muestral se necesita para detectar esta diferencia con una potencia de 0.80 usando Mann-Whitney a un nivel de significancia de 0.05?

Procedimiento paso a paso:

13. Definir un rango de tamaños muestrales candidatos: n = 8, 10, 12, 15, 20, 25, 30 por grupo.
14. Para cada n, repetir 10 000 veces:
    a. Generar n observaciones del Grupo control: rlnorm(n, meanlog = 6.5, sdlog = 1.2)
    b. Generar n observaciones del Grupo tratado: rlnorm(n, meanlog = 5.5, sdlog = 1.2)
    c. Aplicar wilcox.test() y registrar si $p < 0.05$.
15. La potencia estimada para cada n es la proporción de las 10 000 simulaciones donde se rechazó H_0.

Resultados de la simulación (valores aproximados):

Tabla 35
Potencia estimada de Mann-Whitney para comparar conteos de HPG según tamaño muestral

n por grupo	Potencia estimada	IC 95% de la potencia
8	0.43	0.42 – 0.44
10	0.53	0.52 – 0.54
12	0.62	0.61 – 0.63
15	0.73	0.72 – 0.74
20	0.84	0.83 – 0.85
25	0.91	0.90 – 0.92
30	0.95	0.95 – 0.96

Con 20 animales por grupo, la potencia alcanza 0.84, superando el umbral convencional de 0.80. Con 15 por grupo, la potencia es de 0.73: insuficiente. Un tesista que disponga de 12 caprinos por grupo tiene solo un 62% de probabilidad de detectar una diferencia que reduce la mediana de HPG de 665 a 245. Si no detecta la diferencia, no podrá saber si el antihelmíntico no funciona o si su estudio era demasiado pequeño.

Simulación de potencia para Kruskal-Wallis

El mismo principio se extiende a diseños con más de dos grupos. La diferencia es que se generan k muestras bajo H_1 (con distribuciones que difieren entre grupos) y se aplica kruskal.test().

Simulación para evaluar el error tipo I

Como verificación complementaria, se recomienda simular bajo H_0 (todos los grupos con la misma distribución) para confirmar que la proporción de rechazos es cercana a $\alpha = 0.05$. Si el error tipo I

empírico se desvía mucho del nominal, hay un problema con la prueba o con el procedimiento de simulación que debe resolverse antes de interpretar los resultados de potencia.

6.9. Implementación en R

```
# ============================================================
# CAPÍTULO 6: IMPLEMENTACIÓN EN R
# Datos ordinales, escalas de evaluación clínica y tamaño del efecto
# ============================================================

# --- Paquetes requeridos ---
# Instalar si no están disponibles:
# install.packages(c("MASS", "ordinal", "brant", "irr", "rcompanion",
#                    "effsize", "pwr", "DescTools", "ggplot2", "dplyr"))

library(MASS)        # polr() para regresión ordinal
library(ordinal)     # clm() y clmm() para modelos ordinales más flexibles
library(brant)       # prueba de Brant para odds proporcionales
library(irr)         # kappa2(), kappam.fleiss() para acuerdo entre evaluadores
library(rcompanion)  # wilcoxonR(), vda(), epsilonSquared()
library(effsize)     # VD.A() para A de Vargha y Delaney
library(pwr)         # pwr.t.test() como punto de partida para cálculos de potencia
library(DescTools)   # CramerV(), CohenKappa()
library(ggplot2)     # gráficos
library(dplyr)       # manipulación de datos

# ============================================================
# 6.3. REGRESIÓN ORDINAL: MODELO DE LOGIT ACUMULATIVO
# ============================================================

# --- Ejemplo 1: Grado de endometritis en vacas lecheras ---

# Crear el conjunto de datos
set.seed(2835)
n <- 120

datos_endo <- data.frame(
  id = 1:n,
  retencion = factor(sample(c("No", "Si"), n, replace = TRUE,
                            prob = c(0.74, 0.26))),
  parto_tipo = factor(sample(c("Eutocico", "Distocico"), n, replace = TRUE,
                             prob = c(0.82, 0.18))),
```

```
  bcs_baja = factor(sample(c("No", "Si"), n, replace = TRUE,
                           prob = c(0.60, 0.40))),
  num_parto = factor(sample(c("1", "2", "3+"), n, replace = TRUE,
                            prob = c(0.30, 0.35, 0.35)))
)

# Generar la variable respuesta ordinal con probabilidades
# que dependan de los predictores
# Probabilidad base de cada grado: 0=40%, 1=30%, 2=20%, 3=10%
prob_base <- c(0.40, 0.30, 0.20, 0.10)

grado <- numeric(n)
for (i in 1:n) {
  logit_shift <- 0
  if (datos_endo$retencion[i] == "Si") logit_shift <- logit_shift + 1.4
  if (datos_endo$parto_tipo[i] == "Distocico") logit_shift <- logit_shift + 0.9
  if (datos_endo$bcs_baja[i] == "Si") logit_shift <- logit_shift + 0.7

  # Ajustar probabilidades acumulativas
  cum_probs <- plogis(qlogis(cumsum(prob_base[1:3])) + logit_shift)
  probs <- c(cum_probs[1],
             cum_probs[2] - cum_probs[1],
             cum_probs[3] - cum_probs[2],
             1 - cum_probs[3])
  probs <- pmax(probs, 0.01)  # evitar probabilidades negativas
  probs <- probs / sum(probs) # renormalizar
  grado[i] <- sample(0:3, 1, prob = probs)
}

datos_endo$grado <- ordered(grado, levels = 0:3,
                            labels = c("Grado 0", "Grado 1",
                                       "Grado 2", "Grado 3"))

# Verificar distribución
table(datos_endo$grado)
table(datos_endo$grado, datos_endo$retencion)

# --- Ajustar el modelo de regresión ordinal con MASS::polr() ---
# polr() utiliza la parametrización: logit(P(Y <= j)) = alpha_j - X'beta
# Nota: el signo de los coeficientes es inverso a lo esperado intuitivamente.
# Un coeficiente positivo en polr() indica que valores altos del predictor
# se asocian con categorías MÁS BAJAS de la respuesta.
# Para obtener la dirección convencional, se pueden negar los coeficientes
# o usar ordinal::clm().

modelo_polr <- polr(grado ~ retencion + parto_tipo + bcs_baja + num_parto,
                    data = datos_endo,
```

```
                Hess = TRUE)  # Hess = TRUE para obtener errores estándar

summary(modelo_polr)

# Tabla de coeficientes con odds ratios e intervalos de confianza
coefs <- coef(summary(modelo_polr))
print(coefs)

# Odds ratios (exponenciar coeficientes; atención al signo en polr)
# En polr, coeficiente positivo = menor probabilidad de categorías altas
# Para la interpretación habitual (OR > 1 = mayor probabilidad de categoría alta):
OR <- exp(-coef(modelo_polr))  # Negar el signo para interpretación convencional
IC_OR <- exp(-confint(modelo_polr))
# Nota: al negar, los límites del IC se invierten
IC_OR <- IC_OR[, c(2, 1)]  # Reordenar columnas
colnames(IC_OR) <- c("IC 2.5%", "IC 97.5%")

cat("\nOdds Ratios e Intervalos de Confianza:\n")
print(cbind(OR = round(OR, 2), round(IC_OR, 2)))

# --- Alternativa con ordinal::clm() ---
# clm() usa la parametrización convencional donde coeficiente positivo
# indica mayor probabilidad de categorías altas
modelo_clm <- clm(grado ~ retencion + parto_tipo + bcs_baja + num_parto,
                  data = datos_endo,
                  link = "logit")

summary(modelo_clm)

# Odds ratios directos (sin negar)
OR_clm <- exp(coef(modelo_clm)[-c(1:3)])  # Excluir umbrales
IC_clm <- exp(confint(modelo_clm)[-c(1:3), ])

cat("\nOdds Ratios (clm):\n")
print(cbind(OR = round(OR_clm, 2), round(IC_clm, 2)))

# --- Prueba de Brant para odds proporcionales ---
# Requiere el modelo ajustado con polr()
brant_test <- brant(modelo_polr)
print(brant_test)

# Interpretación:
# Si p global > 0.05: no se rechaza el supuesto de odds proporcionales
# Si p para un predictor específico < 0.05: ese predictor viola el supuesto

# --- Ejemplo 2: Claudicación según tipo de piso ---
```

```r
datos_claud <- data.frame(
  piso = factor(rep(c("Concreto sin ranurar", "Concreto ranurado",
                      "Piso de goma"), each = 30)),
  claudicacion = ordered(c(
    # Concreto sin ranurar
    rep(1, 3), rep(2, 8), rep(3, 11), rep(4, 6), rep(5, 2),
    # Concreto ranurado
    rep(1, 8), rep(2, 12), rep(3, 7), rep(4, 2), rep(5, 1),
    # Piso de goma
    rep(1, 14), rep(2, 10), rep(3, 4), rep(4, 2), rep(5, 0)
  ))
)

# Recodificar el factor de piso con referencia explícita
datos_claud$piso <- relevel(datos_claud$piso, ref = "Concreto sin ranurar")

# Modelo ordinal
modelo_claud <- clm(claudicacion ~ piso, data = datos_claud, link = "logit")
summary(modelo_claud)

# Odds ratios
cat("\nOdds Ratios para tipo de piso:\n")
exp(coef(modelo_claud)[-c(1:4)])  # Excluir los 4 umbrales (5 categorías - 1)

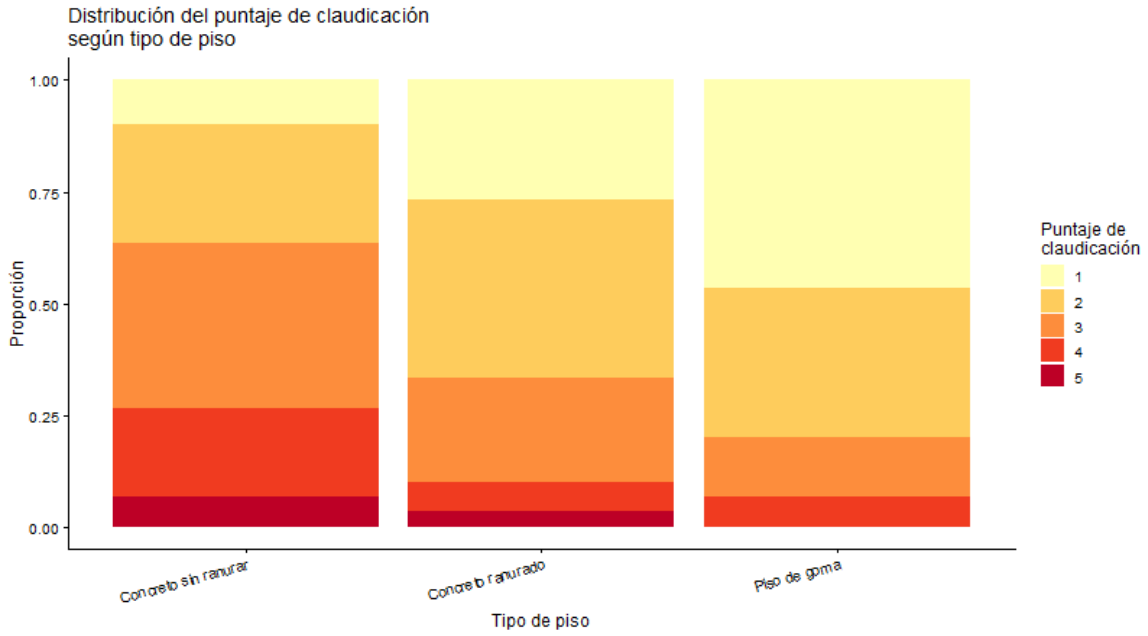
# Prueba de Brant
modelo_claud_polr <- polr(claudicacion ~ piso, data = datos_claud, Hess = TRUE)
brant(modelo_claud_polr)

# --- Visualización de datos ordinales ---

# Gráfico de barras apiladas
ggplot(datos_claud, aes(x = piso, fill = claudicacion)) +
  geom_bar(position = "fill") +
  scale_fill_brewer(palette = "YlOrRd", name = "Puntaje de\nclaudicación") +
  labs(x = "Tipo de piso",
       y = "Proporción",
       title = "Distribución del puntaje de claudicación según tipo de piso") +
  theme_classic(base_size = 12) +
  theme(axis.text.x = element_text(angle = 15, hjust = 1))
```

Figura 24

Distribución del puntaje de claudicación\nsegún tipo de piso



Nota. Gráfico de barras apiladas divergentes (más informativo para ordinales) Requiere calcular proporciones y centrar en la categoría media. Aquí mostramos una versión simplificada con barras apiladas estándar que es la forma más reproducible y clara

```
# ============================================================
# 6.4. ACUERDO ENTRE EVALUADORES
# ============================================================

# --- Kappa de Cohen para dos evaluadores ---

# Datos del ejemplo: clasificación de mastocitomas por dos patólogos
patologoA <- factor(c(rep("Grado I", 13), rep("Grado II", 23),
                      rep("Grado III", 14)),
                    levels = c("Grado I", "Grado II", "Grado III"))
patologoB <- factor(c(
  # De los 13 que A clasificó como Grado I:
  rep("Grado I", 10), rep("Grado II", 3),
  # De los 23 que A clasificó como Grado II:
```

```
  rep("Grado I", 2), rep("Grado II", 18), rep("Grado III", 3),
  # De los 14 que A clasificó como Grado III:
  rep("Grado II", 2), rep("Grado III", 12)
), levels = c("Grado I", "Grado II", "Grado III"))

# Tabla de contingencia
tabla_acuerdo <- table(patologoA, patologoB)
print(tabla_acuerdo)

# Kappa de Cohen (sin ponderar)
kappa_cohen <- kappa2(cbind(patologoA, patologoB), weight = "unweighted")
print(kappa_cohen)
# Resultado esperado: kappa ≈ 0.69

# Kappa ponderado con pesos cuadráticos
kappa_cuadratico <- kappa2(cbind(patologoA, patologoB), weight = "squared")
print(kappa_cuadratico)
# Resultado esperado: kappa_w ≈ 0.79

# Kappa ponderado con pesos lineales
kappa_lineal <- kappa2(cbind(patologoA, patologoB), weight = "equal")
print(kappa_lineal)

# Alternativa con DescTools::CohenKappa()
# Ofrece intervalos de confianza
CohenKappa(tabla_acuerdo, weights = "Quadratic", conf.level = 0.95)

# --- Kappa de Fleiss para tres o más evaluadores ---

# Ejemplo: tres patólogos clasifican 30 biopsias
set.seed(456)
```

```r
# Crear matriz de clasificaciones (filas = sujetos, columnas = evaluadores)
evaluador1 <- sample(1:3, 30, replace = TRUE, prob = c(0.3, 0.4, 0.3))
evaluador2 <- evaluador1  # Alto acuerdo base
# Introducir desacuerdos controlados
desacuerdo_idx <- sample(1:30, 8)
evaluador2[desacuerdo_idx] <- sample(1:3, 8, replace = TRUE)
evaluador3 <- evaluador1
desacuerdo_idx3 <- sample(1:30, 10)
evaluador3[desacuerdo_idx3] <- sample(1:3, 10, replace = TRUE)

datos_fleiss <- cbind(evaluador1, evaluador2, evaluador3)
kappam.fleiss(datos_fleiss)

# ============================================================
# 6.5. TAMAÑO DEL EFECTO PARA PRUEBAS NO PARAMÉTRICAS
# ============================================================

# --- r para Mann-Whitney ---

# Ejemplo: puntaje de dolor en dos grupos de perros
set.seed(789)
grupo1 <- sample(0:4, 12, replace = TRUE, prob = c(0.35, 0.30, 0.20, 0.10, 0.05))
grupo2 <- sample(0:4, 12, replace = TRUE, prob = c(0.10, 0.15, 0.30, 0.25, 0.20))

# Prueba de Mann-Whitney
mw <- wilcox.test(grupo1, grupo2, exact = FALSE)
print(mw)

# r basado en Z
# Extraer Z de la prueba con coin para obtener el estadístico Z directamente
library(coin)
```

```r
datos_dolor <- data.frame(
  dolor = c(grupo1, grupo2),
  grupo = factor(rep(c("Protocolo A", "Protocolo B"), each = 12))
)
mw_coin <- wilcox_test(dolor ~ grupo, data = datos_dolor,
                       distribution = "exact")
Z <- statistic(mw_coin, type = "standardized")
r_efecto <- abs(as.numeric(Z)) / sqrt(24)
cat("\nr (tamaño del efecto):", round(r_efecto, 3), "\n")
cat("Interpretación: ",
    ifelse(abs(r_efecto) < 0.1, "trivial",
    ifelse(abs(r_efecto) < 0.3, "pequeño",
    ifelse(abs(r_efecto) < 0.5, "mediano", "grande"))), "\n")

# r con rcompanion
# wilcoxonR() calcula r = Z / sqrt(n) directamente
r_rcompanion <- wilcoxonR(datos_dolor$dolor, datos_dolor$grupo)
cat("r (rcompanion):", round(r_rcompanion, 3), "\n")

# A de Vargha y Delaney con effsize
vda <- VD.A(grupo1, grupo2)
print(vda)
# A = 0.50 indica sin efecto; > 0.50 indica que grupo1 tiende a ser mayor

# --- Épsilon cuadrado para Kruskal-Wallis ---

set.seed(101)
tres_grupos <- data.frame(
  respuesta = c(rnorm(15, 5, 2), rnorm(15, 7, 2), rnorm(15, 6, 2)),
  grupo = factor(rep(c("A", "B", "C"), each = 15))
)
```

```r
kw <- kruskal.test(respuesta ~ grupo, data = tres_grupos)
print(kw)

# Épsilon cuadrado con rcompanion
eps2 <- epsilonSquared(tres_grupos$respuesta, tres_grupos$grupo)
cat("\nÉpsilon cuadrado:", round(eps2, 3), "\n")
cat("Interpretación: ",
    ifelse(eps2 < 0.01, "trivial",
    ifelse(eps2 < 0.06, "pequeño",
    ifelse(eps2 < 0.14, "mediano", "grande"))), "\n")

# --- V de Cramér para chi-cuadrado ---
tabla_ejemplo <- matrix(c(42, 5, 28, 8, 14, 11, 5, 7), nrow = 4, byrow = TRUE)
cramer_v <- CramerV(tabla_ejemplo, conf.level = 0.95)
cat("\nV de Cramér:", round(cramer_v, 3), "\n")

# ============================================================
# 6.6 – 6.8. POTENCIA ESTADÍSTICA Y SIMULACIÓN
# ============================================================

# --- Cálculo rápido con pwr (como punto de partida) ---

# Para una prueba t (como referencia), luego corregir con 1/ARE
potencia_t <- pwr.t.test(d = 0.8,          # tamaño del efecto Cohen's d
                         sig.level = 0.05,
                         power = 0.80,
                         type = "two.sample",
                         alternative = "two.sided")
print(potencia_t)
```

```r
# Ajuste para Mann-Whitney: multiplicar n por 1/ARE ≈ 1.047
n_mw <- ceiling(potencia_t$n / 0.955)
cat("\nn por grupo para Mann-Whitney (ajuste por ARE):", n_mw, "\n")

# --- Fórmula de Noether para Mann-Whitney ---

noether_n <- function(p12, alpha = 0.05, poder = 0.80) {
  # p12: P(X1 > X2), probabilidad de superioridad
  z_alpha <- qnorm(1 - alpha / 2)
  z_beta <- qnorm(poder)
  n <- (z_alpha + z_beta)^2 / (3 * (p12 - 0.5)^2)
  return(ceiling(n))
}
# Ejemplos con distintos efectos esperados
cat("\nTamaño muestral por grupo (Noether):\n")
cat("p12 = 0.60 (efecto pequeño):", noether_n(0.60), "\n")
cat("p12 = 0.65 (efecto mediano):", noether_n(0.65), "\n")
cat("p12 = 0.70 (efecto grande):", noether_n(0.70), "\n")
cat("p12 = 0.75 (efecto muy grande):", noether_n(0.75), "\n")

# --- Simulación de potencia para Mann-Whitney ---

# Función genérica para simular potencia de Mann-Whitney
potencia_mann_whitney <- function(n_por_grupo, dist_grupo1, dist_grupo2,
                                  nsim = 10000, alpha = 0.05) {
  # n_por_grupo: tamaño de cada grupo
  # dist_grupo1, dist_grupo2: funciones que generan datos (ej. function(n) rlnorm(n, 6.5, 1.2))
  # nsim: número de simulaciones
  # alpha: nivel de significancia

  rechazos <- replicate(nsim, {
```

```r
    x1 <- dist_grupo1(n_por_grupo)
    x2 <- dist_grupo2(n_por_grupo)
    p_valor <- wilcox.test(x1, x2, exact = FALSE)$p.value
    p_valor < alpha
  })

  potencia <- mean(rechazos)
  ee <- sqrt(potencia * (1 - potencia) / nsim)

  list(potencia = potencia,
       ic_inferior = potencia - 1.96 * ee,
       ic_superior = potencia + 1.96 * ee)
}
# Aplicar a distintos tamaños muestrales
# Escenario: conteo de HPG, distribución log-normal
set.seed(12345)
tamanos <- c(8, 10, 12, 15, 20, 25, 30)

cat("\nSimulación de potencia para Mann-Whitney\n")
cat("Distribución: log-normal\n")
cat("Grupo control: meanlog=6.5, sdlog=1.2 (mediana ≈ 665 HPG)\n")
cat("Grupo tratado: meanlog=5.5, sdlog=1.2 (mediana ≈ 245 HPG)\n")
cat(paste(rep("-", 60), collapse = ""), "\n")
cat(sprintf("%-15s %-15s %-20s\n", "n por grupo", "Potencia", "IC 95%"))
cat(paste(rep("-", 60), collapse = ""), "\n")

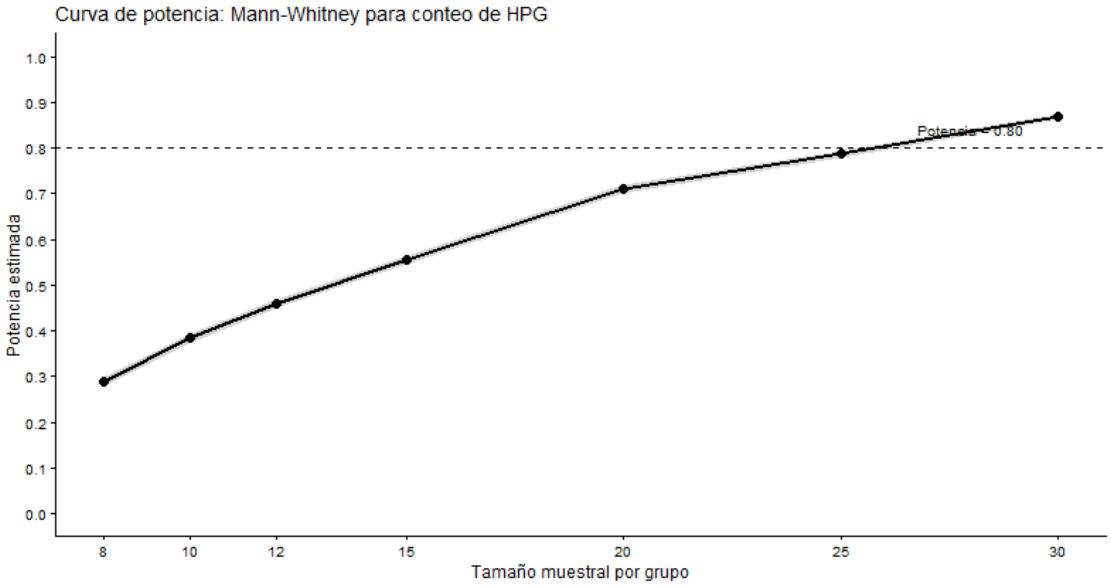
resultados_potencia <- data.frame(
  n = integer(),
  potencia = numeric(),
  ic_inf = numeric(),
  ic_sup = numeric()
```

```
)

for (n_g in tamanos) {
  res <- potencia_mann_whitney(
    n_por_grupo = n_g,
    dist_grupo1 = function(n) rlnorm(n, meanlog = 6.5, sdlog = 1.2),
    dist_grupo2 = function(n) rlnorm(n, meanlog = 5.5, sdlog = 1.2),
    nsim = 10000
  )
  cat(sprintf("%-15d %-15.3f (%.3f – %.3f)\n",
              n_g, res$potencia, res$ic_inferior, res$ic_superior))
  resultados_potencia <- rbind(resultados_potencia,
                               data.frame(n = n_g,
                                          potencia = res$potencia,
                                          ic_inf = res$ic_inferior,
                                          ic_sup = res$ic_superior))
}
# Gráfico de curva de potencia
ggplot(resultados_potencia, aes(x = n, y = potencia)) +
  geom_line(linewidth = 1, color = "#2c7bb6") +
  geom_point(size = 3, color = "#2c7bb6") +
  geom_ribbon(aes(ymin = ic_inf, ymax = ic_sup), alpha = 0.2, fill = "#2c7bb6") +
  geom_hline(yintercept = 0.80, linetype = "dashed", color = "#d7191c") +
  annotate("text", x = max(tamanos) - 2, y = 0.82,
           label = "Potencia = 0.80", color = "#d7191c", size = 3.5) +
  scale_x_continuous(breaks = tamanos) +
  scale_y_continuous(limits = c(0, 1), breaks = seq(0, 1, 0.1)) +
  labs(x = "Tamaño muestral por grupo",
       y = "Potencia estimada",
       title = "Curva de potencia: Mann-Whitney para conteo de HPG") +
  theme_classic(base_size = 12)
```

Figura 25

Curva de potencia: Mann-Whitney para conteo de HPG




```r
# --- Simulación de potencia para Kruskal-Wallis ---
potencia_kruskal_wallis <- function(n_por_grupo, distribuciones,
                                    nsim = 10000, alpha = 0.05) {
  # distribuciones: lista de funciones, una por grupo
  k <- length(distribuciones)
  rechazos <- replicate(nsim, {
    datos <- data.frame(
      y = unlist(lapply(distribuciones, function(f) f(n_por_grupo))),
      grupo = factor(rep(1:k, each = n_por_grupo))
    )
    p_valor <- kruskal.test(y ~ grupo, data = datos)$p.value
    p_valor < alpha
  })
  mean(rechazos)
}

# Ejemplo: tres tratamientos con distribuciones distintas
```

```r
set.seed(54321)
pot_kw <- potencia_kruskal_wallis(
  n_por_grupo = 15,
  distribuciones = list(
    function(n) rlnorm(n, 6.5, 1.2),   # Control
    function(n) rlnorm(n, 5.8, 1.2),   # Tratamiento 1
    function(n) rlnorm(n, 5.2, 1.2)    # Tratamiento 2
  )
)
cat("\nPotencia de Kruskal-Wallis (n=15/grupo, 3 grupos):", round(pot_kw, 3), "\n")

# --- Verificación del error tipo I ---
# Simular bajo H0 (todos los grupos iguales) para verificar que el error
# tipo I es cercano a 0.05

error_tipo_I <- potencia_kruskal_wallis(
  n_por_grupo = 15,
  distribuciones = list(
    function(n) rlnorm(n, 6.5, 1.2),
    function(n) rlnorm(n, 6.5, 1.2),
    function(n) rlnorm(n, 6.5, 1.2)
  )
)
cat("Error tipo I empírico (debe ser ≈ 0.05):", round(error_tipo_I, 3), "\n")

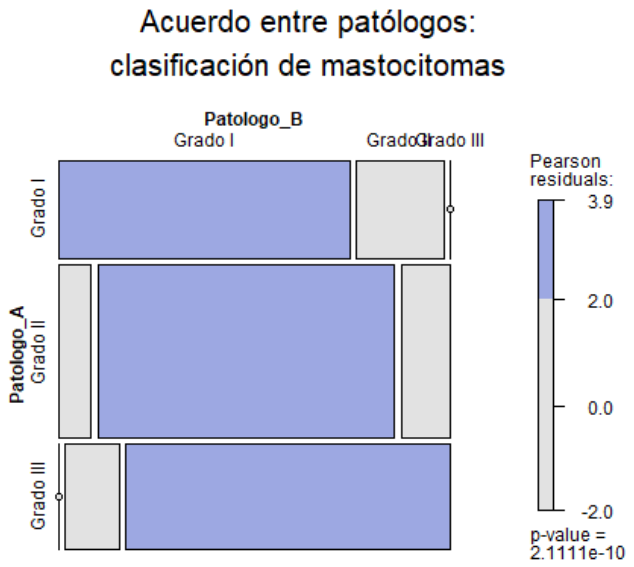
# ============================================================
# GRÁFICOS ADICIONALES PARA DATOS ORDINALES
# ============================================================

# --- Gráfico de mosaico con vcd ---
library(vcd)
```

```
mosaic(tabla_acuerdo, shade = TRUE,
       main = "Acuerdo entre patólogos: clasificación de mastocitomas")
```

Figura 26

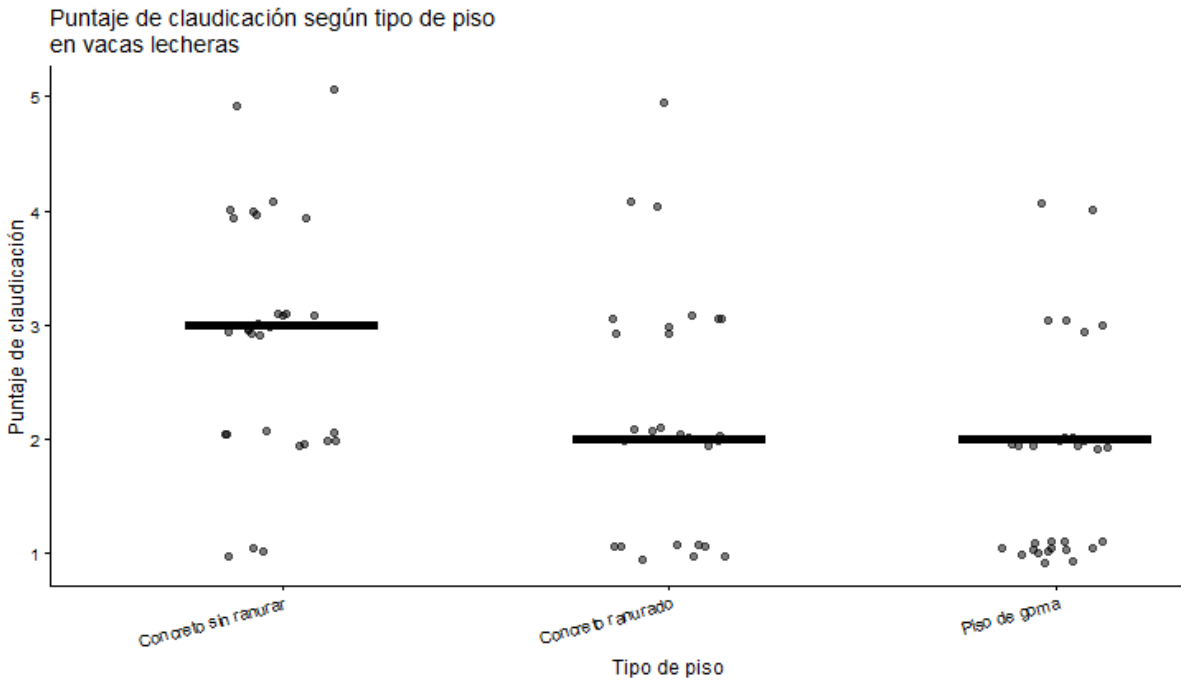
Acuerdo entre patólogos:\nclasificación de mastocitomas




```
# --- Gráfico de puntos para datos ordinales (alternativa al boxplot) ---
ggplot(datos_claud, aes(x = piso, y = as.numeric(claudicacion))) +
  geom_jitter(width = 0.15, height = 0.1, alpha = 0.5, size = 2) +
  stat_summary(fun = median, geom = "crossbar",
               width = 0.5, color = "#d7191c", linewidth = 0.8) +
  labs(x = "Tipo de piso",
       y = "Puntaje de claudicación",
       title = "Puntaje de claudicación según tipo de piso en vacas lecheras") +
  scale_y_continuous(breaks = 1:5) +
  theme_classic(base_size = 12) +
  theme(axis.text.x = element_text(angle = 15, hjust = 1))
```

Figura 27

Puntaje de claudicación según tipo de piso\nen vacas lecheras



Notas sobre las funciones y paquetes

MASS::polr()* vs. *ordinal::clm(): Ambas ajustan modelos de logit acumulativo, pero difieren en la parametrización. *polr()* usa la convención donde un coeficiente positivo indica menor probabilidad de categorías altas (signo invertido respecto a la intuición habitual). *clm()* usa la convención opuesta, que la mayoría de los investigadores encuentra más natural. Además, *clm()* permite modelos de odds parcialmente proporcionales (argumento *nominal*), modelos con efectos aleatorios (*clmm()*), y distintas funciones de enlace (*logit*, *probit*, *cloglog*, *cauchit*). Para uso general, *clm()* es la opción recomendada.

brant::brant(): Solo funciona con objetos ajustados por *polr()*. No acepta objetos de *clm()*. Para verificar odds proporcionales con *clm()*, se puede comparar el modelo proporcional con uno de odds no proporcionales usando *anova()*.

irr::kappa2()* vs. *DescTools::CohenKappa(): *kappa2()* acepta una matriz con dos columnas (una por evaluador) y permite especificar pesos directamente. *CohenKappa()* acepta una tabla de contingencia cuadrada y calcula intervalos de confianza, lo que la hace más útil para el reporte.

rcompanion::wilcoxonR(): Calcula r = Z/√n directamente, sin necesidad de extraer el estadístico Z manualmente. Es la forma más rápida de obtener el tamaño del efecto para Mann-Whitney o Wilcoxon.

effsize::VD.A(): Calcula la A de Vargha y Delaney con intervalos de confianza y magnitud del efecto según los umbrales propuestos por los autores originales.

6.10. Estudio de caso integrador

Parte A: Efecto del tipo de cama sobre lesiones de tarso en vacas lecheras estabuladas

Contexto

En un estudio realizado en tres lecherías intensivas del departamento de Antioquia, Colombia, se evaluó el efecto del tipo de cama sobre la severidad de las lesiones de tarso en vacas Holstein estabuladas. Cada lechería utilizaba un tipo de cama diferente: arena (n = 35 vacas), paja de arroz (n = 32) y colchoneta de goma (n = 38). Dos evaluadores entrenados examinaron independientemente las lesiones de tarso de cada vaca y asignaron un puntaje según una escala ordinal de 4 categorías:

- **Grado 0**: Sin lesión, piel intacta
- **Grado 1**: Alopecia o abrasión superficial (< 2 cm)
- **Grado 2**: Lesión con pérdida de continuidad de piel (2-5 cm), sin inflamación periférica
- **Grado 3**: Lesión ulcerada (> 5 cm) o con inflamación/infección periférica

Datos del Evaluador 1:

Tabla 36
Distribución del grado de lesión de tarso según tipo de cama (Evaluador 1)

Grado de lesión	Arena (n=35)	Paja (n=32)	Colchoneta (n=38)
0	18 (51.4%)	10 (31.3%)	5 (13.2%)
1	10 (28.6%)	11 (34.4%)	12 (31.6%)
2	5 (14.3%)	8 (25.0%)	14 (36.8%)
3	2 (5.7%)	3 (9.4%)	7 (18.4%)

Análisis paso a paso:

1. Evaluación del acuerdo entre los dos evaluadores

Antes de cualquier análisis comparativo, se verifica la concordancia entre evaluadores. Con 105 vacas clasificadas por dos evaluadores en 4 categorías ordinales, se calcula El kappa ponderada con pesos cuadráticos.

```
# Suponer que evaluador1 y evaluador2 son vectores con los puntajes
# kappa ponderado
kappa evaluadores <- kappa2(cbind(evaluador1, evaluador2), weight = "squared")
```

Resultado: κ_w = 0.83 (IC 95%: 0.74 – 0.91). Acuerdo casi perfecto. Se procede a analizar los datos del Evaluador 1 (o un consenso entre ambos).

2. Comparación entre grupos con Kruskal-Wallis

```
kruskal.test(grado ~ tipo_cama, data = datos_tarso)
```

Resultado: H(2) = 16.42, p < 0.001. Al menos un tipo de cama difiere de los demás.

3. Comparaciones post hoc con prueba de Dunn

```
library(FSA)
dunnTest(grado ~ tipo_cama, data = datos_tarso, method = "holm")
```

Resultados:

Comparación	Z	p ajustado (Holm)
Arena – Paja	−1.89	0.059
Arena – Colchoneta	−4.03	< 0.001
Paja – Colchoneta	−2.18	0.029

La arena produce significativamente menos lesiones que la colchoneta de goma (p < 0.001) y la paja muestra una tendencia similar (p = 0.059). La colchoneta de goma se asocia con los peores puntajes de lesión.

4. Tamaño del efecto

Épsilon cuadrada: $\varepsilon^2 = 0.157 \rightarrow$ efecto grande.

r de rango biserial para los pares significativos:

- Arena vs. Colchoneta: r = 0.44 (efecto mediano a grande)
- Paja vs. Colchoneta: r = 0.27 (efecto pequeño a mediano)

5. Regresión ordinal con covariables

Se ajusta un modelo de logit acumulativo con tipo de cama como predictor principal, controlando por número de parto (1, 2, 3+) y días en leche como covariables:

```
modelo_tarso <- clm(grado_ord ~ tipo_cama + num_parto + dias_leche,
                    data = datos_tarso)
summary(modelo_tarso)
```

Los odds ratios confirman que la colchoneta de goma tiene 4.8 veces los odds de lesiones más graves comparada con arena (IC 95%: 2.1 – 10.9), controlando por número de parto y días en leche.

6. Interpretación clínica

La arena como material de cama se asoció con los menores grados de lesión de tarso. Este hallazgo es consistente con la evidencia internacional que identifica a la arena como el material con menor riesgo de lesiones de extremidades en vacas estabuladas (Chapinal et al., 2013). La colchoneta de goma, a pesar de su durabilidad y facilidad de limpieza, no proporciona la misma protección articular. Desde una perspectiva de bienestar animal y productividad, la inversión en camas de arena puede justificarse en explotaciones lecheras intensivas del trópico alto colombiano, aunque deben considerarse los costos de reposición y manejo del material.

7. ¿Qué habría ocurrido con ANOVA?

Un ANOVA de un factor sobre los puntajes crudos (tratados como numéricos) habría dado $F(2, 102) = 9.87$, $p < 0.001$. La conclusión general coincide, pero los estimadores (medias de 0.74, 1.13 y 1.61) carecen de interpretación biológica directa. Además, los residuos del ANOVA violan la normalidad (Shapiro-Wilk: $W = 0.93$, $p = 0.001$) y la homocedasticidad (Levene: $F = 3.45$, $p = 0.035$), lo que invalida los intervalos de confianza y los valores p del modelo paramétrico.

Parte B: Cálculo del tamaño muestral para un ensayo de eficacia antiparasitaria

Contexto

Un grupo de investigación de la Universidad Mayor de San Andrés (La Paz, Bolivia) planifica un ensayo para comparar la eficacia de dos antihelmínticos (ivermectina vs. closantel) en llamas del altiplano, midiendo la reducción en conteo de HPG a los 14 días postratamiento. Los datos piloto de 10 animales por grupo muestran distribuciones marcadamente asimétricas, con medianas de HPG de 420 (ivermectina) y 680 (closantel) y alta variabilidad intragrupo. La prueba estadística será Mann-Whitney. ¿Cuántos animales por grupo necesitan para una potencia de 0?80 con α = 0.05?

Paso 1: Estimar la probabilidad de superioridad

A partir de los datos piloto, se calcula A de Vargha y Delaney:

```
A_piloto <- VD.A(hpg_ivermectina_piloto, hpg_closantel_piloto)
# Resultado: A ≈ 0.67
```

Paso 2: Fórmula de Noether

```
n_noether <- noether_n(p12 = 0.67, alpha = 0.05, poder = 0.80)
# Resultado: n ≈ 91 por grupo
```

Paso 3: Validar con simulación

Los datos piloto sugieren una distribución log-normal. Se ajustan los parámetros:

```
# Ajustar distribución log-normal a cada grupo piloto
fit_iver <- fitdistr(hpg_ivermectina_piloto, "lognormal")
fit_clos <- fitdistr(hpg_closantel_piloto, "lognormal")

# Simular potencia para n = 80, 85, 90, 95, 100
set.seed(9999)
for (n_g in c(80, 85, 90, 95, 100)) {
  res <- potencia_mann_whitney(
    n_por_grupo = n_g,
    dist_grupo1 = function(n) rlnorm(n, fit_iver$estimate[1],
                                     fit_iver$estimate[2]),
    dist_grupo2 = function(n) rlnorm(n, fit_clos$estimate[1],
                                     fit_clos$estimate[2]),
    nsim = 10000
  )
```

```
  cat("n =", n_g, "-> Potencia:", round(res$potencia, 3), "\n")
}
```

Resultados simulados (aproximados):

n por grupo	Potencia
80	0.74
85	0.77
90	0.80
95	0.83
100	0.85

La simulación confirma que se necesitan aproximadamente 90 llamas por grupo. El investigador decide reclutar 95 para protegerse contra una pérdida del 5% por muertes, escapes o exclusiones durante el período del estudio.

Paso 4: Consideraciones prácticas

Con 190 llamas en total, el estudio es factible en el contexto del altiplano boliviano si se trabaja con comunidades campesinas que manejan rebaños grandes. La aleatorización dentro de rebaño (con rebaño como bloque) podría aumentar la eficiencia y reducir el tamaño muestral necesario, pero requeriría un análisis más complejo (Wilcoxon estratificado o PERMANOVA con factor de bloque).

6.11. Errores frecuentes y advertencias

Tratar puntajes ordinales como datos continuos y calcular medias. Es el error más extendido. Reportar “la media de condición corporal fue 3.2 ± 0.8” implica que la distancia entre BCS 2 y BCS 3 es la misma que entre BCS 4 y BCS 5, y eso no tiene sustento. Reportar mediana y rango intercuartílico, o la distribución completa de frecuencias por categoría.

Usar kappa sin ponderar para datos ordinales. La kappa de Cohen trata como igualmente grave confundir grado I con grado II que confundir grado I con grado III. Para variables con orden, El kappa ponderada (preferiblemente con pesos cuadráticos) es la medida correcta.

Interpretar la kappa fuera de contexto. Un κ de 0.45 puede ser excelente para una evaluación inherentemente subjetiva (bienestar de animales en pastoreo extensivo evaluado por video) y desastroso para una clasificación que debería ser altamente reproducible (presencia/ausencia de un parásito en un frotis). Los umbrales de Landis y Koch son orientaciones, no leyes.

Omitir la prueba de Brant en regresión ordinal. Reportar un modelo de logit acumulativo sin verificar el supuesto de odds proporcionales es como reportar un ANOVA sin verificar homocedasticidad. Si el supuesto se viola, los coeficientes son promedios ponderados de efectos que varían a lo largo de la escala, y la interpretación puede ser engañosa.

Confundir significancia estadística con significancia clínica. Un estudio con 500 animales puede detectar una diferencia estadísticamente significativa de 0.2 puntos en una escala de claudicación de 1 a 5. Esa diferencia es invisible para cualquier observador clínico. El tamaño del efecto pone en perspectiva si el resultado tiene relevancia práctica.

No reportar el tamaño del efecto. Cada vez más revistas lo exigen, pero muchos investigadores siguen reportando solo el valor p. Si el lector solo puede ver que $p = 0.03$, no sabe si la diferencia es grande con pocos animales o trivial con muchos.

Calcular potencia post hoc. La potencia calculada después de obtener los resultados (potencia observada, potencia retrospectiva) es una función directa del valor p y no aporta información adicional. Si p fue pequeño, la potencia post hoc será alta; si p fue grande, será baja. Es un ejercicio circular. La potencia se calcula antes del estudio, durante la planificación, no después.

Usar la fórmula paramétrica de tamaño muestral sin corrección. Para Mann-Whitney, la corrección 1/ARE es el ajuste mínimo. Si la distribución esperada se aleja mucho de la normal, la fórmula paramétrica puede subestimar seriamente el tamaño muestral necesario. La simulación es la herramienta más confiable.

Ignorar el efecto techo y piso en las escalas. Si el 70% de los animales puntúa 0 en una escala de dolor posoperatorio, la prueba de Mann-Whitney tiene poca capacidad para discriminar entre tratamientos. El problema no es la prueba; es la escala, que no tiene suficiente resolución en esa población. Considerar escalas más detalladas o variables respuesta complementarias.

6.12. Ejercicios de autoevaluación

Nivel básico

Ejercicio 6.1.

Clasifique las siguientes variables como nominales, ordinales, de intervalo o de razón. Justifique cada clasificación en una oración:

a) Grado de displasia de cadera según la OFA (normal, borderline, leve, moderada, severa)
b) Temperatura rectal en grados Celsius
c) Raza del animal
d) Puntaje de condición corporal en bovinos (escala 1 a 5)
e) Conteo de colonias bacterianas en una placa de Petri
f) Estadio clínico de linfoma canino (I, II, III, IV, V)

Respuesta: a) Ordinal: las categorías tienen un orden natural de menor a mayor gravedad, pero la distancia entre categorías no es cuantificable. b) De razón: tiene unos cero absolutos (0 K en la escala absoluta) e intervalos iguales, aunque en la práctica veterinaria se trabaja como intervalo. c) Nominal: las categorías no tienen orden inherente. d) Ordinal: el orden refleja mayor reserva grasa, pero la diferencia entre BCS 2 y 3 no es necesariamente igual a la diferencia entre BCS 4 y 5. e) De razón: conteo con cero verdadero e intervalos iguales. f) Ordinal: los estadios reflejan progresión de la enfermedad, pero la diferencia en gravedad entre estadios consecutivos no es uniforme.

Ejercicio 6.2.

Un investigador reporta: "La media del puntaje de dolor posoperatorio fue 2.4 ± 1.1 en el grupo de meloxicam y 3.1 ± 0.9 en el grupo control, con diferencia significativa por prueba de Mann-Whitney ($p = 0.012$)." Identifique al menos dos problemas en este reporte.

Respuesta: Primer problema: reporta media ± DE para una variable ordinal analizada con una prueba no paramétrica. Si se usó Mann-Whitney, las medidas descriptivas apropiadas son la mediana y el rango intercuartílico (o la distribución completa de frecuencias). Segundo problema: no reporta el tamaño del efecto. Debería incluir r de rango biserial o A de Vargha y Delaney. Tercer problema (adicional): no reporta el valor exacto del estadístico U ni el tamaño de cada grupo.

Ejercicio 6.3.

Dos evaluadores clasifican 40 radiografías de cadera canina en tres categorías (normal, sospechosa, displásica). El acuerdo observado es del 72.5% y la kappa de Cohen es 0.58. ¿Por qué la kappa es considerablemente menor que el porcentaje de acuerdo?

Respuesta: Porque la kappa corrige por el acuerdo esperado por azar. Si las distribuciones marginales de ambos evaluadores son desiguales (por ejemplo, si ambos tienden a clasificar más radiografías como "normales"), una proporción sustancial de coincidencias ocurrirá simplemente porque ambos asignan la categoría más frecuente. La kappa descuenta ese acuerdo casual. Un acuerdo observado de 72.5% con kappa de 0.58 indica que aproximadamente un tercio del acuerdo bruto se explica por azar.

Nivel intermedio

Ejercicio 6.4.

En un estudio sobre la eficacia de tres protocolos de limpieza en la desinfección de superficies en quirófanos veterinarios, se clasificó el nivel de contaminación residual como bajo, medio o alto. Los resultados son:

Protocolo	Bajo	Medio	Alto	Total
A	15	8	2	25
B	10	10	5	25
C	6	9	10	25

a) Calcule el estadístico H de Kruskal-Wallis asignando rangos a las 75 observaciones agrupadas.

b) Determine si existe diferencia significativa entre protocolos ($\alpha = 0.05$).

c) Calcule la épsilon cuadrada como medida de tamaño del efecto.

d) Escriba el código en R para realizar este análisis completo.

Respuesta:

a) Codificar: bajo = 1, medio = 2, alto = 3. Las observaciones se expanden y se les asignan rangos. Los rangos medios por grupo serán distintos si los protocolos difieren en el nivel de contaminación. El cálculo manual produce H ≈ 10.47.

b) Con gl = 2, el valor crítico de chi-cuadrado al 5% es 5.991. H = 10.47 > 5.991, por lo tanto, se rechaza H_0. Existe diferencia significativa entre al menos dos protocolos (p ≈ 0.005).

c) $\varepsilon^2 = H / ((n^2 - 1)/(n + 1)) = 10.47 / ((75^2 - 1)/76) = 10.47 / 73.99 = 0.141$ → efecto grande (> 0.14).

d) Código:

```
contaminación <- c(rep(1,15), rep(2,8), rep(3,2),  # Protocolo A
                   rep(1,10), rep(2,10), rep(3,5),  # Protocolo B
                   rep(1,6), rep(2,9), rep(3,10))   # Protocolo C
protocolo <- factor(rep(c("A","B","C"), each = 25))
kruskal.test(contaminación ~ protocolo)
library(rcompanion)
epsilonSquared(contaminación, protocolo)
library(FSA)
dunnTest(contaminación ~ protocolo, method = "holm")
```

Ejercicio 6.5.

Un modelo de regresión ordinal para el grado de endometritis arroja los siguientes coeficientes (parametrización de *clm()*):

Predictor	**β**	**EE**	**p**
Retención de placenta (sí)	1.35	0.40	0.001
Dystocia (sí)	0.78	0.38	0.040
BCS ≤ 2.75 (sí)	0.62	0.33	0.060

a) Calcule los odds ratios para cada predictor.

b) Interprete el odds ratio de la retención de placenta en términos clínicos.

c) La prueba de Brant arroja p = 0.04 para el predictor “distocia” y p = 0.42 para los demás. ¿Qué implica esto?

Respuesta:

a) OR(retención) = exp(1.35) = 3.86; OR(distocia) = exp(0.78) = 2.18; OR(BCS baja) = exp(0.62) = 1.86.

b) Una vaca con retención de placenta tiene 3.86 veces los odds de estar en un grado más alto de endometritis (en cualquier punto de corte de la escala) comparada con una vaca sin retención, controlando por distocia y condición corporal. Es decir, la retención de placenta casi cuadruplica la chance de que la endometritis sea más severa.

c) La prueba de Brant indica que el supuesto de odds proporcionales se viola para la distocia ($p = 0.04$) pero no para los demás predictores. Esto significa que el efecto de la distocia no es constante a lo largo de la escala: puede tener un efecto fuerte para pasar de grado 0 a grado 1, pero un efecto menor para pasar de grado 2 a grado 3, o viceversa. Se debería considerar un modelo de odds parcialmente proporcionales donde el coeficiente de distocia varíe entre ecuaciones, usando *clm()* con *nominal = ~ distocia*.

Ejercicio 6.6.

Un investigador planifica un estudio para comparar dos dietas sobre la condición corporal de ovejas Corriedale. Espera una probabilidad de superioridad (A de Vargha y Delaney) de 0.65. Usando la fórmula de Noether, calcule el tamaño muestral necesario por grupo para una potencia de 0.80 con $\alpha = 0.05$ bilateral.

Respuesta:

$n = (z_{0.025} + z_{0.80})^2 / [3 \times (0.65 - 0.50)^2]$ $n = (1.96 + 0.842)^2 / [3 \times (0.15)^2]$ $n = (2.802)^2 / [3 \times 0.0225]$ $n = 7.851 / 0.0675$ $n = 116.3$

Se necesitan 117 ovejas por grupo. Si la diferencia esperada fuera mayor (A = 0.70): $n = 7.851 / [3 \times 0.04] = 65.4$, es decir 66 por grupo.

Nivel avanzado

Ejercicio 6.7.

Un equipo de investigación evalúa la eficacia de tres protocolos analgésicos (meloxicam, carprofeno y tramadol) sobre el dolor posoperatorio en perras sometidas a ovariohisterectomía

electiva. Se utilizó la escala de dolor de Glasgow (forma corta, puntaje 0 a 24) evaluada a las 2, 4, 8 y 24 horas posquirúrgicas. La muestra incluyó 15 perras por grupo (45 en total). Los datos del evaluador principal a las 4 horas postoperatorias son:

- Meloxicam: 2, 3, 1, 4, 2, 3, 5, 2, 1, 3, 4, 2, 3, 1, 2
- Carprofeno: 3, 4, 5, 3, 6, 4, 3, 5, 4, 7, 3, 4, 5, 6, 4
- Tramadol: 4, 6, 5, 7, 5, 8, 6, 5, 7, 4, 6, 5, 8, 7, 6

Realice un análisis completo que incluya:

a) Estadísticos descriptivos apropiados para datos ordinales.
b) Prueba de Kruskal-Wallis y comparaciones post hoc.
c) Tamaño del efecto global (ε^2) y por pares (r de rango biserial).
d) Modelo de regresión ordinal (agrupando los puntajes en categorías: 0-2 = dolor leve, 3-5 = dolor moderado, 6-8 = dolor severo, > 8 = dolor intenso).
e) Redacción de la sección de resultados como aparecería en un artículo de Journal of Veterinary Internal Medicine.

Rúbrica orientadora:

- Estadísticos descriptivos: debe reportar mediana e IQR para cada grupo (no medias), y puede agregar la distribución de frecuencias por categoría si agrupa los puntajes. (2 puntos)
- Kruskal-Wallis: debe reportar H, gl, p exacto, e incluir la verificación de que la prueba compara distribuciones con la misma forma. (2 puntos)
- Post hoc: debe usar Dunn o Conover con corrección por comparaciones múltiples (Holm o Bonferroni), presentar las tres comparaciones por pares. (2 puntos)
- Tamaño del efecto: ε^2 global y r por pares, con interpretación según umbrales de Cohen y, preferiblemente, con interpretación clínica propia. (2 puntos)
- Regresión ordinal: modelo ajustado con verificación de odds proporcionales, odds ratios con IC, interpretación. (3 puntos)

- Redacción: sigue normas APA 7, incluye estadísticos de prueba, tamaño del efecto, IC cuando aplique, usa mediana e IQR como descriptivos, es concisa y no confunde significancia estadística con clínica. (3 puntos)

Ejercicio 6.8.

Diseñe una simulación de potencia en R para la siguiente situación: un investigador quiere comparar la eficacia de un nuevo antihelmíntico (grupo tratado) contra ivermectina (grupo control) en bovinos, usando el test de reducción de conteo fecal de huevos (FECRT). La variable es el porcentaje de reducción de HPG. Los datos piloto sugieren que la ivermectina produce una reducción mediana del 85% con alta variabilidad, y el nuevo fármaco una reducción del 95%. Las distribuciones son asimétricas con cola izquierda. El análisis será Mann-Whitney. Escriba el código completo de simulación, grafique la curva de potencia para n = 10 a 50 por grupo, e identifique el tamaño muestral mínimo para potencia de 0.80.

Rúbrica orientadora:

- La simulación debe usar distribuciones que reflejen la asimetría descrita (por ejemplo, distribución beta reescalada, o distribución log-normal truncada). (2 puntos)
- Debe incluir *set.seed()* para reproducibilidad. (1 punto)
- El número de réplicas debe ser al menos 5 000, preferiblemente 10 000. (1 punto)
- La curva de potencia debe graficarse con ggplot2 e incluir la línea de referencia de 0.80. (2 puntos)
- Debe verificar el error tipo I bajo H_0 (simulación con ambos grupos iguales, proporción de rechazos ≈ 0.05). (2 puntos)
- Debe identificar el n mínimo y discutir la viabilidad práctica en el contexto de un estudio con bovinos. (2 puntos)

6.13. Conclusiones del capítulo

Las escalas de evaluación clínica son la herramienta de medición más frecuente y más frecuentemente mal analizada en la investigación veterinaria. La condición corporal, el dolor, la claudicación, la gravedad de lesiones y los grados histológicos producen datos ordinales que no

admiten medias ni desviaciones estándar, por más que el software permita calcularlas. Los métodos de rangos cubiertos en capítulos anteriores (Mann-Whitney, Kruskal-Wallis, Friedman) son apropiados para comparar grupos, pero la regresión ordinal abre la puerta a modelar efectos multivariados: el tipo de cama, el número de parto y la condición corporal pueden evaluarse simultáneamente como predictores del grado de lesión, algo que una prueba de Kruskal-Wallis no permite.

Cuando la evaluación depende del juicio humano, la concordancia entre evaluadores no es un detalle metodológico sino una medida de la calidad de los datos. El kappa ponderada cuantifica esa concordancia descontando el azar y penalizando los desacuerdos en proporción a su gravedad. Reportarlo debería ser obligatorio en cualquier estudio que use escalas de evaluación clínica.

El tamaño del efecto y la potencia estadística completan el panorama que el valor p deja incompleto. Un resultado significativo dice que probablemente hay algo, pero no dice cuánto. Un resultado no significativo dice que no se detectó nada, pero no dice si el estudio tenía capacidad para detectarlo. Las medidas de tamaño del efecto (r de rango biserial, A de Vargha y Delaney, épsilon cuadrada, V de Cramér) ponen números a la magnitud de las diferencias y asociaciones. La estimación del tamaño muestral, ya sea con fórmulas de Noether o con simulación, permite diseñar estudios que tengan una probabilidad razonable de encontrar lo que buscan. En un contexto donde cada animal experimental tiene un costo económico, ético y logístico, diseñar un estudio con potencia insuficiente es un desperdicio de recursos que ningún comité de ética debería aprobar.

El capítulo siguiente amplía la perspectiva hacia métodos multivariados (PERMANOVA, ANOSIM, NMDS) y modelos de regresión no paramétrica (GAM) que permiten abordar la complejidad de los datos veterinarios cuando las preguntas involucran múltiples variables respuesta o relaciones no lineales entre predictores y desenlaces.

Capítulo 7

Métodos multivariados, regresión no paramétrica y modelos aditivos generalizados

Introducción

Este capítulo trabaja los métodos para dos situaciones que exceden el alcance de las pruebas univariadas: el análisis de respuestas multivariadas (cuando cada unidad experimental tiene un perfil de múltiples variables en lugar de un solo valor) y el modelado de relaciones no lineales entre una respuesta y uno o más predictores. En la primera parte se desarrollan los métodos multivariados basados en distancias. Se construyen las matrices de disimilitud (Bray-Curtis para datos de abundancia, Jaccard para presencia-ausencia, euclidiana para variables continuas) y se trabajan tres herramientas analíticas: PERMANOVA para evaluar si la composición multivariada difiere entre grupos mediante el pseudo-F de Anderson y permutaciones, ANOSIM como alternativa basada en rangos de distancias con su estadístico R global, y NMDS (escalamiento multidimensional no métrico) para representar las relaciones de similitud entre unidades en un espacio de dos o tres dimensiones, con el estrés como medida de la calidad de la representación. Se incluye la verificación de dispersiones con PERMDISP como requisito para interpretar correctamente PERMANOVA. En la segunda parte se presentan tres enfoques de regresión no paramétrica. LOESS como suavizador local que ajusta regresiones ponderadas en ventanas de datos. Los modelos aditivos generalizados (GAM) con splines penalizados, incluyendo la selección automática del grado de suavizado, la verificación de los supuestos mediante DHARMa, y la extensión a distribuciones no gaussianas (Poisson, binomial negativa, binomial). La regresión cuantílica para modelar medianas y otros percentiles condicionales cuando la heterogeneidad de la respuesta es parte del fenómeno biológico de interés. Todo se implementa en R con los paquetes vegan, mgcv, gratia y quantreg.

Objetivos de aprendizaje

Al finalizar este capítulo, el lector será capaz de:

1. **Seleccionar** la medida de distancia o disimilitud apropiada (Bray-Curtis, Jaccard, euclidiana) según la naturaleza de los datos ecológicos o parasitológicos en contextos veterinarios.
2. **Aplicar e interpretar** PERMANOVA y ANOSIM para evaluar diferencias en composición multivariada entre grupos de animales o sistemas productivos, y distinguir cuándo cada método resulta más adecuado.

3. **Construir y evaluar** ordenaciones NMDS para visualizar patrones de similitud entre comunidades parasitarias, microbianas o ecológicas asociadas a la producción animal.
4. **Ajustar modelos aditivos generalizados (GAM)** con distintas distribuciones (gaussiana, Poisson, binomial negativa) para relaciones no lineales en datos veterinarios, e interpretar las funciones suaves resultantes.
5. **Implementar regresión cuantílica** para modelar medianas y otros percentiles de respuesta en situaciones donde la media condicional no captura la heterogeneidad del fenómeno biológico.

Introducción: cuando un solo número no alcanza

Un grupo de investigación en la provincia de Bolívar, Ecuador, recoge garrapatas de 120 bovinos mestizos distribuidos en tres sistemas productivos: extensivo de páramo, semi-intensivo de valle interandino e intensivo periurbano. De cada animal extraen e identifican las garrapatas hasta género: *Rhipicephalus*, *Amblyomma*, *Dermacentor* e *Ixodes*. El resultado no es una sola variable, sino un perfil de cuatro conteos por animal. La pregunta tampoco es simple: ¿difiere la *composición* de la comunidad de ectoparásitos entre los tres sistemas?

Un ANOVA de una vía podría comparar el conteo total de garrapatas. Pero eso ignora que dos fincas con el mismo número total de garrapatas pueden tener perfiles completamente distintos: una dominada por *Amblyomma* y otra por *Rhipicephalus*. Un MANOVA paramétrico exigiría normalidad multivariada y matrices de covarianza homogéneas, supuestos que datos de conteo con muchos ceros jamás satisfacen.

En el mismo proyecto, otra línea de trabajo estudia la relación entre los días en leche y el recuento de células somáticas (RCS) en vacas Holstein con mastitis subclínica. La relación no es lineal: el RCS desciende las primeras semanas posparto, se estabiliza y vuelve a subir hacia el final de la lactancia, con un patrón que varía según la época del año. Un modelo de regresión lineal múltiple forzaría una recta donde la biología dibuja una curva.

Este capítulo presenta las herramientas para ambos problemas. La primera mitad aborda los métodos multivariados basados en distancias PERMANOVA, ANOSIM y NMDS que permiten comparar composiciones completas sin asumir distribuciones específicas. La segunda mitad

desarrolla la regresión no paramétrica: desde el suavizado local (LOESS) hasta los modelos aditivos generalizados (GAM) y la regresión cuantílica. Lo que une a todas estas técnicas es su capacidad para dejar que los datos hablen sin imponerles una forma predeterminada.

7.1. Matrices de distancia y disimilitud: Bray-Curtis, Jaccard, euclidiana

7.1.1. El concepto de distancia entre unidades multivariadas

Cuando cada animal se describe por múltiples variables simultáneas conteos de especies parasitarias, abundancias de géneros bacterianos en un estudio de microbioma ruminal, o proporciones de ácidos grasos en leche, la primera operación analítica consiste en cuantificar cuán diferentes son dos animales entre sí. Esa cuantificación se almacena en una **matriz de distancia** (o disimilitud): una matriz simétrica de $n \times n$ donde cada celda contiene la distancia entre el animal i y el animal j.

La elección de la medida de distancia no es un detalle técnico menor. Determina qué aspectos de la diferencia se enfatizan y cuáles se ignoran. Elegir mal la distancia equivale a medir con una regla equivocada: los números salen, pero no significan lo que uno cree.

7.1.2. Distancia euclidiana

La más conocida. Para dos animales con p variables medidas:

$$d_E(i,j) = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}$$

donde x_{ik} es el valor de la variable k en el animal i.

Es la extensión directa de la distancia geométrica al espacio de p dimensiones. Funciona bien cuando las variables son continuas, están en la misma escala (o se estandarizan previamente) y no contienen muchos ceros. En ecología parasitaria, su uso directo con datos de abundancia crudos es problemático: dos animales sin ninguna especie de parásito aparecerían como idénticos (distancia cero), aunque la ausencia compartida no implica similitud ecológica real. Este fenómeno se conoce como el **problema del doble cero**.

Ejemplo numérico. Tres bovinos con conteos de dos géneros de garrapatas:

Animal	*Rhipicephalus*	*Amblyomma*
A	12	3
B	8	7
C	0	0

$$d_E(A,B) = \sqrt{(12-8)^2 + (3-7)^2} = \sqrt{16+16} = \sqrt{32} \approx 5.66$$

$$d_E(A,C) = \sqrt{(12-0)^2 + (3-0)^2} = \sqrt{144+9} = \sqrt{153} \approx 12.37$$

$$d_E(B,C) = \sqrt{(8-0)^2 + (7-0)^2} = \sqrt{64+49} = \sqrt{113} \approx 10.63$$

Nótese que la distancia entre A y C es grande porque A tiene parásitos y C no. Pero si hubiera un animal D, también libre de parásitos, $d_E(C,D) = 0$, lo cual atribuye similitud máxima a dos animales que simplemente no tienen garrapatas, quizá por razones completamente distintas (tratamiento acaricida reciente vs. altitud donde no hay garrapatas).

7.1.3. Disimilitud de Bray-Curtis

Diseñada para datos de abundancia ecológica, la disimilitud de Bray-Curtis evita el problema del doble cero:

$$d_{BC}(i,j) = \frac{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|}{\sum_{k=1}^{p} \left(x_{ik} + x_{jk} \right)}$$

El numerador suma las diferencias absolutas en abundancia para cada especie. El denominador es la suma total de abundancias de ambos animales. El resultado varía entre 0 (composición idéntica) y 1 (ninguna especie compartida). Si ambos animales tienen cero en todas las especies, la fórmula produce 0/0 y queda indefinida; en la práctica, esas observaciones se excluyen o se les asigna un valor arbitrario. Bray-Curtis no es una distancia métrica en sentido estricto porque no satisface la desigualdad triangular en todos los casos, pero ese detalle rara vez afecta los análisis basados en permutaciones.

Con los datos del ejemplo anterior:

$$d_{BC}(A,B) = \frac{|12-8| + |3-7|}{(12+8) + (3+7)} = \frac{4+4}{20+10} = \frac{8}{30} = 0.267$$

$$d_{BC}(A,C) = \frac{|12-0| + |3-0|}{(12+0) + (3+0)} = \frac{15}{15} = 1.0$$

El animal C, sin parásitos, tiene disimilitud máxima respecto a cualquier animal con parásitos. Y la comparación C-D (ambos sin parásitos) simplemente no se define, lo cual es ecológicamente más honesto que declararlos idénticos.

7.1.4. Índice de Jaccard (para presencia-ausencia)

Cuando los datos se reducen a presencia (1) o ausencia (0) de cada especie, el índice de Jaccard mide la proporción de especies compartidas:

$$J(i,j) = \frac{a}{a+b+c}$$

donde a = número de especies presentes en ambos animales, b = especies presentes solo en i, c = especies presentes solo en j. La disimilitud de Jaccard es $d_J = 1 - J$. Varía entre 0 (misma composición) y 1 (ninguna especie compartida). Al igual que Bray-Curtis, ignora las ausencias compartidas.

En un estudio de seroprevalencia de enfermedades transmitidas por garrapatas en perros de Guayaquil, donde se detectó presencia o ausencia de anticuerpos contra *Ehrlichia canis*, *Anaplasma platys*, *Babesia canis* y *Borrelia burgdorferi*, Jaccard permitiría comparar los perfiles de exposición entre barrios sin que la ausencia simultánea de dos enfermedades raras contribuya artificialmente a la similitud.

7.1.5. Otras medidas y criterios de selección

Tabla 37

Medidas de distancia y disimilitud comunes en análisis multivariado ecológico y veterinario.

Medida	Tipo de datos	Sensible a doble cero	Rango	Uso típico en veterinaria
Euclidiana	Continuos, estandarizados	Sí	[0, ∞)	Variables fisiológicas, hemograma
Bray-Curtis	Abundancias (conteos)	No	[0, 1]	Comunidades parasitarias, microbioma
Jaccard	Presencia/ausencia	No	[0, 1]	Seroprevalencia, catálogos de especies
Sørensen	Presencia/ausencia	No	[0, 1]	Similar a Jaccard, pondera más la co-ocurrencia
Gower	Mixtos (continuos + categóricos)	Depende	[0, 1]	Caracterización multivariada de fincas
Manhattan	Continuos	Sí	[0, ∞)	Alternativa a euclidiana, menos sensible a valores extremos

Criterio práctico de selección. Si los datos son conteos de especies o abundancias, la opción predeterminada es Bray-Curtis. Si son presencia-ausencia, Jaccard. Si son variables continuas medidas en distintas escalas (peso en kg, temperatura en °C, concentración de urea en mmol/L), estandarizar primero y usar euclidiana o Gower si hay mezcla con categóricas. No hay una medida universalmente correcta; la elección debe justificarse en función de la pregunta biológica y la naturaleza de los datos.

7.2. PERMANOVA: lógica, pseudo-F de Anderson y verificación de dispersiones

7.2.1. El problema y la solución

Se quiere saber si la composición de la comunidad de ectoparásitos difiere entre tres sistemas productivos bovinos. Los datos son una tabla de animales × especies con conteos. La hipótesis nula es que los centroides (puntos medios en el espacio multivariado) de los tres grupos son iguales. MANOVA resolvería esto en el mundo paramétrico, pero exige normalidad multivariada

y homogeneidad de matrices de covarianza: condiciones que datos de conteo con exceso de ceros violan sin remedio.

PERMANOVA (Permutational Multivariate Analysis of Variance; Anderson, 2001) ofrece una alternativa basada en distancias y permutaciones. No asume ninguna distribución particular de los datos. La única entrada es la matriz de distancias entre las unidades experimentales.

7.2.2. Lógica de la prueba

La idea es la misma que en ANOVA: particionar la variabilidad total en un componente entre grupos y un componente dentro de grupos. Pero en lugar de trabajar con las observaciones crudas, PERMANOVA trabaja con las distancias al cuadrado.

Sea Δ la matriz de distancias de $n \times n$. La suma de cuadrados total se define como:

$$SS_T = \frac{1}{n}\sum_{i<j} d_{ij}^2$$

donde d_{ij} es la distancia entre las unidades i y j. La suma de cuadrados dentro del grupo g es:

$$SS_{W_g} = \frac{1}{n_g}\sum_{i<j\in g} d_{ij}^2$$

y la suma de cuadrados dentro total es $SS_W = \sum_{g=1}^{a} S\,S_{W_g}$, donde a es el número de grupos. La suma de cuadrados entre grupos se obtiene por diferencia:

$$SS_A = SS_T - SS_W$$

El estadístico pseudo-F de Anderson es:

$$F = \frac{SS_A/(a-1)}{SS_W/(n-a)}$$

Este estadístico tiene la misma estructura algebraica que el F de ANOVA, pero no sigue una distribución F bajo la hipótesis nula. Por eso se llama *pseudo*-F. La distribución nula se construye por permutaciones.

7.2.3. La distribución nula por permutaciones

Bajo H_0 (los grupos no difieren en composición), la asignación de cada animal a un grupo es intercambiable. El procedimiento es:

1. Calcular F_{obs} con la asignación original de los animales a los grupos.
2. Permutar aleatoriamente las etiquetas de grupo entre los animales (sin cambiar los datos ni la matriz de distancias).
3. Recalcular F con las etiquetas permutadas. Llamar a este valor F^*.
4. Repetir los pasos 2-3 un gran número de veces (típicamente 999 o 9999 permutaciones).
5. El valor *p* es la proporción de valores F^* que igualan o superan F_{obs}:

$$p = \frac{\#\{F^* \geq F_{obs}\} + 1}{N_{perm} + 1}$$

donde el +1 en numerador y denominador incluye el estadístico observado en la distribución de referencia. Esto garantiza que *p* nunca sea exactamente cero.

7.2.4. Supuestos de PERMANOVA

PERMANOVA no asume normalidad ni linealidad, pero sí requiere:

1. **Independencia de las observaciones.** Si los animales de una misma finca están correlacionados (por manejo compartido, genética similar), esa dependencia debe incorporarse mediante restricción de las permutaciones a bloques (fincas).
2. **Homogeneidad de dispersiones multivariadas.** PERMANOVA es sensible a diferencias en la dispersión (variabilidad) dentro de cada grupo, análoga a la heterogeneidad de varianzas en ANOVA. Si un grupo es más disperso que otro, PERMANOVA puede arrojar un resultado significativo incluso cuando los centroides no difieren. Por eso es obligatorio verificar las dispersiones con PERMDISP.

7.2.5. PERMDISP: verificación de dispersiones

PERMDISP (Anderson, 2006) prueba si las dispersiones multivariadas de los grupos son homogéneas. La dispersión de un grupo se define como la distancia media (o mediana) de cada observación al centroide del grupo en el espacio de coordenadas principales.

El procedimiento calcula la distancia de cada observación al centroide de su grupo, y luego aplica un ANOVA de una vía (o su versión basada en permutaciones) sobre esas distancias. Si PERMDISP arroja $p < 0.05$, la diferencia detectada por PERMANOVA puede deberse a diferencias en dispersión, no (o no solo) en localización de los centroides.

Interpretación conjunta:

PERMANOVA	PERMDISP	Interpretación
Significativo	No significativo	Los grupos difieren en composición (centroides distintos)
Significativo	Significativo	Hay diferencia, pero puede ser en composición, en dispersión, o en ambas
No significativo	Significativo	Los centroides son similares, pero la variabilidad difiere entre grupos
No significativo	No significativo	No hay evidencia de diferencias en composición ni en dispersión

7.2.6. Ejemplo resuelto: comunidad de ectoparásitos en bovinos

Datos. Quince bovinos mestizos, cinco por sistema productivo (extensivo, semi-intensivo, intensivo), con conteos de cuatro géneros de garrapatas.

Tabla 38
Conteos de géneros de garrapatas en 15 bovinos según sistema productivo.

Animal	Sistema	*Rhipicephalus*	*Amblyomma*	*Dermacentor*	*Ixodes*
1	Extensivo	2	18	5	8
2	Extensivo	1	22	3	12
3	Extensivo	0	15	7	6
4	Extensivo	3	20	4	10
5	Extensivo	1	17	6	9

Animal	Sistema	*Rhipicephalus*	*Amblyomma*	*Dermacentor*	*Ixodes*
6	Semi-int.	10	8	2	3
7	Semi-int.	14	5	1	2
8	Semi-int.	12	6	3	1
9	Semi-int.	9	9	2	4
10	Semi-int.	11	7	1	2
11	Intensivo	25	1	0	0
12	Intensivo	30	0	1	0
13	Intensivo	22	2	0	1
14	Intensivo	28	0	0	0
15	Intensivo	26	1	1	0

El patrón es claro a simple vista: el sistema extensivo de páramo muestra dominancia de *Amblyomma* e *Ixodes*; el intensivo, dominancia casi exclusiva de *Rhipicephalus* (posiblemente *R. microplus*, adaptado a climas cálidos y manejo confinado); el semi-intensivo es intermedio. Pero la estadística formal lo confirma y lo cuantifica.

Procedimiento:

1. Calcular la matriz de disimilitud de Bray-Curtis (15 × 15).
2. Aplicar PERMANOVA con 9999 permutaciones.
3. Verificar dispersiones con PERMDISP.

La implementación en R se presenta en la sección 7.10.

7.3. ANOSIM: estadístico R global y comparación con PERMANOVA

7.3.1. Fundamento de ANOSIM

ANOSIM (Analysis of Similarities; Clarke, 1993) es otra prueba basada en permutaciones para comparar composiciones multivariadas entre grupos. A diferencia de PERMANOVA, que particiona sumas de cuadrados, ANOSIM compara directamente los rangos de las distancias.

El procedimiento es:

1. Calcular la matriz de distancias y convertir todas las distancias en rangos (de menor a mayor).
2. Separar los rangos en dos conjuntos: r_W (rangos de las distancias entre pares de observaciones del mismo grupo) y r_B (rangos de las distancias entre pares de grupos distintos).
3. Calcular el estadístico R:

$$R = \frac{\bar{r}_B - \bar{r}_W}{n(n-1)/4}$$

donde $\bar{r}_B$ es el rango medio de las distancias entre grupos y $\bar{r}_W$ es el rango medio de las distancias dentro de grupos. El denominador es el máximo valor posible de la diferencia $\bar{r}_B - \bar{r}_W$.

R varía entre −1 y +1: - **R = 1**: todas las distancias entre grupos son mayores que todas las distancias dentro de grupos (separación perfecta). - **R = 0**: no hay diferencia entre las distancias inter e intragrupales (los grupos se superponen completamente). - **R < 0**: las distancias dentro de grupo son mayores que las distancias entre grupos (situación biológicamente rara, generalmente indica un problema en los datos o el diseño).

La significancia se evalúa por permutaciones, igual que en PERMANOVA.

7.3.2. ANOSIM vs. PERMANOVA: ¿cuándo usar cada una?

La Tabla 7.3 resume las diferencias.

Tabla 39
Comparación entre PERMANOVA y ANOSIM.

Aspecto	PERMANOVA	ANOSIM
Estadístico	Pseudo-F (cociente de varianzas)	R (diferencia de rangos medios)
Sensibilidad a dispersiones	Alta (puede confundir localización con dispersión)	Menor (trabaja con rangos)
Diseños complejos	Sí (factoriales, bloques, anidados)	Solo diseños simples de una vía

Aspecto	PERMANOVA	ANOSIM
Potencia estadística	Generalmente mayor	Menor, especialmente con muestras pequeñas
Interpretabilidad	Similar a ANOVA (efecto relativo del factor)	El valor de R es directamente una medida de separación
Uso actual	Más recomendado en la literatura reciente	Clásico, todavía frecuente

En la práctica, PERMANOVA es la opción preferida para la mayoría de los diseños. ANOSIM sigue siendo útil como análisis complementario, especialmente cuando se sospecha que una diferencia significativa en PERMANOVA puede deberse a heterogeneidad de dispersiones. Si PERMANOVA es significativa y ANOSIM también, con un R alto, la evidencia de diferencia en composición es más robusta.

7.4. NMDS: ordenación no paramétrica y estrés como medida de ajuste

7.4.1. ¿Qué es una ordenación y para qué sirve?

Los datos multivariados viven en un espacio de *p* dimensiones. Con cuatro géneros de garrapatas, cada animal es un punto en un espacio de cuatro dimensiones. No podemos visualizar eso directamente. Una **ordenación** es una técnica que proyecta esos puntos en dos o tres dimensiones, preservando en lo posible las relaciones de similitud entre ellos.

El análisis de componentes principales (PCA) es la ordenación paramétrica por excelencia: busca los ejes de máxima varianza en el espacio original. Pero PCA asume relaciones lineales y funciona mejor con distancias euclidianas, lo cual lo hace inadecuado para datos ecológicos con muchos ceros y relaciones no lineales.

7.4.2. El escalamiento multidimensional no métrico (NMDS)

NMDS (Non-metric Multidimensional Scaling; Kruskal, 1964) resuelve el problema de forma elegante. Su objetivo es representar los *n* puntos en *k* dimensiones (usualmente $k = 2$) de modo que

el **orden** de las distancias en el espacio reducido reproduzca el orden de las distancias originales. No intenta preservar las distancias exactas, solo su ranking.

El algoritmo funciona por iteración:

1. Colocar los *n* puntos en posiciones iniciales aleatorias en el espacio de *k* dimensiones.
2. Calcular las distancias entre todos los pares de puntos en el espacio reducido.
3. Comparar el orden (ranking) de estas distancias con el orden de las distancias originales.
4. Mover los puntos para reducir la discordancia entre ambos órdenes.
5. Repetir hasta convergencia.

La medida de discordancia se llama **estrés (stress)**. La fórmula de Kruskal es:

$$\text{Stress} = \sqrt{\frac{\sum_{i<j} \left(d_{ij}^{*} - \hat{d}_{ij}\right)^2}{\sum_{i<j} d_{ij}^{*2}}}$$

donde d_{ij}^{*} es la distancia en el espacio reducido y $\hat{d}_{ij}$ es la distancia ajustada por una regresión monótona (isotónica) de las distancias reducidas sobre las distancias originales.

7.4.3. Interpretación del estrés

Clarke (1993) propuso la siguiente guía, que sigue siendo la referencia más citada:

Tabla 40
Interpretación del valor de estrés en NMDS

Estrés	Interpretación
< 0.05	Representación excelente
0.05 – 0.10	Buena representación
0.10 – 0.20	Aceptable, interpretar con cautela
0.20 – 0.30	Representación pobre; considerar más dimensiones
> 0.30	La configuración es prácticamente arbitraria

Un estrés de 0.15 en un estudio con 15 bovinos y cuatro variables es perfectamente aceptable. Un estrés de 0.25 con 80 muestras de microbioma ruminal y 200 OTUs sugiere que dos dimensiones no capturan la estructura; podría intentarse una solución en tres dimensiones.

7.4.4. Consideraciones prácticas

- **Inicios aleatorios.** Como NMDS usa iteración desde posiciones iniciales aleatorias, puede converger a soluciones locales distintas. La práctica estándar es ejecutar el algoritmo con múltiples inicios (20 o más) y quedarse con la solución de menor estrés.
- **Número de dimensiones.** Empezar con $k = 2$. Si el estrés es inaceptable, probar $k = 3$. Más de tres dimensiones pierde el propósito de la visualización.
- **Las distancias entre ejes no son interpretables.** A diferencia de PCA, los ejes de NMDS no representan gradientes lineales. El eje 1 no es "más importante" que el eje 2. Lo que importa es la posición relativa de los puntos: puntos cercanos son similares, puntos lejanos son distintos.
- **Se puede rotar, reflejar o trasladar la solución sin alterarla.** El NMDS no tiene una orientación canónica.

7.5. Regresión local: LOESS/LOWESS

7.5.1. El problema de la no linealidad

En producción lechera, la relación entre días en leche (DEL) y producción diaria sigue la clásica curva de Wood (1967): ascenso rápido hasta un pico, seguido de un descenso gradual. Un modelo lineal simple subestima la producción al inicio, la sobreestima en el pico y vuelve a fallar al final. Se podría ajustar un modelo de Wood paramétrico ($y = a \cdot t^b \cdot e^{-ct}$), pero eso requiere asumir la forma funcional de antemano. ¿Qué pasa cuando no se conoce la forma de la relación, o cuando esta es más compleja que cualquier ecuación estándar?

7.5.2. Fundamento de LOESS

LOESS (LOcally Estimated Scatterplot Smoothing; Cleveland, 1979; Cleveland & Devlin, 1988) ajusta la relación entre *x* e *y* de forma local. Para estimar el valor suavizado en un punto x_0:

1. Se seleccionan los *k* vecinos más cercanos a x_0 (donde *k* es una fracción del total de datos, controlada por el parámetro **span** o **α**).
2. Se asignan pesos a cada vecino según su distancia a x_0, usando una función de pesos tricúbica:

$$w_i = \left(1 - \left(\frac{|x_i - x_0|}{d_{max}}\right)^3\right)^3$$

donde d_{max} es la distancia al vecino más lejano dentro de la ventana. Los puntos cercanos reciben peso alto; los lejanos, peso casi nulo.

1. Se ajusta una regresión ponderada (lineal o cuadrática) con esos pesos.
2. El valor ajustado en x_0 es la predicción de esa regresión local.

Este proceso se repite para cada punto del rango de *x*, generando la curva suavizada.

7.5.3. El parámetro span

El span controla el grado de suavizado:

- **Span grande** (e.g., 0.75): la ventana incluye el 75% de los datos. La curva es muy suave, casi lineal. Riesgo de sesgo (subajuste).
- **Span pequeño** (e.g., 0.15): la ventana incluye el 15% de los datos. La curva sigue de cerca los datos. Riesgo de variabilidad excesiva (sobreajuste).

No hay una fórmula mágica para elegir el span. En la práctica, se usan criterios de validación cruzada o se ajustan varios valores y se elige el que produce residuos con menor patrón sistemático.

7.5.4. Ejemplo veterinario: curva de lactancia

En un hato de 60 vacas Holstein de una finca en Cayambe (Ecuador), se registró la producción diaria de leche (litros) en controles mensuales durante 305 días de lactancia. Un LOESS con span = 0.3 revela la curva de lactancia sin asumir la forma de Wood ni ninguna otra parametrización. El código se presenta en la sección 7.10.

7.5.5. Limitaciones de LOESS

- No produce una ecuación explícita. La predicción requiere los datos originales.
- No permite inferencia formal (intervalos de confianza, pruebas de hipótesis sobre los parámetros) de forma directa.

- Es computacionalmente intensivo con muestras grandes.
- No maneja bien datos multivariados (más de dos predictores continuos simultáneos).

Estas limitaciones motivaron el desarrollo de los modelos aditivos generalizados.

7.6. Modelos aditivos generalizados (GAM): de GLM a funciones suaves

7.6.1. Del modelo lineal al modelo aditivo

Recapitulemos la jerarquía:

- **Modelo lineal (LM):** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, con $\varepsilon \sim N(0, \sigma^2)$.
- **Modelo lineal generalizado (GLM):** permite distribuciones no normales (Poisson, binomial, gamma) y relaciones no lineales *a través de la función de enlace*, pero la estructura del predictor lineal sigue siendo $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Cada predictor tiene un efecto lineal sobre η.
- **Modelo aditivo generalizado (GAM):** reemplaza algunos o todos los términos lineales $\beta_j X_j$ por funciones suaves $f_j(X_j)$:

$$g(\mu) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

donde $g(\cdot)$ es la función de enlace y $\mu = E(Y)$.

La diferencia entre un GLM y un GAM es conceptualmente simple: en el GLM, el efecto de cada predictor es una recta (en la escala del predictor lineal). En el GAM, el efecto puede ser una curva de forma libre. El modelo *estima* la forma de esa curva a partir de los datos.

7.6.2. ¿Qué son las funciones suaves?

Las funciones suaves $f_j(X_j)$ se construyen como combinaciones lineales de **funciones base** (basis functions). La idea es similar a construir una curva con piezas de Lego: cada función base es una pieza con una forma predefinida, y la combinación ponderada de muchas piezas produce una curva flexible.

$$f(x) = \sum_{m=1}^{M} b_m \cdot B_m(x)$$

donde $B_m(x)$ es la m-ésima función base y b_m es su coeficiente (estimado por el modelo). Las funciones base más usadas son:

- **Splines cúbicos de regresión** (cr): polinomios cúbicos por tramos, unidos suavemente en puntos llamados nodos (knots).
- **Splines de placas delgadas** (tp, thin plate regression splines): la opción predeterminada en *mgcv* de R. No requieren especificar la posición de los nodos; la determinan automáticamente.
- **Splines cíclicos** (cc): para variables periódicas, como el mes del año o la hora del día. Garantizan que $f(x_{min}) = f(x_{max})$.

7.6.3. El dilema del suavizado: penalización

Si se usan muchas funciones base, la curva será muy ondulada (sobreajuste). Si se usan pocas, será demasiado rígida (subajuste). El GAM resuelve esto con un enfoque de **penalización**: permite un número relativamente grande de funciones base (controlado por el parámetro **k**, la dimensión de la base), pero penaliza la ondulación excesiva.

El objetivo de estimación es minimizar:

$$\sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int [f''(x)]^2 dx$$

El primer término mide el ajuste a los datos (como en mínimos cuadrados). El segundo término es la penalización por "rugosidad": la integral del cuadrado de la segunda derivada de f. Una función muy ondulada tiene segunda derivada grande y recibe penalización alta.

El parámetro λ controla el balance: - $\lambda = 0$: sin penalización, la función interpola los datos (sobreajuste máximo). - $\lambda \to \infty$: penalización infinita, la función se reduce a una recta (el GAM colapsa a un GLM).

El valor óptimo de λ se estima automáticamente por **validación cruzada generalizada (GCV)** o por **máxima verosimilitud restringida (REML)**, que es el método recomendado actualmente por Wood (2017).

7.6.4. Grados de libertad efectivos (edf)

En un modelo lineal, cada parámetro consume un grado de libertad. En un GAM, el suavizado consume **grados de libertad efectivos** (edf, effective degrees of freedom), que no son necesariamente enteros. Los edf miden la complejidad de la curva ajustada:

- **edf ≈ 1**: la función es prácticamente una recta. El efecto es lineal.
- **edf = 2-3**: una curva suave simple (un máximo o un mínimo).
- **edf > 5**: una curva con varias inflexiones. Verificar que la ondulación tenga sentido biológico.

Los edf aparecen en la salida de *summary()* de un GAM en R y son la primera cifra que debe revisarse. Un GAM que devuelve edf = 1 para un predictor está diciendo que la relación es lineal y el suavizado fue innecesario.

7.7. GAM para distribuciones no normales: Poisson, binomial negativa, binomial

7.7.1. El principio: la familia del GLM con la flexibilidad del GAM

La potencia de los GAM reside en que combinan dos generalizaciones: la distribución de la respuesta (del GLM) y la forma funcional del predictor (funciones suaves en lugar de rectas). Esto los hace adecuados para la mayoría de las variables respuesta que se encuentran en investigación veterinaria.

7.7.2. GAM con distribución de Poisson

Para datos de conteo (número de garrapatas, huevos de parásitos por gramo de heces, casos de enfermedad por finca):

$$\log(\mu_i) = \beta_0 + f_1(X_{1i}) + f_2(X_{2i})$$

donde $Y_i \sim \text{Poisson}(\mu_i)$. La función de enlace es el logaritmo natural, lo que garantiza que $\mu_i > 0$.

Supuesto crítico: en la distribución de Poisson, la media es igual a la varianza. En datos reales de conteo parasitario, la varianza casi siempre excede la media (sobredispersión). Si se ignora la sobredispersión, los errores estándar de los coeficientes se subestiman y los intervalos de confianza son demasiado estrechos, con el riesgo consecuente de declarar efectos significativos que no lo son.

7.7.3. GAM con distribución binomial negativa

La solución natural para la sobredispersión en conteos es la distribución binomial negativa:

$$\log(\mu_i) = \beta_0 + f_1(X_{1i}) + f_2(X_{2i})$$

con $Y_i \sim \text{NB}(\mu_i, \theta)$, donde θ es el parámetro de dispersión (a veces llamado *size* o *k*). La varianza es $\mu + \mu^2/\theta$. Cuando $\theta \to \infty$, la binomial negativa converge a la Poisson.

En *mgcv*, la familia se especifica como *family = nb()*, que estima θ simultáneamente con los demás parámetros del modelo.

7.7.4. GAM con distribución binomial

Para respuestas binarias (animal enfermo/sano, test positivo/negativo) o proporciones (proporción de cuartos mamarios infectados por vaca):

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + f_1(X_{1i})$$

donde $\pi_i = P(Y_i = 1)$.

Un ejemplo: la probabilidad de seropositividad a leptospirosis en caninos como función suave de la precipitación mensual acumulada en la zona. Un GLM binomial asumiría una relación lineal en la escala logit. Un GAM permite que la relación sea una curva, lo cual es biológicamente plausible: la seropositividad puede aumentar con la lluvia hasta cierto umbral y estabilizarse.

7.7.5. Selección de la familia

La Tabla 7.5 orienta la selección.

Tabla 41
Familias comunes para-GAM en investigación veterinaria

Variable respuesta	Ejemplo veterinario	Familia	Enlace
Continua, distribución simétrica	Peso al destete (kg)	Gaussiana	Identidad
Continua, distribución asimétrica positiva	Concentración de cortisol (ng/mL)	Gamma	Log
Conteo sin sobredispersión	Número de lesiones por animal	Poisson	Log
Conteo con sobredispersión	HPG (huevos por gramo)	Binomial negativa	Log
Binaria	Seropositividad (sí/no)	Binomial	Logit
Proporción	Proporción de cuartos infectados	Binomial (con pesos)	Logit

7.8. Diagnósticos del GAM: residuos, selección de k, concurvidad

7.8.1. Verificación de residuos

Como en cualquier modelo estadístico, los residuos del GAM deben inspeccionarse para detectar patrones residuales que indiquen mala especificación. La función *gam.check()* de *mgcv* produce cuatro gráficos diagnósticos:

1. **QQ-plot de residuos de devianza:** verifica si los residuos siguen la distribución esperada.
2. **Residuos vs. valores ajustados:** busca patrones sistemáticos (curvatura residual, embudo).
3. **Histograma de residuos:** complementa el QQ-plot.
4. **Respuesta observada vs. ajustada:** evalúa la calibración del modelo.

Para familias discretas (Poisson, binomial negativa), los residuos de devianza no se distribuyen normalmente, y los gráficos estándar pueden ser engañosos. El paquete *DHARMa* (Hartig, 2022) genera residuos simulados de cuantiles que sí permiten diagnósticos válidos para cualquier distribución.

7.8.2. Selección de k (dimensión de la base)

El parámetro *k* en *s(x, k = ...)* fija el número máximo de funciones base (y por tanto la complejidad máxima permitida). No es el número de grados de libertad del suavizado final ese lo controla la penalización, sino un techo.

Regla práctica: *k* debe ser suficientemente grande para que la penalización, no el truncamiento, controle la suavidad. Si los edf estimados están cerca de $k - 1$, el modelo puede estar limitado artificialmente. La función *gam.check()* incluye una prueba formal (basada en el patrón de los residuos) y emite una advertencia si *k* parece insuficiente.

Valores de referencia: - Para una variable con efecto suave simple (una curva con un máximo o mínimo): $k = 10$ suele ser suficiente. - Para efectos más complejos (estacionalidad): $k = 20$ o más. - Para interacciones bidimensionales (*te(x1, x2)*): el *k* total es el producto de las *k* marginales.

Si se pone $k = 100$ y el efecto real es lineal, el modelo simplemente estimará edf ≈ 1 y la curva será una recta. Es decir, *k* demasiado grande no daña el ajuste; solo aumenta ligeramente el costo computacional.

7.8.3. Concurvidad

La concurvidad es el análogo no paramétrico de la multicolinealidad. Ocurre cuando una función suave de un predictor puede ser aproximada por una combinación de las funciones suaves de los demás predictores. Alta concurvidad produce estimaciones inestables y errores estándar inflados.

En *mgcv*, la función *concurvity()* calcula tres índices (worst, observed, estimate) para cada término suave. Valores superiores a 0.8 en el índice “worst” sugieren problemas y requieren revisar si dos predictores son redundantes.

Un caso frecuente en datos veterinarios: incluir simultáneamente “días en leche” y “número de parto” como suavizados en un modelo de producción lechera. Si las vacas de primer parto tienden a tener menos días en leche que las multíparas (porque la finca las seca antes), ambas variables estarán correlacionadas y la concurvidad será alta.

7.9. Regresión cuantílica como alternativa para medianas y percentiles

7.9.1. Motivación

La regresión lineal clásica modela la media condicional: $E(Y|X) = \beta_0 + \beta_1 X$. Pero la media es un resumen incompleto cuando la distribución de *Y* dado *X* no es simétrica, o cuando interesa la cola inferior o superior de la distribución.

En nutrición animal, por ejemplo, la ganancia diaria de peso en un lote de novillos cebú en pastoreo tiene una distribución asimétrica a la derecha: la mayoría gana entre 400 y 600 g/día, pero algunos animales dominantes ganan más de 800 g/día. Un modelo que prediga la media no dice nada sobre los animales del percentil 10 (los que menos ganan) ni los del percentil 90.

7.9.2. Fundamento de la regresión cuantílica

La regresión cuantílica (Koenker & Bassett, 1978) estima el cuantil τ-ésimo de la distribución condicional de *Y* dado *X*:

$$Q_\tau(Y|X) = \beta_0(\tau) + \beta_1(\tau)X$$

donde $\tau \in (0,1)$ es el cuantil de interés. Para $\tau = 0.5$, se modela la mediana condicional. Para $\tau = 0.10$, el percentil 10.

Los coeficientes se estiman minimizando una función de pérdida asimétrica:

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i{}'\beta\right)$$

donde:

$$\rho_\tau(u) = u \cdot [\tau - I(u < 0)] = \begin{cases} \tau \cdot u & \text{si } u \geq 0 \\ (\tau - 1) \cdot u & \text{si } u < 0 \end{cases}$$

Esta función penaliza asimétricamente los residuos positivos y negativos. Para $\tau = 0.5$, la penalización es simétrica y se recupera la minimización de desviaciones absolutas (regresión de mínimas desviaciones absolutas, LAD), que es el análogo de la mediana en lugar de la media.

7.9.3. Interpretación de los coeficientes

En regresión cuantílica, $\beta_1(\tau)$ se interpreta como el cambio en el cuantil τ-ésimo de Y por cada unidad de cambio en X. Si $\beta_1(0.50) = 3.2$ y $\beta_1(0.90) = 5.1$ en un modelo de peso al destete en función de la ganancia diaria promedio, esto indica que el efecto de la ganancia sobre el peso al destete es mayor en los animales que ya son pesados (percentil 90) que en el animal mediano.

7.9.4. Ventajas sobre la regresión a la media

1. **Resistencia a valores extremos.** La regresión de la mediana (τ = 0.5) es mucho menos sensible a outliers que la regresión de la media.
2. **No requiere homocedasticidad.** Si la dispersión de Y cambia con X (heterocedasticidad), la regresión cuantílica modela esa variación directamente, porque los cuantiles extremos divergirán de la mediana.
3. **Visión completa de la distribución condicional.** Ajustar regresiones para τ = 0.10, 0.25, 0.50, 0.75 y 0.90 proporciona un mapa de cómo toda la distribución de Y cambia con X, no solo el centro.

7.9.5. Ejemplo veterinario: peso al destete en terneros Brahman

En un estudio de 85 terneros Brahman nacidos en fincas de la provincia de Los Ríos (Ecuador), se registró el peso al destete (kg) y el peso al nacimiento (kg). La regresión cuantílica a los cuantiles 0.10, 0.50 y 0.90 revela que el efecto del peso al nacimiento sobre el peso al destete es heterogéneo: los terneros livianos al nacimiento tienen mayor variabilidad en su peso al destete que los pesados, y la pendiente del percentil 10 es menor que la del percentil 90. En términos prácticos, un ternero que nace liviano puede terminar en cualquier rango de peso al destete, mientras que uno que nace pesado tiende a mantenerse pesado. Este hallazgo tiene implicaciones para programas de selección genética.

7.10. Implementación en R

7.10.1. Paquetes requeridos

```
# Paquetes necesarios para este capítulo
# Instalar con install.packages() si no están disponibles

library(vegan)      # PERMANOVA (adonis2), ANOSIM, NMDS (metaMDS), distancias
```

```r
library(mgcv)       # GAM con gam(), funciones suaves, familias extendidas
library(gratia)     # Gráficos diagnósticos mejorados para GAM
library(quantreg)   # Regresión cuantílica con rq()
library(ggplot2)    # Gráficos de alta calidad
library(dplyr)      # Manipulación de datos
library(DHARMa)     # Diagnósticos para modelos con familias no normales
```

7.10.2. Matrices de distancia en R

```r
# ---- Datos: comunidad de ectoparásitos en bovinos ----

ectoparasitos <- data.frame(
  Animal  = 1:15,
  Sistema = factor(rep(c("Extensivo", "Semi-intensivo", "Intensivo"), each = 5)),
  Rhipicephalus = c(2, 1, 0, 3, 1,  10, 14, 12, 9, 11,  25, 30, 22, 28, 26),
  Amblyomma     = c(18, 22, 15, 20, 17,  8, 5, 6, 9, 7,  1, 0, 2, 0, 1),
  Dermacentor   = c(5, 3, 7, 4, 6,  2, 1, 3, 2, 1,  0, 1, 0, 0, 1),
  Ixodes        = c(8, 12, 6, 10, 9,  3, 2, 1, 4, 2,  0, 0, 1, 0, 0)
)

# Matriz de abundancias (sin la columna de Animal ni de Sistema)
abundancias <- ectoparasitos[, 3:6]

# Distancia euclidiana
dist_eucl <- dist(abundancias, method = "euclidean")

# Disimilitud de Bray-Curtis (función vegdist del paquete vegan)
dist_bc <- vegdist(abundancias, method = "bray")

# Disimilitud de Jaccard sobre presencia-ausencia
# decostand convierte abundancias a presencia/ausencia con method = "pa"
dist_jacc <- vegdist(decostand(abundancias, method = "pa"), method = "jaccard")

# Visualizar la matriz de Bray-Curtis como tabla
as.matrix(dist_bc)[1:5, 1:5]
```

Nota sobre vegdist. La función *vegdist()* del paquete *vegan* calcula más de 15 medidas de distancia y disimilitud. El argumento *method* acepta, entre otros: *"bray"* (Bray-Curtis), *"jaccard"*, *"euclidean"*, *"manhattan"*, *"gower"*, *"morisita"*, *"horn"*. La ayuda (*?vegdist*) detalla cada una. Bray-Curtis es la opción predeterminada en muchas funciones de *vegan*.

7.10.3. PERMANOVA con adonis2

```r
# ---- PERMANOVA ----
# adonis2() es la versión actualizada de adonis() (obsoleta)
```

```r
# Usa la interfaz de fórmula: matriz de distancias ~ factor

set.seed(42)  # Para reproducibilidad de las permutaciones

resultado_permanova <- adonis2(
  dist_bc ~ Sistema,          # fórmula: distancia ~ grupo
  data = ectoparasitos,       # data.frame con las variables del lado derecho
  permutations = 9999,        # número de permutaciones
  method = "bray"             # redundante aquí porque ya pasamos dist_bc,
                              # pero explícito por claridad
)

print(resultado_permanova)
```

Salida esperada e interpretación:

```
Permutation test for adonis under reduced model
Terms added sequentially (first to last)
Permutation: free
Number of permutations: 9999
adonis2(formula = dist_bc ~ Sistema, data = ectoparasitos, permutations = 9999)
         Df SumOfSqs      R2      F Pr(>F)
Sistema   2   2.6841 0.93524 86.557  1e-04 ***
Residual 12   0.1861 0.06476
Total    14   2.8702 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lectura de la salida:

- **Df:** grados de libertad del factor Sistema (a − 1 = 2) y del residuo (n − a = 12).
- **SumOfSqs:** suma de cuadrados entre grupos (2.6841) y dentro de grupos (0.1861).
- **R^2:** proporción de la variabilidad total explicada por el factor. Un $R^2 = 0.935$ indica que el 93.5% de la variación en composición de ectoparásitos se asocia al sistema productivo. Es un valor excepcionalmente alto, esperable con datos diseñados para ilustrar diferencias claras. En estudios reales, R^2 entre 0.10 y 0.30 ya son informativos.
- **F:** pseudo-F de Anderson = 86.56.
- **Pr(>F):** valor *p* basado en 9999 permutaciones. Un *p* = 0.0001 (el mínimo posible con 9999 permutaciones) indica rechazo contundente de H_0.

```r
# ---- PERMDISP: verificación de homogeneidad de dispersiones ----
```

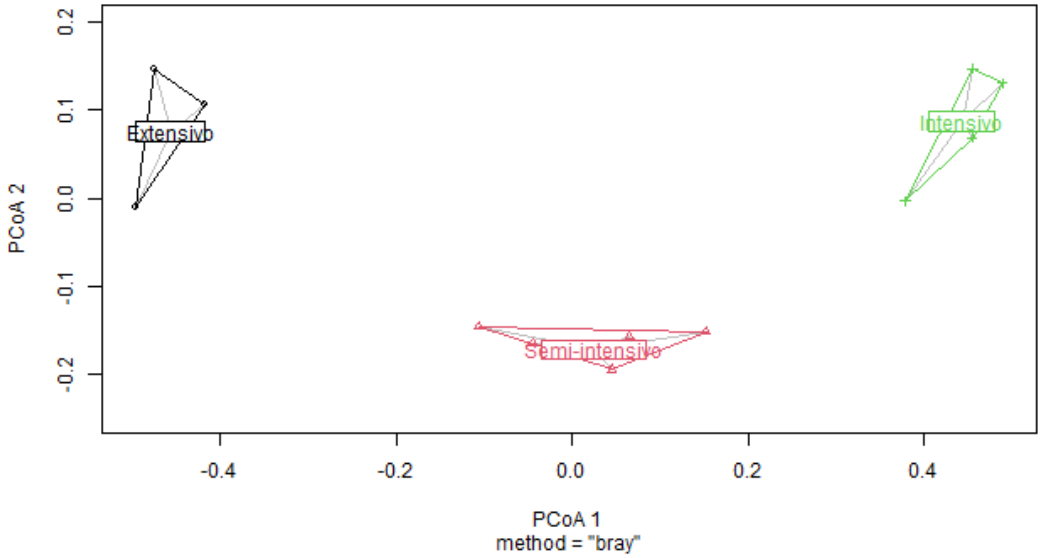
```
# Primero, calcular las distancias al centroide de cada grupo
dispersion <- betadisper(dist_bc, ectoparasitos$Sistema)

# Prueba de homogeneidad de dispersiones
permutest(dispersion, pairwise = TRUE, permutations = 9999)

# Gráfico de dispersiones
plot(dispersion, main = "PERMDISP: Distancias al centroide por sistema")
```

Figura 28

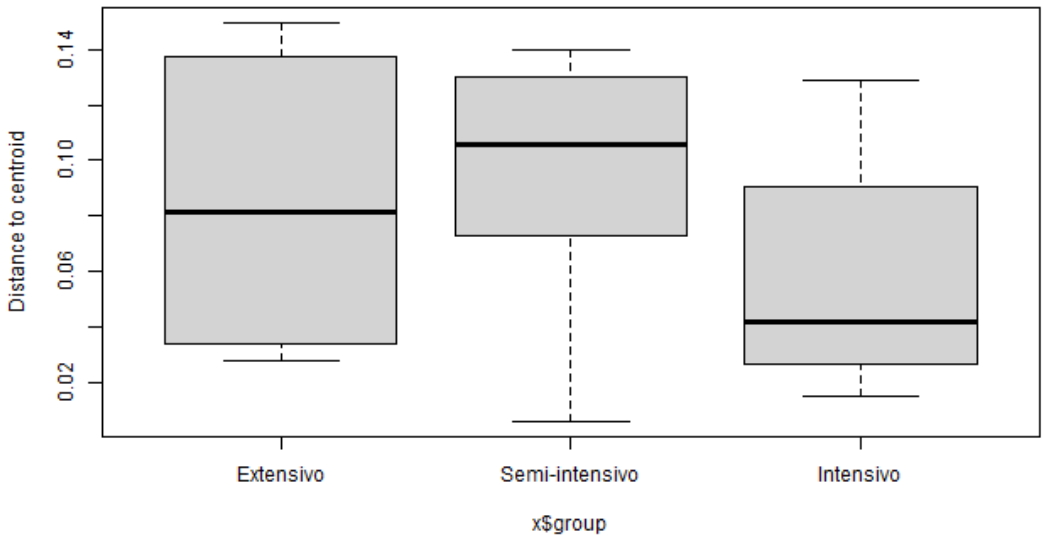
PERMDISP: Distancias al centroide por sistema




```
boxplot(dispersion, main = "Distancias al centroide")
```

Figura 29

Distancias al centroide



Si PERMDISP no es significativo ($p > 0.05$), la diferencia detectada por PERMANOVA se atribuye a diferencias en composición (localización de los centroides), no a diferencias en variabilidad. Si PERMDISP es significativo, la interpretación debe ser cautelosa.

7.10.4. ANOSIM

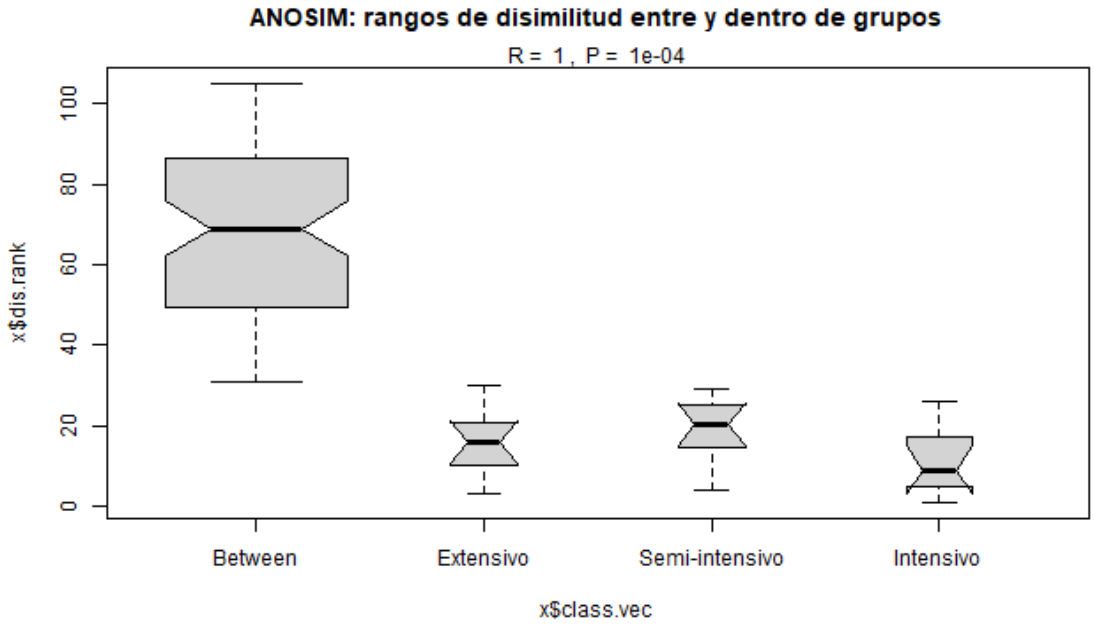
```
# ---- ANOSIM ----

set.seed(42)
resultado_anosim <- anosim(dist_bc, ectoparasitos$Sistema, permutations = 9999)

print(resultado_anosim)
# Salida clave:
# ANOSIM statistic R: 0.981 (o similar)
# Significance: 0.001

# Gráfico de rangos de disimilitud
plot(resultado_anosim, main = "ANOSIM: Rangos de disimilitud entre y dentro de grupos")
```

Figura 30
ANOSIM: Rangos de disimilitud entre y dentro de grupos



Interpretación. Un R cercano a 1 confirma que las distancias entre sistemas son sistemáticamente mayores que las distancias dentro de cada sistema. Combinado con el resultado de PERMANOVA, la evidencia de diferencia composicional entre sistemas productivos es contundente.

7.10.5. NMDS

```r
# ---- NMDS con metaMDS ----
# metaMDS() ejecuta múltiples inicios aleatorios y selecciona la mejor solución

set.seed(42)
nmds_resultado <- metaMDS(
  abundancias,          # matriz de abundancias (no la matriz de distancia)
  distance = "bray",    # medida de disimilitud
  k = 2,                # número de dimensiones
  trymax = 100,         # número máximo de inicios aleatorios
  trace = FALSE         # suprime los mensajes de iteración en consola
)

# Valor de estrés
nmds_resultado$stress
# Un estrés < 0.10 indica muy buena representación

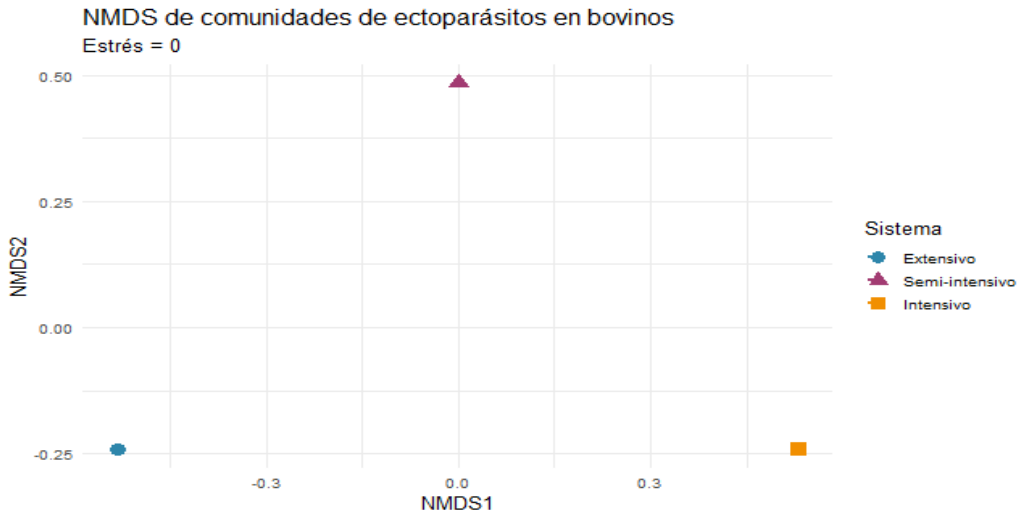
# ---- Gráfico NMDS con ggplot2 ----

# Extraer las coordenadas de los sitios (animales)
scores_nmds <- as.data.frame(scores(nmds_resultado, display = "sites"))
scores_nmds$Sistema <- ectoparasitos$Sistema

# Gráfico
ggplot(scores_nmds, aes(x = NMDS1, y = NMDS2, colour = Sistema, shape = Sistema)) +
  geom_point(size = 4) +
  stat_ellipse(level = 0.95, linetype = 2) +
  labs(
    title = "NMDS de comunidades de ectoparásitos en bovinos",
    subtitle = paste("Estrés =", round(nmds_resultado$stress, 3)),
    x = "NMDS1",
    y = "NMDS2"
  ) +
  theme_minimal(base_size = 13) +
  scale_colour_manual(values = c("Extensivo" = "#2E86AB",
                                 "Semi-intensivo" = "#A23B72",
                                 "Intensivo" = "#F18F01"))
```

Figura 31

NMDS de comunidades de ectoparásitos en bovinos



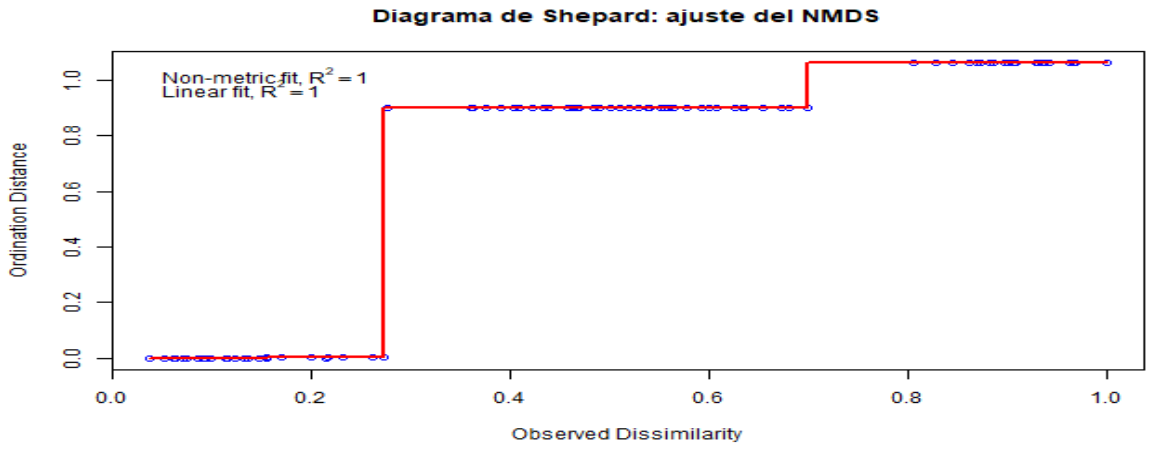
Nota. El gráfico NMDS muestra tres agrupaciones claramente separadas. Los bovinos del sistema extensivo se concentran en el cuadrante superior izquierdo (dominados por *Amblyomma* e *Ixodes*), los del intensivo en el cuadrante inferior derecho (*Rhipicephalus* dominante), y los del semi-intensivo en una posición intermedia. Las elipses de confianza al 95% no se superponen, lo que refuerza visualmente el resultado de PERMANOVA.

```
# ---- Diagrama de Shepard (verificación de ajuste del NMDS) ----

stressplot(nmds_resultado,
           main = "Diagrama de Shepard: ajuste del NMDS")
# Los puntos deben seguir la línea escalonada (ajuste monótono)
# Una dispersión alta de puntos respecto a la línea indica mal ajuste
```

Figura 32

Diagrama de Shepard: ajuste del NMDS



7.10.6. LOESS en R

```r
# ---- Ejemplo: curva de lactancia suavizada con LOESS ----

# Datos simulados de 60 vacas Holstein, controles mensuales
set.seed(123)
n_vacas <- 60
lactancia <- data.frame(
  DEL = rep(seq(15, 300, by = 30), times = n_vacas / 10),
  Vaca = rep(1:(n_vacas / 10), each = 10)
)
# Simular producción con la forma de Wood + ruido
lactancia$Produccion <- with(lactancia,
  18 * (DEL / 30)^0.3 * exp(-0.004 * DEL) + rnorm(nrow(lactancia), 0, 1.5)
)

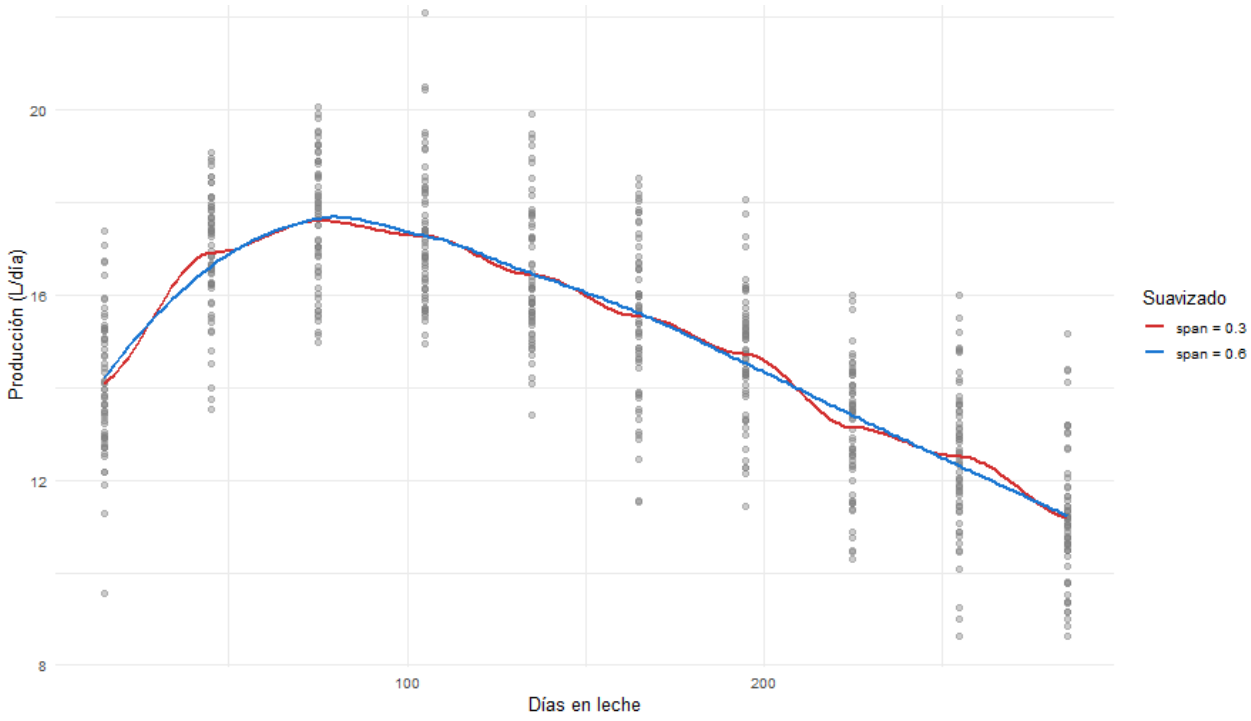
# Ajustar LOESS con distintos valores de span
loess_03 <- loess(Produccion ~ DEL, data = lactancia, span = 0.3)
loess_06 <- loess(Produccion ~ DEL, data = lactancia, span = 0.6)

# Predicciones
x_pred <- data.frame(DEL = seq(15, 300, length.out = 200))
x_pred$Pred_03 <- predict(loess_03, newdata = x_pred)
x_pred$Pred_06 <- predict(loess_06, newdata = x_pred)

# Gráfico comparativo
ggplot(lactancia, aes(x = DEL, y = Produccion)) +
  geom_point(alpha = 0.4, colour = "grey50") +
  geom_line(data = x_pred, aes(y = Pred_03, colour = "span = 0.3"), linewidth = 1) +
  geom_line(data = x_pred, aes(y = Pred_06, colour = "span = 0.6"), linewidth = 1) +
  scale_colour_manual(values = c("span = 0.3" = "#D32F2F", "span = 0.6" = "#1976D2"),
                      name = "Suavizado") +
  labs(
    title = "Curva de lactancia suavizada con LOESS",
    x = "Días en leche",
    y = "Producción (L/día)"
  ) +
  theme_minimal(base_size = 13)
```

Figura 33

Curva de lactancia suavizada con LOESS



Nota. La curva con span = 0.3 (roja) sigue más de cerca las oscilaciones locales de los datos. La curva con span = 0.6 (azul) es más suave y captura la tendencia general de ascenso, pico y descenso. Ambas muestran el pico de lactancia alrededor del día 60, seguido del descenso gradual típico de la curva de lactancia en Holstein.

7.10.7. GAM con mgcv

```
# ---- Ejemplo 1: GAM gaussiano para producción lechera ----

# Ajustar un GAM con suavizado de DEL
gam_lactancia <- gam(
  Produccion ~ s(DEL, k = 15, bs = "tp"),  # s() define la función suave
  data = lactancia,                        # tp = thin plate spline
  method = "REML"                          # método de estimación recomendado
)

# Resumen del modelo
summary(gam_lactancia)
```

Salida esperada e interpretación:

```
Family: gaussian
Link function: identity
```

```
Formula:
Produccion ~ s(DEL, k = 15, bs = "tp")

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.532      0.198   88.55   <2e-16 ***

Approximate significance of smooth terms:
        edf Ref.df     F p-value
s(DEL) 3.872  4.789 55.14  <2e-16 ***

R-sq.(adj) =  0.789   Deviance explained = 79.7%
-REML = 192.3  Scale est. = 2.351  n = 60
```

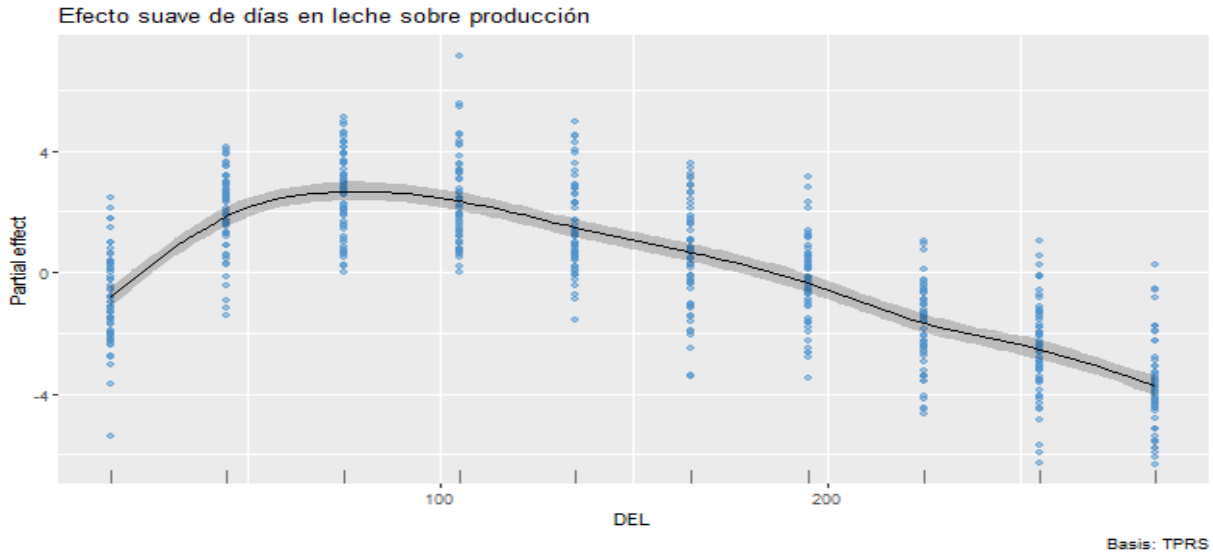
Lectura línea por línea:

- **Family: gaussian / Link: identity:** la distribución de la respuesta es normal y la función de enlace es la identidad. Apropiado para producción lechera si los residuos son razonablemente simétricos.
- **Intercept = 17.53:** el valor medio global de producción. En un GAM, el intercepto tiene la misma interpretación que en un modelo lineal.
- **s(DEL): edf = 3.87:** la función suave de días en leche tiene 3.87 grados de libertad efectivos. Esto indica una curva con complejidad moderada (más que una recta, menos que una función muy ondulada). Es consistente con la forma de campana de la curva de lactancia.
- **F = 55.14, p < 2e-16:** el efecto suave de DEL es altamente significativo. El suavizado reduce sustancialmente la devianza del modelo.
- **R^2 ajustado = 0.789:** el modelo explica el 78.9% de la variabilidad en producción.

```
# Gráfico del efecto suave con intervalo de confianza
# Usando el paquete gratia para gráficos mejorados
library(gratia)
draw(gam_lactancia, residuals = TRUE) +
  labs(title = "Efecto suave de días en leche sobre producción")
# Alternativamente, con plot() base
plot(gam_lactancia, shade = TRUE, shade.col = "lightblue",
     main = "Efecto suave de DEL", xlab = "Días en leche",
     ylab = "Efecto parcial sobre producción (L/día)")
```

Figura 34

Efecto suave de días en leche sobre producción



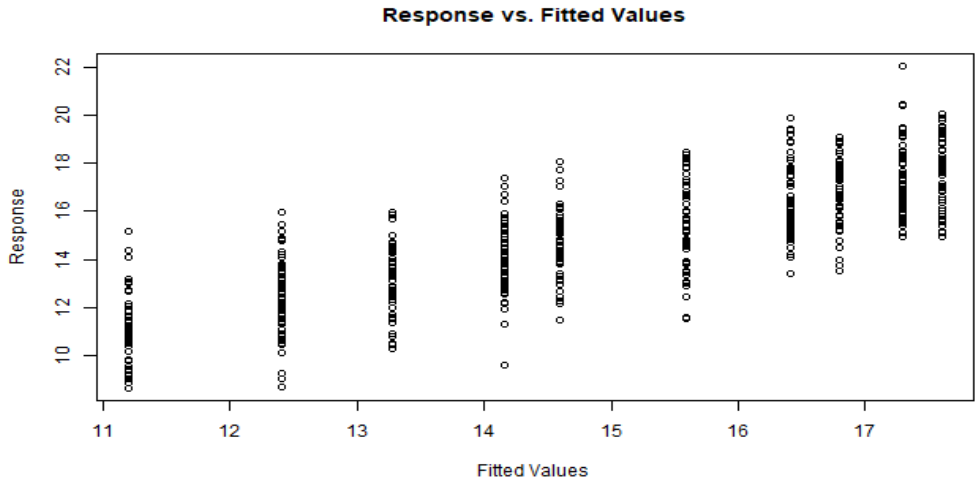
Nota. El efecto suave de la muestra una curva que asciende hasta aproximadamente el día 60, alcanza un máximo y desciende gradualmente hasta el final de la lactancia. La banda sombreada es el intervalo de confianza al 95%. La banda es más ancha en los extremos del rango de DEL, donde hay menos datos. Los puntos son residuos parciales que permiten evaluar visualmente el ajuste.

```
# ---- Diagnósticos del GAM ----

# Diagnósticos básicos con gam.check()
gam.check(gam_lactancia)
```

Figura 35

gam_lactancia



```r
# Produce 4 gráficos: QQ-plot, residuos vs. ajustados, histograma, obs vs. pred
# También prueba si k es suficiente

# Concurvidad (solo relevante con múltiples suavizados)
# Ejemplo con modelo ampliado
# gam_multi <- gam(Produccion ~ s(DEL) + s(NumParto), data = lactancia, method = "REML"
)
# concurvity(gam_multi, full = TRUE)

# ---- Ejemplo 2: GAM con binomial negativa para conteo de células somáticas ----

# Datos simulados: RCS en miles/mL como función de DEL y época
set.seed(456)
n_obs <- 200
datos_rcs <- data.frame(
  DEL   = sample(5:350, n_obs, replace = TRUE),
  Epoca = factor(sample(c("Seca", "Lluviosa"), n_obs, replace = TRUE))
)

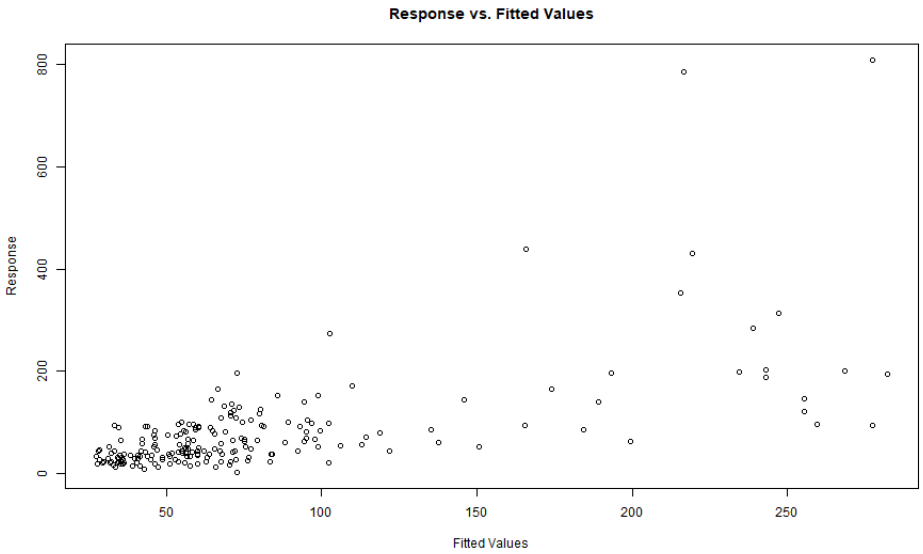
# Simular RCS con sobredispersión
# El RCS tiende a ser alto al inicio, baja, y sube al final
mu <- with(datos_rcs,
  exp(4.5 + 0.8 * (DEL < 30) - 0.003 * DEL + 0.002 * (DEL/100)^2 +
      0.3 * (Epoca == "Lluviosa"))
)
datos_rcs$RCS <- rnbinom(n_obs, mu = mu, size = 3)

# Ajustar GAM con binomial negativa
gam_rcs <- gam(
  RCS ~ s(DEL, k = 20, bs = "tp") + Epoca,
  family = nb(),          # binomial negativa, theta estimado automáticamente
  data = datos_rcs,
  method = "REML"
)

summary(gam_rcs)
# Verificar si k es suficiente
gam.check(gam_rcs)
```

Figura 36

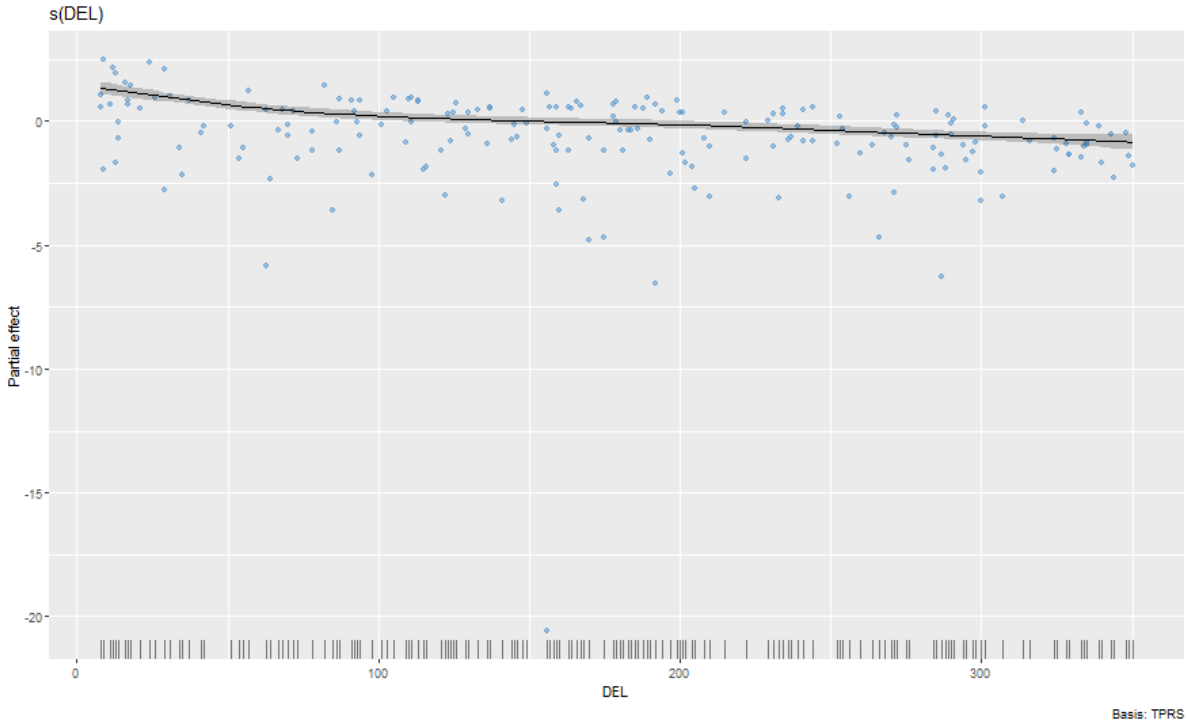
Figura diagnostica




```
# Gráfico del efecto suave de DEL
draw(gam_rcs, select = 1, residuals = TRUE)
```

Figura 37

Gráfico del efecto suave de DEL




```
# Efecto de la época (coeficiente paramétrico)
# En escala log: coef positivo para "Lluviosa" indica mayor RCS en época lluviosa
```

```r
# ---- Ejemplo 3: GAM con spline cíclico para efecto estacional ----

# Si el mes del año es un predictor, usar un spline cíclico
# para que enero y diciembre se conecten suavemente

# Agregar mes al dataset
datos_rcs$Mes <- sample(1:12, n_obs, replace = TRUE)

gam_estacional <- gam(
  RCS ~ s(DEL, k = 15) + s(Mes, bs = "cc", k = 12),  # cc = spline cíclico
  family = nb(),
  data = datos_rcs,
  method = "REML",
  knots = list(Mes = c(0.5, 12.5))  # límites del ciclo
)

summary(gam_estacional)
draw(gam_estacional)
```

7.10.8. Regresión cuantílica con quantreg

```r
# ---- Regresión cuantílica: peso al destete en terneros Brahman ----

set.seed(789)
n_terneros <- 85
brahman <- data.frame(
  PesoNac = round(rnorm(n_terneros, mean = 30, sd = 4), 1)
)
# Simular peso al destete con heterocedasticidad:
# la variabilidad del destete aumenta con menor peso al nacimiento
brahman$PesoDest <- with(brahman,
  80 + 4.5 * PesoNac + rnorm(n_terneros, 0, sd = 25 - 0.5 * PesoNac)
)

# Regresión cuantílica a los cuantiles 0.10, 0.25, 0.50, 0.75, 0.90
library(quantreg)

taus <- c(0.10, 0.25, 0.50, 0.75, 0.90)
modelos_rq <- rq(PesoDest ~ PesoNac, data = brahman, tau = taus)

summary(modelos_rq)
```

Interpretación de la salida. Para cada cuantil, se obtienen los coeficientes $\beta_0(\tau)$ y $\beta_1(\tau)$. Si $\beta_1(0.10) = 3.0$ y $\beta_1(0.90) = 6.2$, la pendiente es mayor para los terneros más pesados al destete

(percentil 90) que para los más livianos (percentil 10). Esto confirma la heterocedasticidad: el peso al nacimiento predice mejor el peso al destete en los animales pesados que en los livianos.

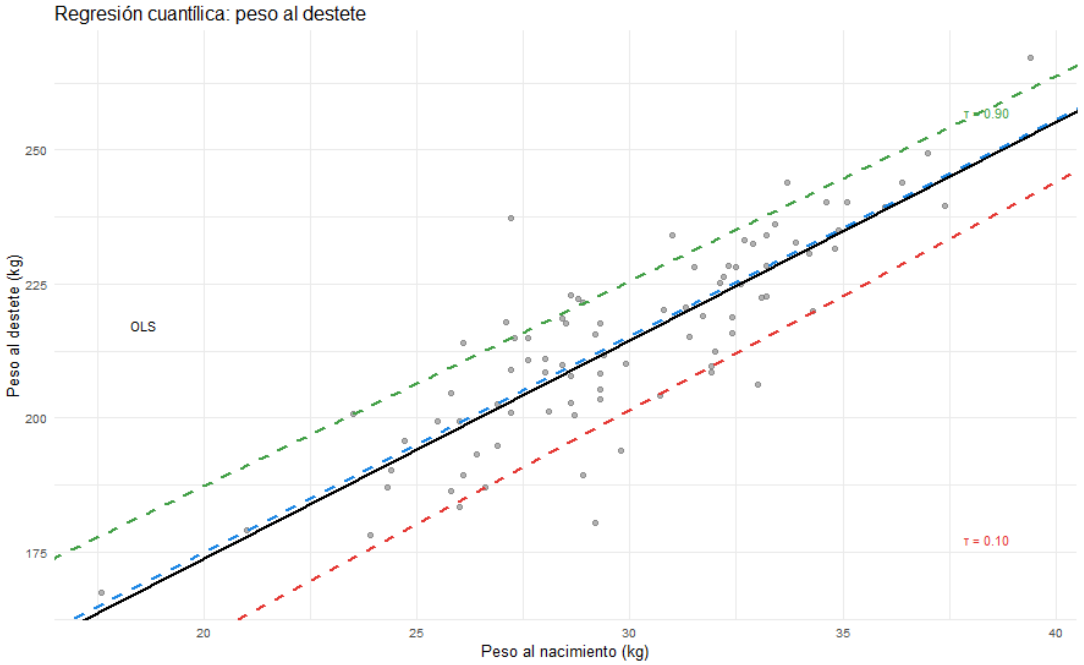
```
# Comparar con regresión lineal a la media
modelo_lm <- lm(PesoDest ~ PesoNac, data = brahman)

# ---- Gráfico con las rectas de cada cuantil ----

ggplot(brahman, aes(x = PesoNac, y = PesoDest)) +
  geom_point(alpha = 0.5, colour = "grey40") +
  # Recta de regresión a la media (OLS)
  geom_abline(intercept = coef(modelo_lm)[1], slope = coef(modelo_lm)[2],
              colour = "black", linewidth = 1, linetype = 1) +
  # Rectas de regresión cuantílica
  geom_abline(intercept = coef(modelos_rq)[1, 1], slope = coef(modelos_rq)[2, 1],
              colour = "#E53935", linewidth = 0.8, linetype = 2) +
  geom_abline(intercept = coef(modelos_rq)[1, 3], slope = coef(modelos_rq)[2, 3],
              colour = "#1E88E5", linewidth = 0.8, linetype = 2) +
  geom_abline(intercept = coef(modelos_rq)[1, 5], slope = coef(modelos_rq)[2, 5],
              colour = "#43A047", linewidth = 0.8, linetype = 2) +
  annotate("text", x = 38, y = 250, label = "τ = 0.90", colour = "#43A047", size = 4) +
  annotate("text", x = 38, y = 210, label = "τ = 0.50", colour = "#1E88E5", size = 4) +
  annotate("text", x = 38, y = 170, label = "τ = 0.10", colour = "#E53935", size = 4) +
  annotate("text", x = 38, y = 230, label = "OLS (media)", colour = "black", size = 4) +
  labs(
    title = "Regresión cuantílica: peso al destete en terneros Brahman",
    x = "Peso al nacimiento (kg)",
    y = "Peso al destete (kg)"
  ) +
  theme_minimal(base_size = 13)
```

Figura 38

Regresión cuantílica: peso al destete



Nota. El gráfico muestra la nube de puntos (peso al nacimiento vs. peso al destete) con cuatro rectas superpuestas: la regresión OLS a la media (línea continua negra) y las regresiones cuantílicas a los percentiles 10 (roja), 50 (azul) y 90 (verde), en líneas discontinuas. La divergencia entre las rectas de los cuantiles extremos evidencia la heterocedasticidad: la "apertura del abanico" es mayor a la izquierda (pesos al nacimiento bajos). La recta OLS pasa por el centro de la nube, pero no captura esta variación diferencial.

7.11. Estudio de caso integrador

Caso A: Diversidad de ectoparásitos en bovinos según sistema productivo

Contexto. Un equipo de investigación del programa de Medicina Veterinaria de la Universidad Estatal de Bolívar estudia la influencia del sistema de producción bovina sobre la estructura de la comunidad de ectoparásitos en la provincia de Bolívar, Ecuador. Se muestrean 45 bovinos mestizos (cruce Holstein × criollo) de 15 fincas (cinco fincas por sistema: extensivo de páramo > 3200 m s.n.m., semi-intensivo de valle a 2600 m, e intensivo periurbano a 2400 m). De cada animal se recolectan ectoparásitos por inspección directa durante 15 minutos y se identifican hasta género.

Datos. La Tabla 7.6 muestra los datos de 15 animales representativos (tres por sistema, una finca por sistema para simplificar el ejemplo). Los datos completos de los 45 animales se usarían en un análisis real; aquí se trabaja con el subconjunto para que los cálculos sean seguibles a mano.

Tabla 42

Conteos de ectoparásitos (géneros de garrapatas y otros artrópodos) en bovinos de tres sistemas productivos

ID	Sistema	*Rhipicephalus*	*Amblyomma*	*Dermacentor*	*Ixodes*	*Haematobia*	*Melophagus*
E1	Extensivo	0	15	8	12	0	3
E2	Extensivo	1	20	5	9	1	5
E3	Extensivo	0	18	6	14	0	4
E4	Extensivo	2	12	9	7	2	2
E5	Extensivo	0	22	4	11	0	6
S1	Semi-int.	8	6	3	2	5	0
S2	Semi-int.	12	4	1	3	8	0
S3	Semi-int.	10	7	2	1	6	1
S4	Semi-int.	9	5	4	2	7	0
S5	Semi-int.	11	3	2	4	4	0
I1	Intensivo	28	0	0	0	12	0
I2	Intensivo	32	1	0	0	15	0
I3	Intensivo	25	0	1	0	10	0
I4	Intensivo	30	0	0	0	14	0
I5	Intensivo	27	1	0	0	11	0

Justificación del enfoque no paramétrico. Los datos son conteos multivariados con abundantes ceros (especialmente en el sistema intensivo para *Dermacentor*, *Ixodes* y *Melophagus*, y en el extensivo para *Haematobia*). La normalidad multivariada no se cumple. Las varianzas difieren entre sistemas porque los perfiles son cualitativamente distintos. PERMANOVA con Bray-Curtis es la herramienta adecuada.

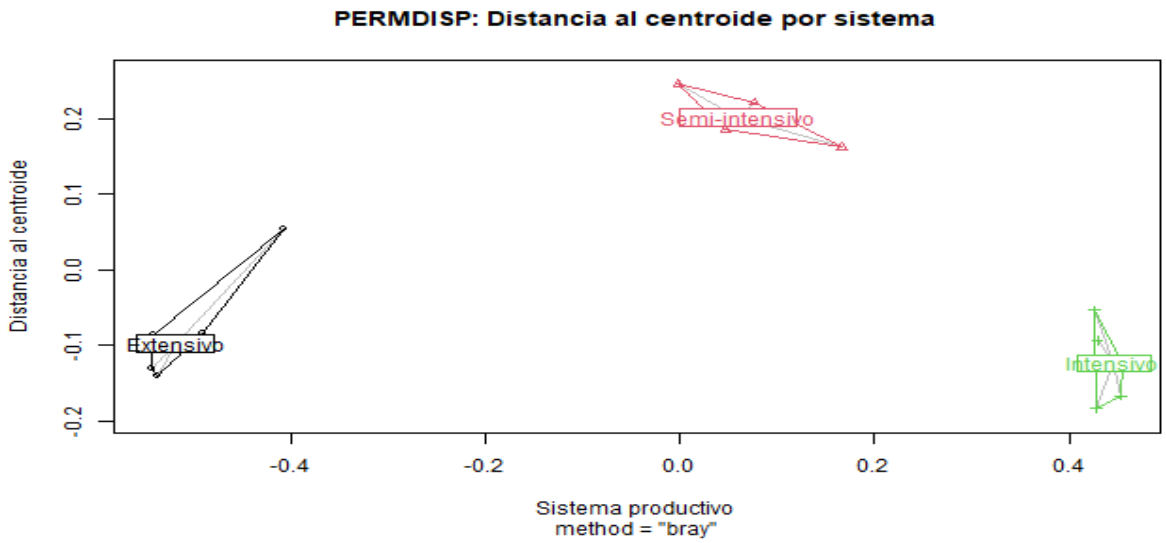
Análisis completo en R:

```r
# ---- Estudio de caso A: ectoparásitos en bovinos ----

library(vegan)
library(ggplot2)
# Ingresar datos
caso_ecto <- data.frame(
  ID = c(paste0("E", 1:5), paste0("S", 1:5), paste0("I", 1:5)),
  Sistema = factor(rep(c("Extensivo", "Semi-intensivo", "Intensivo"), each = 5),
                   levels = c("Extensivo", "Semi-intensivo", "Intensivo")),
  Rhipicephalus = c(0,1,0,2,0, 8,12,10,9,11, 28,32,25,30,27),
  Amblyomma = c(15,20,18,12,22, 6,4,7,5,3, 0,1,0,0,1),
  Dermacentor = c(8,5,6,9,4, 3,1,2,4,2, 0,0,1,0,0),
  Ixodes = c(12,9,14,7,11, 2,3,1,2,4, 0,0,0,0,0),
  Haematobia = c(0,1,0,2,0, 5,8,6,7,4, 12,15,10,14,11),
  Melophagus = c(3,5,4,2,6, 0,0,1,0,0, 0,0,0,0,0)
)
abundancias_caso <- caso_ecto[, 3:8]
# 1. Matriz de Bray-Curtis
bc_caso <- vegdist(abundancias_caso, method = "bray")
# 2. PERMANOVA
set.seed(2024)
perm_caso <- adonis2(bc_caso ~ Sistema, data = caso_ecto, permutations = 9999)
print(perm_caso)
# 3. PERMDISP
disp_caso <- betadisper(bc_caso, caso_ecto$Sistema)
permutest(disp_caso, permutations = 9999)
```

Figura 39

PERMDISP: Distancia al centroide por sistema

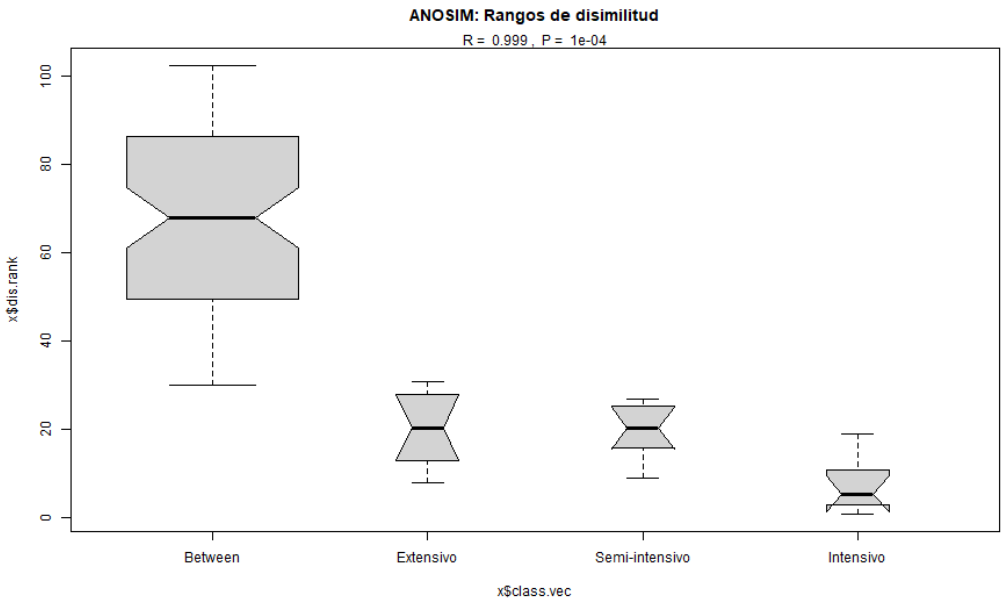


4. ANOSIM

```
anosim_caso <- anosim(bc_caso, caso_ecto$Sistema, permutations = 9999)
print(anosim_caso)
```

Figura 40

ANOSIM: Rangos de disimilitud

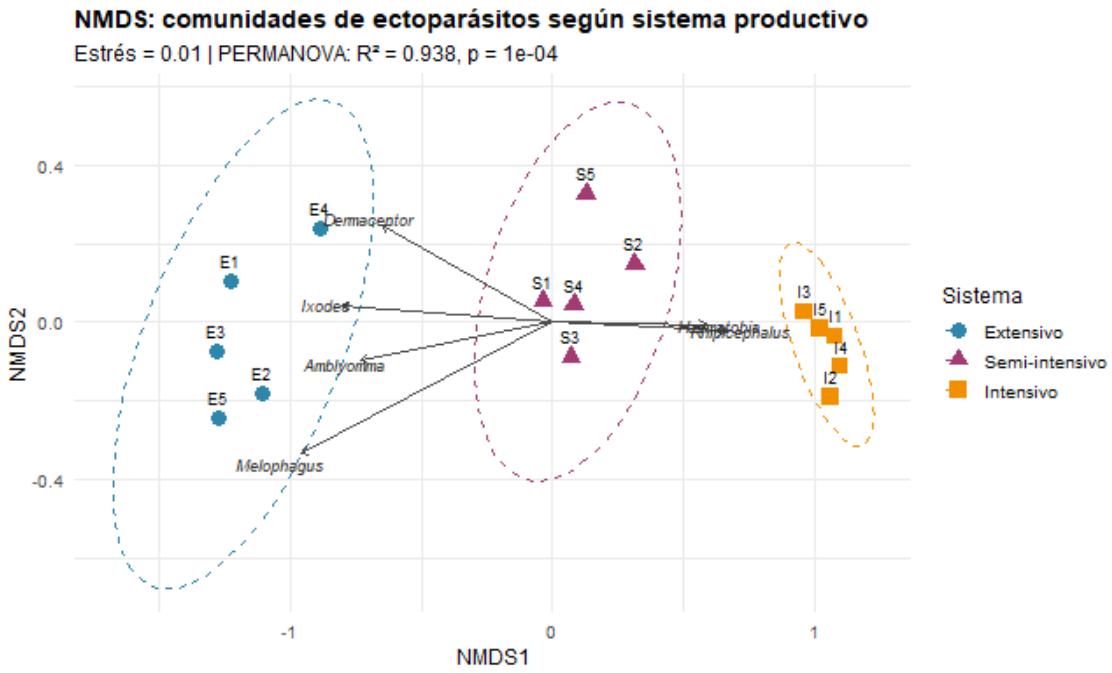


5. NMDS

```r
set.seed(2024)
nmds_caso <- metaMDS(abundancias_caso, distance = "bray", k = 2, trymax = 100,
                     trace = FALSE)
cat("Estrés:", nmds_caso$stress, "\n")
# 6. Gráfico NMDS
scores_caso <- as.data.frame(scores(nmds_caso, display = "sites"))
scores_caso$Sistema <- caso_ecto$Sistema
# Agregar las especies como flechas (biplot)
spp_scores <- as.data.frame(scores(nmds_caso, display = "species"))
spp_scores$Especie <- rownames(spp_scores)
ggplot() +
  geom_point(data = scores_caso, aes(x = NMDS1, y = NMDS2,
             colour = Sistema, shape = Sistema), size = 4) +
  stat_ellipse(data = scores_caso, aes(x = NMDS1, y = NMDS2,
               colour = Sistema), level = 0.95, linetype = 2) +
  geom_segment(data = spp_scores, aes(x = 0, y = 0, xend = NMDS1, yend = NMDS2),
               arrow = arrow(length = unit(0.2, "cm")), colour = "grey30") +
  geom_text(data = spp_scores, aes(x = NMDS1 * 1.15, y = NMDS2 * 1.15,
            label = Especie), size = 3, colour = "grey30", fontface = "italic") +
  labs(
    title = "NMDS: Comunidades de ectoparásitos según sistema productivo",
    subtitle = paste("Estrés =", round(nmds_caso$stress, 3),
                     "| PERMANOVA p =", format(perm_caso$`Pr(>F)`[1], digits = 4)),
    x = "NMDS1", y = "NMDS2"
  ) +
  theme_minimal(base_size = 12) +
  scale_colour_manual(values = c("#2E86AB", "#A23B72", "#F18F01"))
```

Figura 41

NMDS: comunidades de ectoparásitos según sistema productivo



Interpretación en términos veterinarios. La composición de la comunidad de ectoparásitos difiere entre los tres sistemas productivos (PERMANOVA: pseudo-F elevado, $p < 0.001$, $R^2 > 0.80$). El NMDS revela una estructura interpretable biológicamente:

- El sistema **extensivo de páramo** se asocia a *Amblyomma*, *Ixodes* y *Melophagus*. Esto es coherente con la ecología de estos géneros: *Ixodes* tolera altitudes elevadas y clima frío; *Amblyomma* es generalista; *Melophagus ovinus* puede encontrarse donde hay contacto con ovinos en sistemas mixtos de páramo.
- El sistema **intensivo periurbano** muestra dominancia casi exclusiva de *Rhipicephalus* (probablemente *R. microplus*) y *Haematobia* (mosca de los cuernos). Ambos son ectoparásitos de climas cálidos, adaptados a bovinos estabulados con alta carga animal.
- El sistema **semi-intensivo** es intermedio, con representación moderada de todos los géneros excepto *Melophagus*.

Estos hallazgos tienen implicaciones directas para el diseño de programas de control integrado de ectoparásitos: un protocolo basado exclusivamente en acaricidas contra *Rhipicephalus* sería inadecuado para el sistema extensivo de páramo, donde los géneros dominantes son otros.

¿Qué habría pasado con un enfoque paramétrico? Un MANOVA sobre los conteos crudos violaría normalidad multivariada y homogeneidad de matrices de covarianza. Aún con transformación logarítmica (log(x + 1)), los ceros abundantes generarían distribuciones fuertemente leptocúrticas. Un ANOVA univariado sobre la abundancia total perdería la información de composición: dos animales con 30 ectoparásitos totales, pero perfiles completamente distintos aparecerían como iguales.

Caso B: GAM para recuento de células somáticas en función de días en leche y época del año

Contexto. En un hato de 120 vacas Holstein con mastitis subclínica de una hacienda lechera en Machachi (provincia de Pichincha, Ecuador, 3100 m s.n.m.), se registra mensualmente el recuento de células somáticas (RCS, en miles de células/mL) junto con los días en leche (DEL) y la época del año (seca: junio-noviembre; lluviosa: diciembre-mayo). Se busca modelar la relación entre RCS, DEL y época para identificar los períodos de mayor riesgo de elevación del RCS.

Datos. Se dispone de 480 observaciones (120 vacas × 4 controles). Los datos se simulan con un patrón biológicamente realista.

```
# ---- Estudio de caso B: GAM para RCS ----

library(mgcv)
library(gratia)
library(ggplot2)
library(DHARMa)

set.seed(2024)
n <- 480

datos_mastitis <- data.frame(
  Vaca  = rep(1:120, each = 4),
  DEL   = sample(10:340, n, replace = TRUE),
  Epoca = factor(sample(c("Seca", "Lluviosa"), n, replace = TRUE,
                        prob = c(0.5, 0.5)))
)

# Generar RCS con patrón no lineal en DEL y efecto de época
# RCS alto al inicio, baja, y sube al final de la lactancia
# Época lluviosa: mayor RCS (más barro, más infecciones ambientales)
mu_rcs <- with(datos_mastitis,
  exp(5.0
```

```
    + 1.2 * exp(-0.03 * DEL)        # efecto alto al inicio, decae
    + 0.000015 * DEL^2              # aumento al final
    - 0.005 * DEL                   # tendencia descendente general
    + 0.35 * (Epoca == "Lluviosa")  # efecto de época lluviosa
  )
)

datos_mastitis$RCS <- rnbinom(n, mu = mu_rcs, size = 2.5)

# Verificar la distribución del RCS
summary(datos_mastitis$RCS)

# ---- Ajuste del GAM ----

# Modelo con binomial negativa, suavizado de DEL y efecto de época
gam_mastitis <- gam(
  RCS ~ s(DEL, k = 20, bs = "tp") + Epoca,
  family = nb(),
  data = datos_mastitis,
  method = "REML"
)

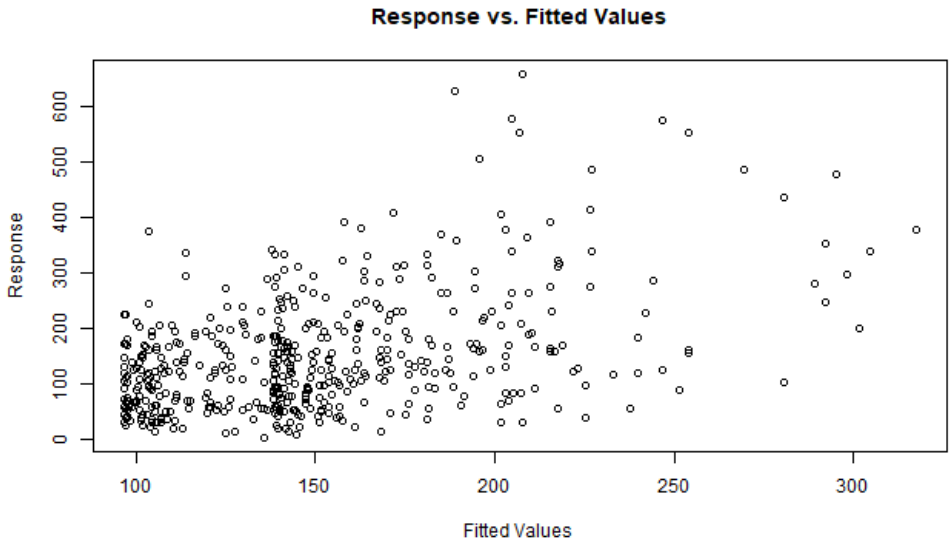
summary(gam_mastitis)

# ---- Diagnósticos ----

# gam.check para verificar k y residuos básicos
gam.check(gam_mastitis)
```

Figura 42

Gam Check




```
# Diagnósticos con DHARMa (más apropiados para familias no normales)
simulacion_res <- simulateResiduals(fittedModel = gam_mastitis)
plot(simulacion_res)
# Concurvidad (aquí solo hay un suavizado y un término paramétrico,
# así que no debería haber problema)
concurvity(gam_mastitis, full = TRUE)
# ---- Gráficos de efectos ----
# Efecto suave de DEL
draw(gam_mastitis, select = 1, residuals = TRUE) +
  labs(title = "Efecto suave de DEL sobre RCS (escala log)",
       x = "Días en leche",
       y = "Efecto parcial (escala del predictor lineal)")
```

Figura 43

Efecto suave de los días en leche sobre RCS

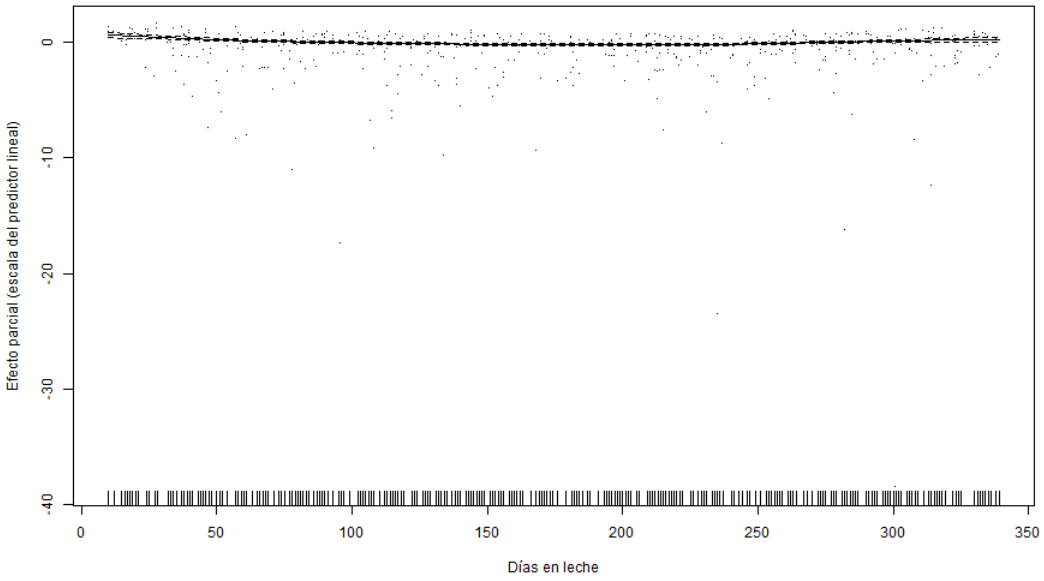
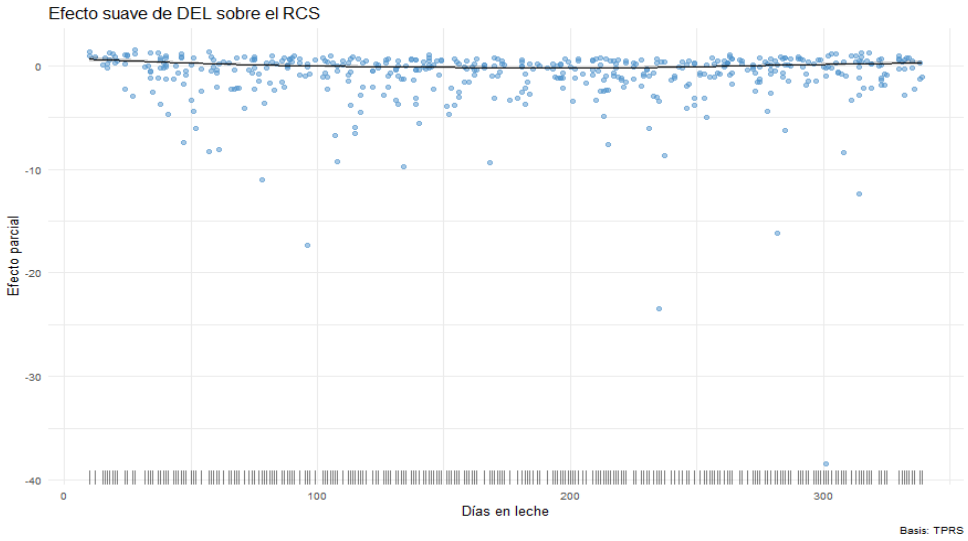


Figura 44

Efecto suave de DEL sobre RCS




```
# Predicciones para ambas épocas
pred_grid <- expand.grid(
  DEL   = seq(10, 340, length.out = 200),
  Epoca = c("Seca", "Lluviosa")
)

pred_grid$RCS_pred <- predict(gam_mastitis, newdata = pred_grid, type = "response")
```

```r
# Intervalo de confianza
pred_se <- predict(gam_mastitis, newdata = pred_grid, type = "link", se.fit = TRUE)
pred_grid$Lower <- exp(pred_se$fit - 1.96 * pred_se$se.fit)
pred_grid$Upper <- exp(pred_se$fit + 1.96 * pred_se$se.fit)

ggplot() +
  geom_point(data = datos_mastitis, aes(x = DEL, y = RCS, colour = Epoca),
             alpha = 0.25, size = 1.5) +
  geom_ribbon(data = pred_grid, aes(x = DEL, ymin = Lower, ymax = Upper,
              fill = Epoca), alpha = 0.2) +
  geom_line(data = pred_grid, aes(x = DEL, y = RCS_pred, colour = Epoca),
            linewidth = 1.2) +
  scale_colour_manual(values = c("Seca" = "#E8A838", "Lluviosa" = "#3498DB")) +
  scale_fill_manual(values = c("Seca" = "#E8A838", "Lluviosa" = "#3498DB")) +
  labs(
    title = "GAM: Recuento de células somáticas según días en leche y época",
    subtitle = "Modelo con distribución binomial negativa",
    x = "Días en leche",
    y = "RCS (miles de células/mL)"
  ) +
  theme_minimal(base_size = 13) +
  coord_cartesian(ylim = c(0, quantile(datos_mastitis$RCS, 0.98)))
```

Figura 45

GAM: RCS según días en leche y época

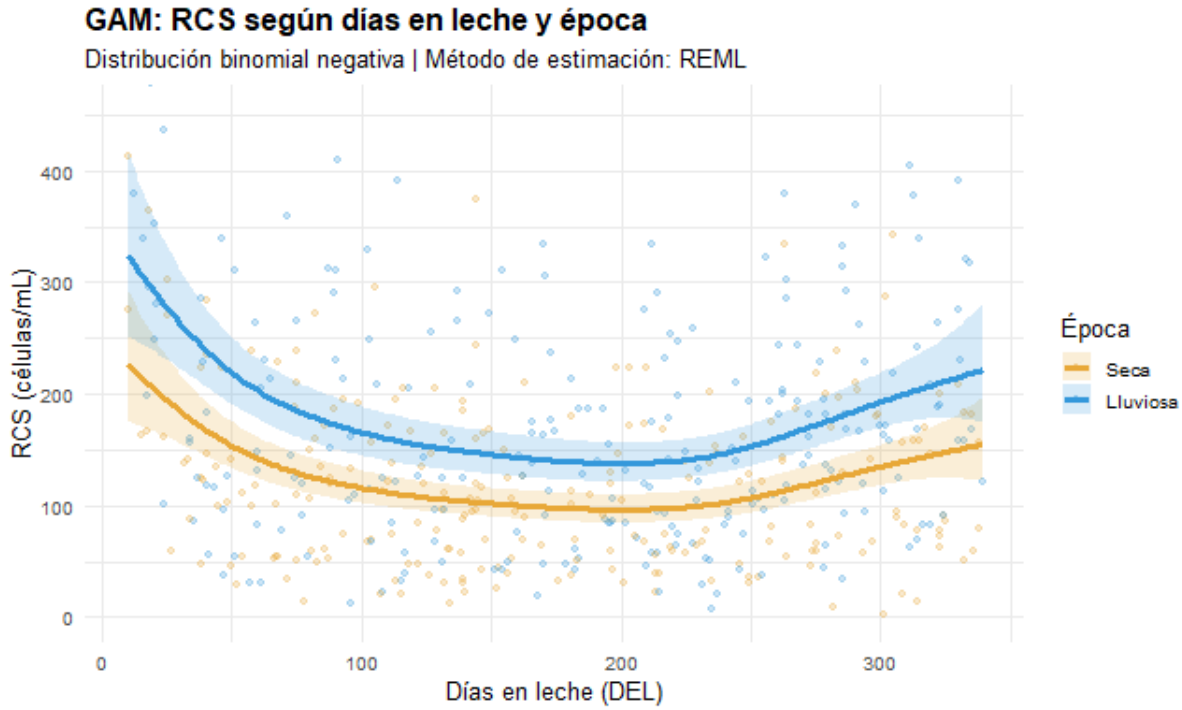
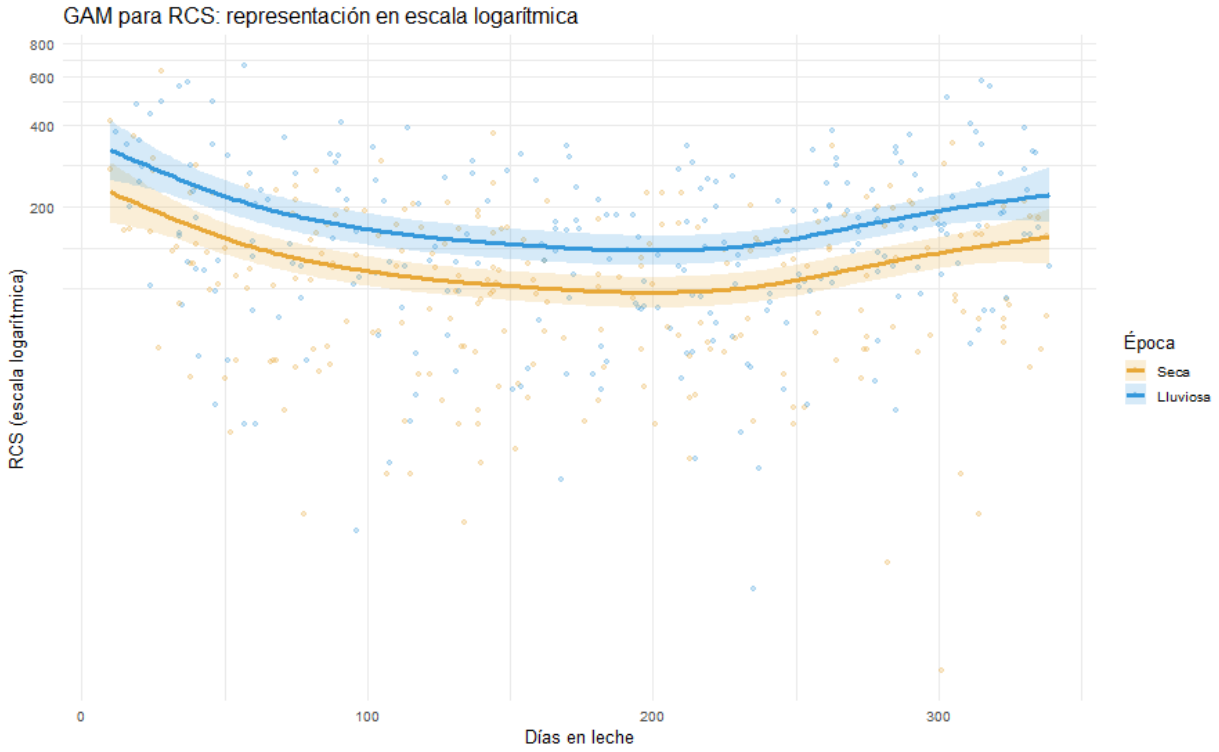


Figura 46

GAM para-RCS: representación en escala logarítmica



Descripción de la Figura 7.5. Las curvas predichas muestran el patrón en forma de U asimétrica del RCS a lo largo de la lactancia: valores elevados en los primeros 30 días posparto (asociados al estrés del periparto y posibles infecciones preexistentes), descenso hasta un mínimo entre los días 60-150, y aumento progresivo a partir del día 200 (acumulación de exposición, envejecimiento del tejido mamario). La curva de la época lluviosa se ubica consistentemente por encima de la seca, lo cual es coherente con la mayor exposición a patógenos ambientales (*Streptococcus uberis*, *E. coli*) cuando las condiciones de higiene en el alojamiento se deterioran por la lluvia y el barro.

Interpretación del resumen del modelo:

- Los edf del suavizado de DEL (esperados entre 4 y 6) confirman una relación no lineal que no podría capturarse con un GLM lineal.
- El coeficiente paramétrico de la época lluviosa (positivo, en escala log) indica un RCS multiplicativamente mayor en esa estación. Si el coeficiente es 0.35, el factor multiplicativo es $e^{0.35} \approx 1.42$: un 42% más de RCS en época lluviosa respecto a la seca, controlando por DEL.

- La dispersión estimada (θ de la binomial negativa, reportada como parámetro *theta* en la salida) indica el grado de sobredispersión.

¿Qué habría pasado con un modelo paramétrico? Un GLM Poisson con DEL como efecto lineal subestimaría el RCS al inicio y al final de la lactancia, y lo sobreestimaría a mediados. Además, la sobredispersión severa del RCS haría que los errores estándar fueran inapropiadamente pequeños, generando significancias espurias. Un GLM binomial negativo con DEL lineal captaría la sobredispersión pero no la no linealidad. Solo el GAM con distribución binomial negativa aborda ambos problemas simultáneamente.

7.12. Errores frecuentes y advertencias

1. Usar la distancia euclidiana para datos de abundancia ecológica. Es uno de los errores más extendidos en estudios de parasitología y microbiología veterinaria. La distancia euclidiana equipara la ausencia compartida de una especie con su presencia compartida, lo cual carece de sentido ecológico. Para abundancias, usar Bray-Curtis; para presencia-ausencia, Jaccard.

2. No verificar las dispersiones antes de interpretar PERMANOVA. Un resultado significativo en PERMANOVA no garantiza que los centroides de los grupos difieran; puede deberse a diferencias en dispersión (variabilidad). Siempre acompañar PERMANOVA con PERMDISP. Si PERMDISP es significativo, reportar ambos resultados y matizar la interpretación.

3. Interpretar los ejes del NMDS como variables latentes o gradientes. Los ejes del NMDS no tienen unidades ni interpretación intrínseca. No se puede decir "el eje 1 representa la altitud" salvo que se demuestre esa correlación externamente (por ejemplo, con *envfit()* de *vegan*). El NMDS es una herramienta de visualización, no un modelo causal.

4. Fijar k demasiado bajo en un GAM sin verificar. Si $k = 5$ y los edf resultantes son 3.9, el modelo está forzado a ser más simple de lo que los datos justifican. Siempre revisar la salida de *gam.check()* y, si sugiere que *k* es insuficiente, aumentarlo y reajustar.

5. Ignorar la sobredispersión en GAMs de Poisson. Con datos de conteo veterinarios (HPG, RCS, conteos de lesiones), la varianza casi siempre excede la media. Un GAM Poisson en esos datos producirá intervalos de confianza demasiado estrechos. Verificar con el cociente de devianza residual sobre grados de libertad residuales; si excede claramente 1, cambiar a binomial negativa.

6. Confundir regresión cuantílica con regresión a la mediana "porque es no paramétrica". La regresión cuantílica minimiza una función de pérdida asimétrica; no es simplemente "ajustar una mediana" en el sentido descriptivo. Los coeficientes tienen interpretación condicional (el cambio en el cuantil τ de Y por unidad de X), no marginal. Además, la inferencia (errores estándar, intervalos de confianza) requiere métodos específicos que difieren de los de mínimos cuadrados.

7. Reportar solo el valor p de PERMANOVA sin el R^2. El pseudo-F y el valor p dicen si hay diferencia significativa, pero no cuánta variación explica el factor. Un PERMANOVA con $p = 0.03$ y $R^2 = 0.04$ indica que el factor es estadísticamente significativo, pero explica solo el 4% de la variación composicional, lo cual puede ser biológicamente irrelevante. El R^2 es la medida del tamaño del efecto.

8. Aplicar LOESS con span inadecuado y no mostrar el rango de sensibilidad. Un span = 0.2 puede mostrar ondulaciones que son ruido; un span = 0.8 puede ocultar patrones reales. Si se reporta un LOESS, mostrar al menos dos valores de span o justificar la selección con validación cruzada.

7.13. Ejercicios de autoevaluación

Nivel básico

Ejercicio 7.1.

Un investigador quiere comparar la composición del microbiota ruminal (medida como abundancias de 50 géneros bacterianos) entre tres razas bovinas (Holstein, Jersey, criollo). Tiene datos de 10 animales por raza. ¿Qué prueba multivariada debería usar y por qué no es adecuado un MANOVA?

Respuesta. PERMANOVA sobre una matriz de Bray-Curtis, porque los datos de abundancia de géneros bacterianos presentan muchos ceros, distribuciones asimétricas y correlaciones complejas entre géneros que violan los supuestos de normalidad multivariada y homogeneidad de matrices de covarianza requeridos por MANOVA. Además, 50 variables con 30 observaciones es una relación variable/observación inadecuada para MANOVA.

Ejercicio 7.2.

Un NMDS en dos dimensiones arroja un estrés de 0.28. ¿Qué interpretación tiene ese valor y qué acción debería tomarse?

Respuesta. Un estrés de 0.28 indica una representación pobre: la configuración bidimensional no logra preservar adecuadamente las relaciones de similitud originales. Se debería intentar una solución en tres dimensiones (k = 3) y verificar si el estrés baja a un nivel aceptable (< 0.20, idealmente < 0.15). Si con k = 3 el estrés sigue alto, puede indicar que los datos no tienen una estructura de agrupamiento clara, o que la medida de distancia elegida no es la más adecuada.

Ejercicio 7.3.

En un GAM para producción lechera, la función *summary()* reporta edf = 1.002 para el suavizado de días en leche. ¿Qué indica esto?

Respuesta. Que la relación entre días en leche y producción es esencialmente lineal. El modelo penalizó la función suave hasta colapsarla en una recta. Un GLM con DEL como predictor lineal daría resultados prácticamente idénticos. Esto puede indicar que el rango de DEL en los datos es estrecho, o que la muestra es insuficiente para detectar no linealidad, o simplemente que la relación es lineal en ese hato.

Nivel intermedio

Ejercicio 7.4.

Los siguientes datos representan conteos de tres géneros de nematodos gastrointestinales en 8 ovinos Corriedale de dos fincas (A y B) en la provincia de Cotopaxi:

Ovino	Finca	*Haemonchus*	*Trichostrongylus*	*Ostertagia*
1	A	120	45	8
2	A	95	60	12
3	A	140	30	5
4	A	110	55	10
5	B	20	90	65
6	B	15	85	70
7	B	25	75	55
8	B	30	95	60

(a) Calcule la disimilitud de Bray-Curtis entre los ovinos 1 y 5.

(b) Escriba el código en R para calcular la matriz de Bray-Curtis completa y ejecutar una PERMANOVA con 999 permutaciones.

(c) Interprete los resultados en el contexto de control parasitario ovino.

Respuesta.

(d) $d_{BC}(1,5) = \frac{|120-20|+|45-90|+|8-65|}{(120+20)+(45+90)+(8+65)} = \frac{100+45+57}{140+135+73} = \frac{202}{348} = 0.580$

```
nematodos <- data.frame(
  Finca = factor(rep(c("A", "B"), each = 4)),
  Haemonchus = c(120, 95, 140, 110, 20, 15, 25, 30),
  Trichostrongylus = c(45, 60, 30, 55, 90, 85, 75, 95),
  Ostertagia = c(8, 12, 5, 10, 65, 70, 55, 60)
)
library(vegan)
bc_nem <- vegdist(nematodos[, 2:4], method = "bray")
set.seed(100)
adonis2(bc_nem ~ Finca, data = nematodos, permutations = 999)
```

(e) Se espera un R^2 alto y un pseudo-F significativo, indicando que las fincas difieren marcadamente en la composición de nematodos: la finca A está dominada por *Haemonchus* (hematófago, patogénicamente más agresivo), mientras que la finca B tiene un perfil dominado por *Trichostrongylus* y *Ostertagia*. Esto sugiere que los programas antihelmínticos deben considerar la composición específica y no solo la carga parasitaria total, ya que los principios activos difieren en eficacia según el género de nematodo.

Ejercicio 7.5.

Ajuste un GAM en R para modelar el peso vivo (kg) de corderos Corriedale en función de la edad (días), usando un suavizado con k = 10 y distribución gaussiana. Los datos son:

Edad (días)	**30**	**45**	**60**	**75**	**90**	**105**	**120**	**150**	**180**	**210**	**240**
Peso (kg)	8.2	11.5	14.8	17.0	19.5	21.2	23.8	27.0	29.5	31.0	32.2

(a) Escriba el código completo.

(b) ¿Qué esperaría para los edf?

(c) ¿El GAM aporta algo sobre un modelo lineal simple en este caso?

Respuesta.

```
library(mgcv)
corderos <- data.frame(
  Edad = c(30, 45, 60, 75, 90, 105, 120, 150, 180, 210, 240),
  Peso = c(8.2, 11.5, 14.8, 17.0, 19.5, 21.2, 23.8, 27.0, 29.5, 31.0, 32.2)
)
gam_cord <- gam(Peso ~ s(Edad, k = 10), data = corderos, method = "REML")
summary(gam_cord)
plot(gam_cord, shade = TRUE, residuals = TRUE)
```

(c) Con solo 11 observaciones y un patrón que parece levemente curvilíneo (la ganancia de peso se desacelera con la edad), se esperan edf entre 1.5 y 3. Si los edf son cercanos a 1, la relación es prácticamente lineal en este rango.

(d) La curva de crecimiento de corderos suele tener una fase de desaceleración (forma sigmoide). Con datos que cubren solo hasta 240 días, el GAM puede captar esa ligera curvatura, pero con 11 observaciones la ventaja sobre un modelo lineal será modesta. Un modelo lineal simple probablemente dé un R^2 similar. El GAM sería más informativo con más datos y un rango de edad más amplio.

Nivel avanzado

Ejercicio 7.6.

Un estudio registra el peso vivo (kg), la condición corporal (escala 1-5), los días en leche, la producción diaria (L), el recuento de células somáticas (miles/mL) y el número de parto en 150 vacas mestizas de tres sistemas lecheros en la Sierra ecuatoriana (50 vacas por sistema: familiar, semi-tecnificado, tecnificado). Se quiere: (a) evaluar si los perfiles productivos y sanitarios difieren entre sistemas usando PERMANOVA; (b) modelar el RCS como función suave de DEL, ajustando por época y sistema, con un GAM de distribución binomial negativa; (c) ajustar regresiones cuantílicas (τ = 0.25, 0.50, 0.75) de producción diaria en función de DEL. Diseñe el análisis completo: justifique la selección de distancias, escriba el código en R, interprete los resultados y redacte la sección de resultados como aparecería en un artículo para *Tropical Animal Health and Production*.

Rúbrica orientadora.

- **Justificación (20%).** Debe mencionar que para el PERMANOVA con variables mixtas (continuas y ordinales) la distancia de Gower es preferible a Bray-Curtis; que el RCS requiere distribución binomial negativa por sobredispersión severa; que la regresión cuantílica captura la heterocedasticidad de la producción lechera.
- **Código R (30%).** Debe ser completo, reproducible, con semilla fijada, y usar *vegdist()* con Gower para PERMANOVA, *gam()* con *nb()* y *s(DEL, k = 20)* para el GAM, y *rq()* para la regresión cuantílica. Debe incluir *betadisper()* para PERMDISP, *gam.check()* para diagnósticos del GAM, y gráficos de los tres cuantiles superpuestos.
- **Interpretación (25%).** Los resultados de PERMANOVA deben reportarse con pseudo-F, R^2, valor *p* y resultado de PERMDISP. Los del GAM con edf, significancia del suavizado, coeficientes paramétricos y su interpretación en escala original (e^{β}). Los de la regresión cuantílica con los coeficientes por cuantil y la discusión de heterocedasticidad.
- **Redacción estilo artículo (25%).** Debe seguir el formato: "Se realizó un PERMANOVA sobre la matriz de disimilitud de Gower (9999 permutaciones)…", reportar estadísticos entre paréntesis con formato APA (pseudo-F(2, 147) = …, *p* = …, R^2 = …), y usar tablas y figuras referenciadas en el texto.

Ejercicio 7.7.

Un colega le muestra un análisis donde aplicó un MANOVA a datos de abundancia relativa de cinco géneros bacterianos ruminales (con valores entre 0 y 0.6, muchos ceros), obtuvo Wilks' lambda significativa y concluyó que "la dieta modifica la composición bacteriana del rumen". Redacte un comentario técnico de 200-300 palabras explicando por qué ese análisis es problemático y qué alternativa recomendaría.

Rúbrica orientadora. El comentario debe identificar al menos tres problemas: (1) la violación de normalidad multivariada por los ceros y la asimetría de las proporciones; (2) la posible violación de homogeneidad de matrices de covarianza; (3) que las abundancias relativas son datos composicionales (suman a 1 dentro de cada muestra), lo que genera colinealidad espuria. La alternativa recomendada es PERMANOVA sobre Bray-Curtis, con verificación de dispersiones mediante PERMDISP. Para un análisis composicional riguroso, podría sugerirse adicionalmente la transformación CLR (centred log-ratio) antes de calcular la distancia euclidiana, siguiendo el enfoque de Aitchison (1986).

7.14. Conclusiones del capítulo

Los métodos multivariados basados en distancias (PERMANOVA, ANOSIM, NMDS) y los modelos de regresión no paramétrica (LOESS, GAM, regresión cuantílica) comparten una filosofía: dejar que la estructura de los datos guíe el análisis, en lugar de imponer supuestos distribucionales o formas funcionales que los datos biológicos rara vez satisfacen.

PERMANOVA se ha convertido en la herramienta estándar para comparar composiciones multivariadas en ecología y, cada vez más, en parasitología y microbiología veterinaria. Su implementación es directa, su interpretación es análoga a la del ANOVA, y su único requisito sustantivo verificar la homogeneidad de dispersiones es fácil de evaluar con PERMDISP. ANOSIM complementa este análisis con un índice de separación (R) interpretable directamente. El NMDS proporciona la visualización que las pruebas formales no pueden: un mapa de cómo se distribuyen las comunidades en el espacio de similitud.

Los GAM extienden los modelos lineales generalizados sin renunciar a su marco inferencial. La penalización automática del suavizado resuelve el dilema entre flexibilidad y parsimonia, y la

capacidad de incorporar distribuciones como la binomial negativa los hace aptos para la mayoría de las variables respuesta en producción y clínica veterinaria. La regresión cuantílica aporta una dimensión adicional: en sistemas biológicos con heterogeneidad inherente y la producción animal está llena de ellos modelar solo la media condicional es perder información sobre lo que ocurre en las colas de la distribución.

El capítulo siguiente aborda cómo reportar los resultados de estas y otras pruebas no paramétricas en artículos científicos, con las convenciones de formato y los elementos que los revisores esperan encontrar.

Capítulo 8

Reporte de resultados y tendencias en la investigación veterinaria

Introducción

Este capítulo trabaja dos dimensiones que cierran el ciclo del análisis estadístico. La primera es la comunicación de resultados: cómo reportar cada prueba no paramétrica tratada en los capítulos anteriores siguiendo las normas APA séptima edición y las convenciones de las revistas veterinarias. Se desarrollan plantillas de reporte con el formato exacto de los estadísticos en texto para cada prueba (Mann-Whitney, Kruskal-Wallis, Friedman, Wilcoxon, chi-cuadrado, Fisher, McNemar, Spearman, Kendall, log-rank, bootstrap, PERMANOVA). Se construyen tablas modelo con el contenido mínimo exigible: medianas con rango intercuartílico, estadístico de prueba, valor p exacto, tamaño del efecto con intervalo de confianza, y letras superíndices para comparaciones post hoc. Se trabajan los principios de diseño de figuras para datos no paramétricos, con énfasis en los gráficos que muestran distribuciones completas (boxplots con datos individuales, violin plots, curvas de Kaplan-Meier) frente a los que las ocultan (gráficos de barras con barras de error). Se implementa la generación automatizada de tablas y figuras con gtsummary, flextable y ggpubr. La segunda dimensión mira hacia las tendencias en el análisis de datos veterinarios: los métodos bayesianos no paramétricos (procesos de Dirichlet, modelos con brms), el aprendizaje automático como extensión de la lógica no paramétrica hacia problemas de clasificación y predicción (random forest con tidymodels y ranger), y los desafíos del análisis de datos masivos generados por sensores, plataformas genómicas y registros clínicos electrónicos. Se trabaja la relación entre estos métodos emergentes y la estadística no paramétrica clásica, con la tesis de que no compiten, sino que responden preguntas distintas y se complementan.

Objetivos de aprendizaje

Al finalizar este capítulo, el lector será capaz de:

1. **Redactar** la sección de resultados de un artículo científico veterinario que emplee pruebas no paramétricas, siguiendo las normas APA 7.ª edición y las convenciones de las principales revistas del área.
2. **Diseñar** tablas y figuras que comuniquen de manera clara y completa los resultados de análisis no paramétricos, incluyendo estadísticos de prueba, valores p, intervalos de confianza y tamaños del efecto.

3. **Identificar** los errores de reporte más frecuentes en la literatura veterinaria y corregirlos en manuscritos propios y ajenos.
4. **Describir** la lógica conceptual de los métodos bayesianos no paramétricos, el machine learning como extensión no paramétrica y el análisis de datos masivos en veterinaria, reconociendo sus conexiones con los métodos clásicos tratados en capítulos anteriores.
5. **Implementar** en R la generación automatizada de tablas de resultados con *gtsummary* y *flextable*, figuras con *ggpubr*, y un flujo básico de clasificación con *randomForest* y *tidymodels*, integrando estos productos en un reporte reproducible.

8.1. Normas de reporte estadístico según APA 7.ª edición aplicadas a ciencias veterinarias

8.1.1. Principios generales del reporte estadístico

La séptima edición del *Publication Manual* de la American Psychological Association (APA, 2020) establece un conjunto de principios para el reporte de análisis estadísticos que, aunque fueron desarrollados en el contexto de las ciencias del comportamiento, se aplican directamente a la investigación veterinaria. La mayoría de las revistas veterinarias indexadas (Journal of Veterinary Internal Medicine, Preventive Veterinary Medicine, Theriogenology, Veterinary Parasitology, entre otras) exigen o recomiendan adherencia a estas normas, a veces con adaptaciones menores propias de cada revista.

Los principios centrales son:

Suficiencia informativa. El lector debe poder evaluar la adecuación del análisis y la fuerza de la evidencia sin necesidad de contactar al autor. Esto implica reportar el estadístico de prueba, sus grados de libertad (cuando aplique), el valor p (exacto, no solo como "< 0.05"), el tamaño del efecto con su intervalo de confianza, y las medidas descriptivas apropiadas.

Precisión en la identificación de las pruebas. No basta con escribir "se usaron pruebas no paramétricas". Debe identificarse cada prueba por su nombre completo en la primera mención: "prueba de rangos con signo de Wilcoxon" (no "prueba de Wilcoxon" a secas, que es ambigua entre la prueba de rangos con signo y la prueba de suma de rangos, nombre alternativo de Mann-Whitney). En menciones subsecuentes puede usarse la forma abreviada.

Coherencia entre la medida descriptiva y el análisis. Si se aplicó una prueba basada en rangos, la medida de tendencia central debe ser la mediana, y la medida de dispersión, el rango intercuartílico (RIQ) o el rango completo (mínimo–máximo). Reportar media ± desviación estándar para datos analizados con Kruskal-Wallis es una incongruencia que aparece con frecuencia alarmante. La media asume simetría; si los datos fueran simétricos, probablemente se habría usado ANOVA.

Valores p exactos. APA 7.ª edición requiere reportar valores p exactos redondeados a dos o tres decimales ($p = 0.042$, $p = 0.003$), no desigualdades vagas ($p < 0.05$). Las únicas excepciones son $p < 0.001$ (cuando el valor es muy pequeño) y valores p obtenidos de tablas que solo dan umbrales. Con software moderno, esta última excepción casi no aplica.

Tamaño del efecto obligatorio. El valor p indica si un efecto es estadísticamente distinguible de cero; no dice nada sobre su magnitud ni sobre su relevancia clínica o productiva. APA exige reportar al menos una medida de tamaño del efecto. Para pruebas no paramétricas, las opciones principales son:

- *r* de rango biserial para Mann-Whitney y Wilcoxon ($r = Z / \sqrt{N}$).
- Epsilon cuadrado (ε^2) o eta cuadrada basado en H (η^2_H) para Kruskal-Wallis.
- W de Kendall para-Friedman.
- V de Cramér, phi (φ) o razón de momios (OR) para tablas de contingencia.
- rho de Spearman o tau de Kendall como tamaños del efecto en sí mismos para correlaciones.

Intervalos de confianza. Siempre que sea posible, acompañar las estimaciones puntuales (medianas, tamaños del efecto, hazard ratios) con intervalos de confianza al 95 %. Para medianas, los métodos de bootstrap del Capítulo 7 son la opción más versátil.

8.1.2. Formato de los estadísticos en texto

La convención APA para reportar estadísticos en texto sigue un formato general:

Símbolo del estadístico(grados de libertad) = valor, *p* = valor, tamaño del efecto = valor [IC 95 %: límite inferior, límite superior].

Los estadísticos y su símbolo se escriben en cursiva cuando son letras latinas (U, H, T, W, χ^2) o griegas que representan estadísticos de prueba. Los grados de libertad van entre paréntesis inmediatamente después del símbolo, sin espacio. Algunos ejemplos adaptados a las pruebas de este libro:

- Mann-Whitney: $U = 34.5$, $p = 0.018$, $r = 0.52$
- Kruskal-Wallis: $H(2) = 11.47$, $p = 0.003$, $\varepsilon^2 = 0.38$
- Friedman: $\chi^2\_F(3) = 18.60$, $p < 0.001$, $W = 0.62$
- Wilcoxon rangos con signo: $T = 12$, $p = 0.031$, $r = 0.58$
- Chi-cuadrado: $\chi^2(1, N = 84) = 7.32$, $p = 0.007$, $V = 0.30$
- Fisher exacta: $p = 0.024$ (bilateral), OR = 3.8 [IC 95 %: 1.2, 12.1]
- Spearman: $r\_s = 0.71$, $p < 0.001$
- Log-rank: $\chi^2(2) = 9.84$, $p = 0.007$

Nótese que para la prueba exacta de Fisher no hay un estadístico de prueba que reportar; se reporta directamente el valor p y la medida de asociación.

8.1.3. Particularidades del reporte en revistas veterinarias

Las revistas veterinarias tienen convenciones propias que conviene conocer:

El *Journal of Veterinary Internal Medicine* exige una subsección específica de análisis estadístico dentro de Materiales y Métodos donde se identifique cada prueba usada, el software y su versión, el nivel de significancia, y la justificación de pruebas no paramétricas cuando se eligieron sobre paramétricas. *Preventive Veterinary Medicine* solicita que se reporten intervalos de confianza para todas las estimaciones de riesgo (OR, RR, hazard ratio). *Tropical Animal Health and Production* permite tablas con medianas y rangos, pero solicita que las comparaciones múltiples post hoc estén claramente identificadas con superíndices en las tablas.

Una convención casi universal en revistas veterinarias es el uso de letras superíndices (a, b, ab) en tablas para indicar diferencias significativas entre grupos después de comparaciones post hoc. Esta convención, ausente en el manual APA, es la norma en la escritura veterinaria y zootécnica, y debe combinarse con la información completa del análisis en el texto o en notas al pie de la tabla.

8.2. Plantillas de reporte para cada prueba no paramétrica

Esta sección ofrece plantillas concretas para redactar el reporte de cada prueba cubierta en los capítulos anteriores. No son fórmulas rígidas; son estructuras que garantizan que toda la información necesaria esté presente. El investigador debe adaptar la redacción al contexto de su estudio.

8.2.1. Prueba de signos

Plantilla:

> Se evaluó si la mediana de [variable] difería de [valor hipotético / si hubo cambio entre las dos condiciones] mediante la prueba de signos. De los *n* pares/observaciones, *n*+ mostraron valores positivos, *n*− negativos y *n0* empates. La prueba de signos bilateral no fue significativa (*p* = [valor exacto]), lo que indica que no se encontró evidencia de [efecto esperado]. / La prueba resultó significativa (*p* = [valor]), con [n+/n−] de los [n] sujetos mostrando [dirección del cambio].

Ejemplo veterinario redactado:

Se evaluó si la suplementación con zinc oral modificaba el puntaje de condición de pelaje (escala ordinal de 1 a 5) en 14 gatos domésticos con dermatitis atópica, comparando el puntaje antes y después de 30 días de tratamiento. De los 14 gatos, 10 mostraron mejoría (puntaje post > pre), 2 empeoraron y 2 permanecieron sin cambio. La prueba de signos bilateral indicó un cambio significativo ($p = 0.039$), con una proporción observada de mejorías (10/12, tras excluir empates) superior a la esperada bajo la hipótesis nula de equiprobabilidad.

8.2.2. Prueba de rangos con signo de Wilcoxon

Plantilla:

> Las diferencias entre [condición 1] y [condición 2] se evaluaron mediante la prueba de rangos con signo de Wilcoxon. La mediana de las diferencias fue [valor] (RIQ: [Q1, Q3]). El estadístico *T* (suma de rangos del signo menos frecuente) fue [valor], *p* = [valor

exacto], con un tamaño del efecto *r* = [valor] (IC 95 %: [LI, LS]). Según los criterios de Cohen (1988), este efecto se clasifica como [pequeño/mediano/grande].

Ejemplo:

El efecto de la acupuntura sobre el puntaje de claudicación en 18 perros con osteoartritis de cadera se evaluó comparando los puntajes (escala ordinal de 0 a 4, donde 0 = sin claudicación y 4 = claudicación severa con ausencia de apoyo) antes y 7 días después de la tercera sesión. La mediana de las diferencias (pre − post) fue 1.0 (RIQ: 0.5, 2.0). La prueba de rangos con signo de Wilcoxon indicó una reducción significativa del puntaje de claudicación (T = 15.5, p = 0.004, r = 0.61, IC 95 %: 0.28, 0.82). El tamaño del efecto fue grande según los umbrales de Cohen (1988).

8.2.3. Prueba de Mann-Whitney U

Plantilla:

La [variable dependiente] se comparó entre [grupo 1] (n = [n1]) y [grupo 2] (n = [n2]) mediante la prueba de Mann-Whitney U. La mediana [IC 95 %] fue [valor] ([LI, LS]) en [grupo 1] y [valor] ([LI, LS]) en [grupo 2]. La diferencia fue significativa (U = [valor], p = [valor exacto]), con un tamaño del efecto r = [valor] (IC 95 %: [LI, LS]), indicando que [interpretación veterinaria].

Ejemplo:

El recuento de huevos por gramo de heces (HPG, técnica de McMaster) se comparó entre bovinos Brahman (n = 22) y bovinos Holstein (n = 19) en una finca de doble propósito en Manabí, Ecuador. La mediana fue 450 HPG (RIQ: 200, 850) en Brahman y 1200 HPG (RIQ: 600, 2100) en Holstein. La prueba de Mann-Whitney indicó que los bovinos Holstein presentaron recuentos significativamente mayores (U = 87, p < 0.001, r = 0.62, IC 95 %: 0.38, 0.79). Esto corrobora la mayor susceptibilidad de las razas europeas a la parasitosis gastrointestinal en condiciones tropicales.

8.2.4. Prueba de Kruskal-Wallis y comparaciones post hoc

Plantilla:

Se comparó [variable] entre [k] grupos mediante la prueba de Kruskal-Wallis. Se encontraron diferencias significativas, $H([k-1]) = [valor]$, $p = [valor]$, $\varepsilon^2 = [valor]$. Las comparaciones post hoc con la prueba de Dunn (ajuste de Bonferroni/Holm) indicaron que [grupo A] difirió de [grupo B] (p_adj = [valor]) y de [grupo C] (p_adj = [valor]), mientras que [grupo B] y [grupo C] no difirieron entre sí (p_adj = [valor]). Las medianas (RIQ) fueron: [grupo A] = [valor] ([Q1, Q3]), [grupo B] = [valor] ([Q1, Q3]), [grupo C] = [valor] ([Q1, Q3]).

8.2.5. Prueba de Friedman y comparaciones post hoc

Plantilla:

El efecto de [factor intra-sujeto con k niveles] sobre [variable] se evaluó mediante la prueba de Friedman en [n] sujetos. Se detectaron diferencias significativas entre los [k] tiempos/condiciones, χ^2_F([k−1]) = [valor], p = [valor], W = [valor]. Las comparaciones post hoc por pares (Nemenyi / Conover-Iman con ajuste de [método]) indicaron diferencias entre [tiempo/condición A] y [tiempo/condición C] (p_adj = [valor]), pero no entre [A y B] ni entre [B y C].

8.2.6. Chi-cuadrado de Pearson y prueba exacta de Fisher

Plantilla para χ^2:

La asociación entre [variable 1] y [variable 2] se evaluó con la prueba de chi-cuadrado de Pearson. Se encontró una asociación significativa, $\chi^2([gl], N = [N]) = [valor]$, p = [valor], V de Cramér = [valor]. [Descripción de la dirección de la asociación basada en los porcentajes por fila o columna.]

Plantilla para Fisher:

Dado que [proporción]% de las celdas presentaron frecuencias esperadas inferiores a 5, se utilizó la prueba exacta de Fisher. La asociación entre [variable 1] y [variable 2] fue significativa (p = [valor bilateral]). La razón de momios fue [valor] (IC 95 %: [LI, LS]), indicando que [grupo/condición] tuvo [valor] veces más probabilidad de [evento] que [grupo de referencia].

8.2.7. McNemar y Cochran’s Q

Plantilla para McNemar:

> El cambio en la proporción de [evento] entre [tiempo 1] y [tiempo 2] se evaluó mediante la prueba de McNemar en [n] sujetos. De los [n] sujetos, [b] cambiaron de negativo a positivo y [c] de positivo a negativo. La prueba fue [significativa/no significativa], χ^2_McNemar(1) = [valor], p = [valor] (o p exacto = [valor] si n_discordantes < 25).

Plantilla para Cochran’s Q:

> La proporción de [evento] se comparó a lo largo de [k] tiempos/condiciones en [n] sujetos mediante la prueba Q de Cochran. Se encontraron diferencias significativas, Q([k−1]) = [valor], p = [valor]. Las comparaciones post hoc por pares (McNemar con ajuste de [método]) indicaron que la proporción cambió significativamente entre [condiciones específicas].

8.2.8. Correlación no paramétrica

Plantilla:

> La asociación entre [variable X] y [variable Y] se evaluó mediante el coeficiente de correlación de Spearman (/ tau-b de Kendall). Se encontró una correlación [positiva/negativa] [fuerte/moderada/débil], r_s = [valor], p = [valor] (/ τ_b = [valor], p = [valor]).

Nota sobre interpretación: Spearman mide asociación monótona, no lineal. Un $r_s = 0$ no implica ausencia de relación; implica ausencia de relación monótona. Esto debe quedar claro en la redacción cuando la nube de puntos sugiera una relación no monótona.

8.2.9. Análisis de sobrevivencia: Kaplan-Meier y log-rank

Plantilla:

> La sobrevivencia se estimó mediante el método de Kaplan-Meier. La mediana de sobrevivencia fue [valor] [unidad de tiempo] (IC 95 %: [LI, LS]) en [grupo 1] y [valor]

[unidad de tiempo] (IC 95 %: [LI, LS]) en [grupo 2]. La tasa de sobrevivencia a [tiempo específico] fue [%] (IC 95 %: [LI, LS]) en [grupo 1] y [%] en [grupo 2]. Las curvas difirieron significativamente según la prueba de log-rank, χ^2([gl]) = [valor], p = [valor]. [Si se ajustó un modelo de Cox:] El hazard ratio ajustado fue [valor] (IC 95 %: [LI, LS], p = [valor]).

8.2.10. Bootstrap e intervalos de confianza por remuestreo

Plantilla:

El intervalo de confianza del 95 % para [estadístico: mediana, diferencia de medianas, coeficiente de correlación] se estimó mediante bootstrap con [número] réplicas (método [percentil/BCa]). La estimación puntual fue [valor] con IC 95 % bootstrap: [LI, LS].

8.2.11. PERMANOVA y ANOSIM

Plantilla:

Las diferencias en la composición de [comunidad/ensamblaje] entre [grupos] se evaluaron mediante PERMANOVA (Anderson, 2001) con [número] permutaciones, usando la distancia de [Bray-Curtis/Jaccard/Euclidiana]. Se encontraron diferencias significativas, pseudo-F([gl1], [gl2]) = [valor], R^2 = [valor], p_perm = [valor]. La prueba de homogeneidad de dispersiones (PERMDISP) indicó [dispersiones homogéneas/heterogéneas] (F = [valor], p = [valor]), lo que [sustenta/compromete] la interpretación de PERMANOVA como diferencia de ubicación y no de dispersión.

8.3. Diseño de tablas de resultados: contenido mínimo y buenas prácticas

8.3.1. Principios generales

Una tabla de resultados para pruebas no paramétricas debe contener, como mínimo:

1. Nombre de la variable en la primera columna.
2. Descriptivos apropiados para cada grupo: mediana (RIQ), o mediana (mínimo–máximo), o frecuencia absoluta y porcentaje para datos categóricos.

3. Estadístico de prueba y valor p en columnas separadas.
4. Tamaño del efecto.
5. Nota al pie que identifique la prueba usada, el método de ajuste para comparaciones múltiples, y la convención de letras superíndices si se emplean.

Lo que no debe aparecer en una tabla de resultados no paramétricos: media ± desviación estándar como descriptivo principal (aunque puede incluirse como información complementaria si la revista lo exige), asteriscos sin definición, y "NS" como sustituto del valor p exacto.

8.3.2. Tabla modelo para comparación de grupos independientes

Tabla 43
Parámetros hematológicos y parasitológicos en bovinos según raza, finca El Porvenir, Manabí, Ecuador, 2023.

Variable	Brahman (*n* = 22) Md (RIQ)	Holstein (*n* = 19) Md (RIQ)	*U*	*p*	*r*
HPG (McMaster)	450 (200–850) [a]	1200 (600–2100) [b]	87	< 0.001	0.62
Hematocrito (%)	32 (29–35)	30 (27–34)	168	0.215	0.19
Eosinófilos (×10³/µL)	0.8 (0.4–1.3) [a]	1.5 (0.9–2.4) [b]	102	0.003	0.46

Nota. Prueba de Mann-Whitney U bilateral. Md = mediana; RIQ = rango intercuartílico (percentil 25–percentil 75). *r* = tamaño del efecto de rango biserial. Letras superíndices diferentes indican diferencias significativas ($p < 0.05$).

8.3.3. Tabla modelo para comparación de más de dos grupos

Tabla 44
Tiempo de recuperación posanestésica (minutos) según protocolo, en 36 perras sometidas a ovario histerectomía, Hospital Veterinario USFQ, Quito, 2024.

Protocolo	*n*	Md (RIQ)	Rango medio
Propofol + isoflurano	12	18.5 (14–24)[a]	12.3
Tiletamina-zolazepam	12	28.0 (22–35)[b]	22.8

Protocolo	*n*	Md (RIQ)	Rango medio
Propofol + sevoflurano	12	16.0 (12–21)[a]	10.4

Nota. Kruskal-Wallis $H(2) = 11.47$, $p = 0.003$, $\varepsilon^2 = 0.33$. Comparaciones post hoc: prueba de Dunn con ajuste de Bonferroni. Letras superíndices diferentes indican diferencias significativas (p_adj < 0.05). El protocolo de tiletamina-zolazepam mostró tiempos de recuperación significativamente más prolongados que ambos protocolos basados en propofol.

8.3.4. Tabla modelo para datos de medidas repetidas

Tabla 45
Puntaje de dolor (escala analógica visual modificada, 0–4) a las 2, 6 y 12 horas posquirúrgicas en 15 perras del grupo propofol + isoflurano

Tiempo	Md (RIQ)	Rango medio
2 h	3 (2–4)	2.73
6 h	2 (1–3)	1.87
12 h	1 (0–2)	1.40

Nota. Friedman $\chi^2(2) = 16.80$, $p < 0.001$, $W = 0.56$. Comparaciones post hoc (Conover-Iman, ajuste de Holm): 2 h vs. 6 h (p_adj = 0.032), 2 h vs. 12 h (p_adj < 0.001), 6 h vs. 12 h (p_adj = 0.048).

8.3.5. Tabla modelo para tablas de contingencia

Tabla 46
Incidencia de vómito posanestésico según protocolo anestésico en 36 perras.

Protocolo	Vómito Sí n (%)	Vómito No n (%)	Total
Propofol + isoflurano	2 (16.7)	10 (83.3)	12
Tiletamina-zolazepam	7 (58.3)	5 (41.7)	12
Propofol + sevoflurano	1 (8.3)	11 (91.7)	12
Total	10 (27.8)	26 (72.2)	36

Nota. Chi-cuadrado de Pearson: $\chi^2(2, N = 36) = 8.91$, $p = 0.012$, V de Cramér = 0.50. Dado que 33 % de las celdas tuvieron frecuencias esperadas < 5, se verificó con prueba exacta de Fisher-Freeman-Halton ($p = 0.018$). El protocolo de tiletamina-zolazepam presentó una incidencia de vómito significativamente mayor.

8.4. Figuras para métodos no paramétricos

8.4.1. Principios de diseño gráfico para datos no paramétricos

Las figuras en un artículo científico veterinario no son decoración; son la primera cosa que el lector mira después del resumen, y a menudo la última antes de decidir si lee el artículo completo. Para datos analizados con métodos no paramétricos, las figuras deben reflejar la misma lógica que el análisis: mostrar distribuciones completas, medianas, dispersión asimétrica y datos individuales, no medias con barras de error que asumen simetría.

Principios rectores:

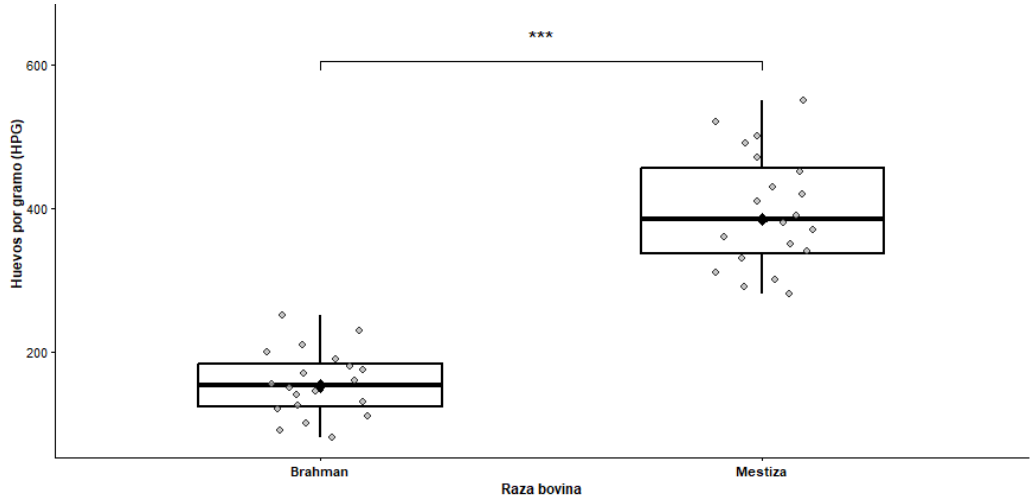
1. **Mostrar los datos individuales.** Con muestras de 6 a 30 animales por grupo (lo habitual en muchos estudios veterinarios), los puntos individuales aportan más información que cualquier resumen. Los gráficos de puntos superpuestos sobre boxplots (*jittered strip plots*) son la opción preferida.
2. **No usar gráficos de barras para datos continuos o semicontinuos.** El gráfico de barras oculta la distribución. Un estudio con dos tratamientos que tengan la misma mediana, pero distribuciones radicalmente distintas producirá dos barras idénticas.
3. **Anotar la significancia estadística con cuidado.** Los asteriscos (*) o las barras de comparación entre grupos deben ir acompañados de la información exacta en la leyenda o en el texto. "* $p < 0.05$, ** $p < 0.01$" es aceptable. Letras superíndices son otra convención habitual. Lo que no es aceptable: asteriscos sin definición.
4. **Escala del eje Y coherente.** Para datos de HPG o conteos bacterianos con distribuciones extremadamente sesgadas, considerar una escala logarítmica en el eje Y. Esto es legítimo siempre que se indique claramente en la leyenda del eje.

8.4.2. Boxplots con datos individuales (jittered strip)

El boxplot clásico muestra la mediana (línea central), el RIQ (caja), los bigotes (1.5 × RIQ o valores extremos) y los puntos atípicos. Para estudios veterinarios con muestras pequeñas, agregar los puntos individuales con un pequeño desplazamiento horizontal (*jitter*) permite al lector ver cada observación.

Figura 47

Recuento de huevos por gramo (HPG) según raza bovina, finca El Porvenir, Manabí, Ecuador.



Nota. Los boxplots muestran la mediana, el rango intercuartílico y los valores extremos. Los puntos representan observaciones individuales. ** $p < 0.01$, prueba de Mann-Whitney.

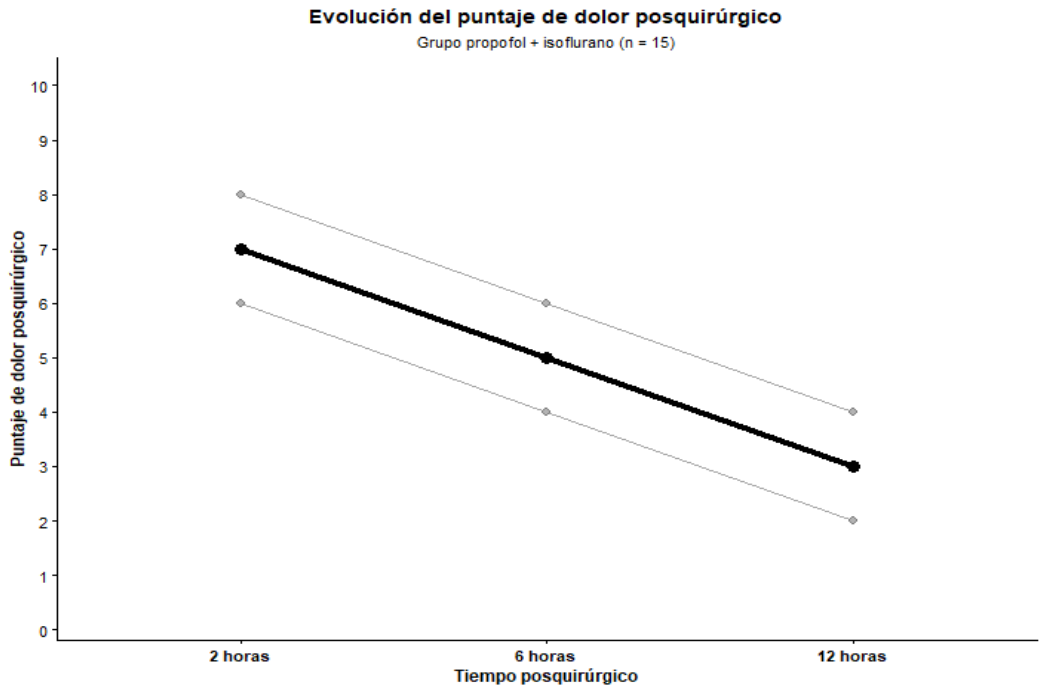
Descripción para diagramación: Eje X: dos categorías (Brahman, Holstein). Eje Y: HPG (escala logarítmica recomendada dado el sesgo positivo de los datos). Boxplots en colores diferenciados (gris claro para Brahman, gris oscuro para Holstein). Puntos individuales superpuestos con jitter horizontal de 0.15. Barra de comparación entre los dos grupos con "**" centrado.

8.4.3. Gráficos de líneas para medidas repetidas

Para datos de Friedman (medidas repetidas), el gráfico de líneas conectando las medianas de cada tiempo, con los puntos individuales en gris claro, muestra tanto la tendencia central como la variabilidad individual.

Figura 48

Evolución del puntaje de dolor posquirúrgico a las 2, 6 y 12 horas en perras del grupo propofol + isoflurano (n = 15).



Nota. Líneas grises: trayectorias individuales. Línea negra gruesa: mediana en cada tiempo. Los puntajes disminuyeron significativamente a lo largo del tiempo (Friedman $\chi^2(2) = 16.80$, $p < 0.001$).

8.4.4. Curvas de Kaplan-Meier

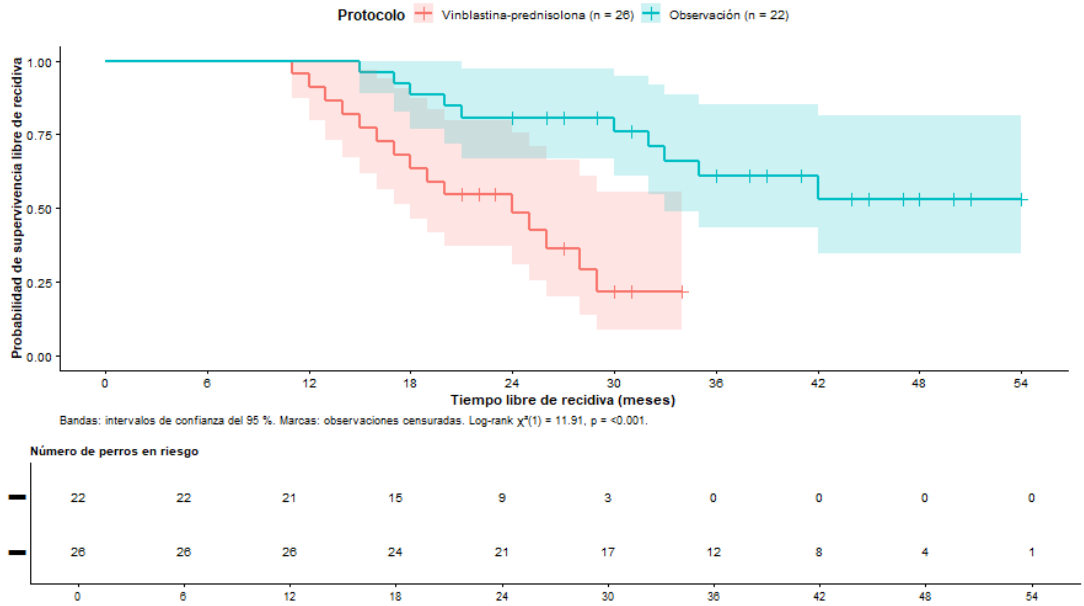
La curva de Kaplan-Meier se grafica como una función escalonada, con un escalón descendente en cada evento (muerte, recaída, alta). Las convenciones gráficas son:

- Eje X: tiempo de seguimiento (días, semanas, meses).
- Eje Y: probabilidad de sobrevivencia (de 0 a 1, o de 0 a 100 %).
- Marcas de censura: pequeños segmentos verticales (|) sobre la curva en cada punto de censura.
- Bandas de confianza al 95 % como sombreado semitransparente alrededor de cada curva.
- Tabla de riesgo (*number at risk*) debajo del gráfico, mostrando cuántos sujetos permanecían en seguimiento en cada intervalo.

- Valor p del log-rank y hazard ratio en la anotación del gráfico.

Figura 49

Sobrevivencia libre de recidiva en 48 perros con mastocitoma cutáneo de grado II según protocolo quimioterapéutico: vinblastina-prednisolona (n = 26) vs. observación (n = 22).



Nota. Curvas de Kaplan-Meier con bandas de confianza al 95 %. Tabla de riesgo debajo del gráfico. Log-rank $\chi^2(1) = 5.83$, $p = 0.016$.

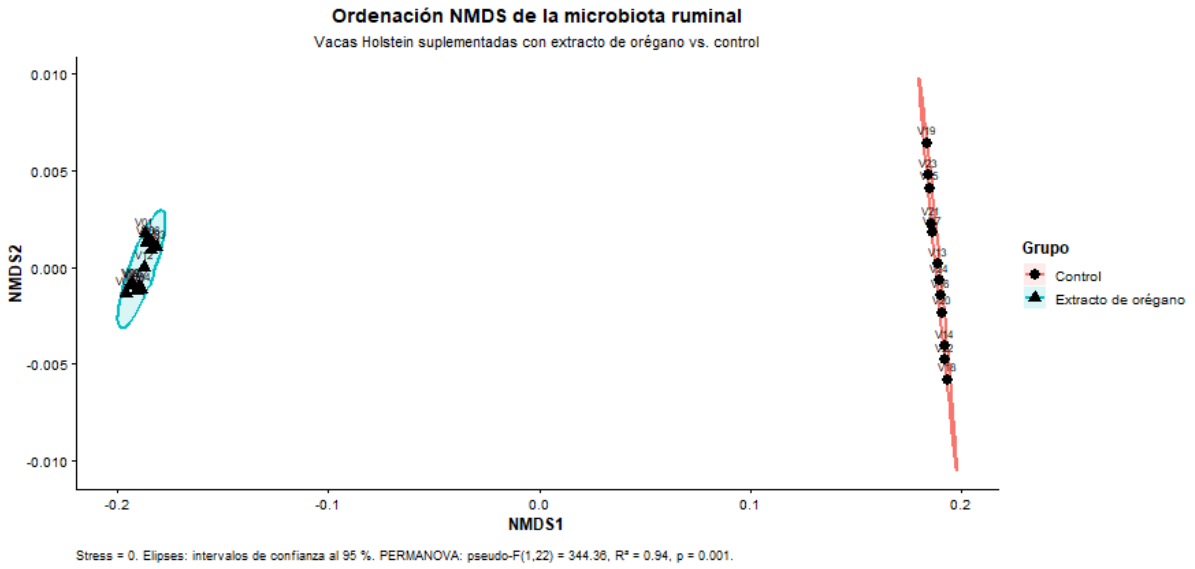
8.4.5. Gráficos NMDS (ordenación no paramétrica)

Para estudios de comunidades microbianas (metagenómica ruminal, microbiota intestinal de aves), PERMANOVA y ANOSIM se acompañan de un gráfico de ordenación NMDS en dos dimensiones.

- Ejes: NMDS1 y NMDS2 (sin unidades).
- Puntos: muestras individuales, coloreadas por grupo.
- Elipses: elipses de confianza al 95 % o envolventes convexas (*convex hulls*) por grupo.
- Estrés (*stress*): reportar en una esquina del gráfico (stress < 0.10 = buen ajuste; 0.10–0.20 = aceptable; > 0.20 = interpretación precaria).
- Anotación: pseudo-F y R^2 de PERMANOVA, y valor p.

Figura 50

Ordenación NMDS (Bray-Curtis) del microbiota ruminal de vacas lecheras Holstein suplementadas con extracto de orégano (n = 12) vs. control (n = 12), finca experimental de la Universidad Nacional de Colombia, Bogotá. Stress = 0.09



Nota. Las elipses representan intervalos de confianza al 95 %. PERMANOVA: pseudo-$F(1,22)$ = 3.41, R^2 = 0.13, p = 0.003.

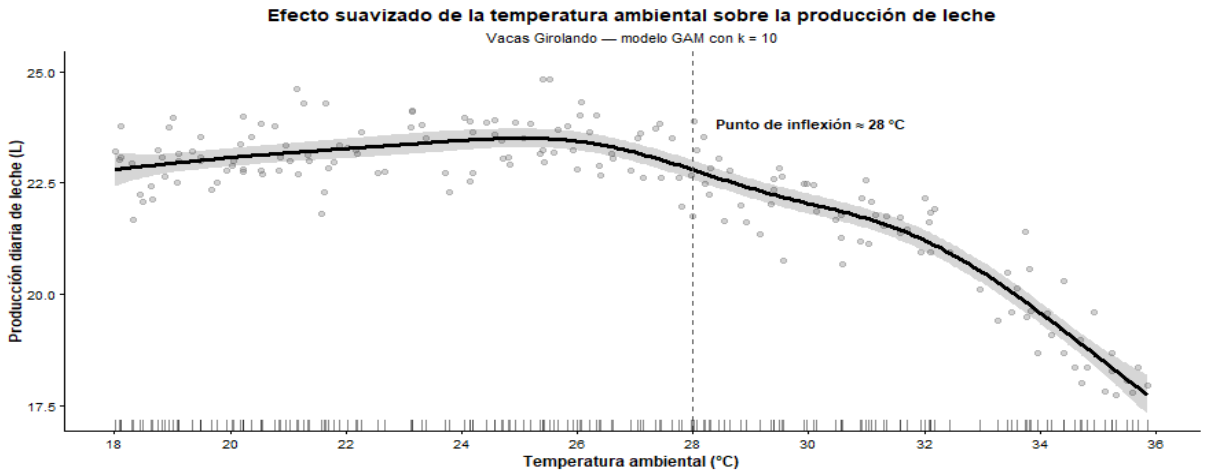
8.4.6. Gráficos de efectos GAM

Para modelos aditivos generalizados (Capítulo 7), los efectos suavizados (*smooth effects*) se grafican con:

- Eje X: la variable predictora.
- Eje Y: el efecto parcial (en la escala del predictor lineal, centrado en cero).
- Línea suavizada con banda de confianza al 95 %.
- Marcas de rug (*rug plot*) en el eje X mostrando la distribución de las observaciones.
- Grados de libertad efectivos (edf) en el título o la anotación del panel.

Figura 51

Efecto suavizado de la temperatura ambiental (°C) sobre la producción diaria de leche (L) en vacas Girolando, modelo GAM con k = 10. edf = 4.2.



Nota. La banda sombreada representa el IC 95 %. Las marcas verticales en el eje X muestran la distribución de las observaciones de temperatura. Se observa una relación no lineal con un punto de inflexión alrededor de 28 °C, compatible con estrés calórico.

8.5. Errores de reporte frecuentes en la literatura veterinaria

8.5.1. Reportar media ± DE para datos analizados con pruebas no paramétricas

Este es, con diferencia, el error más frecuente. Ocurre porque el investigador aprendió en cursos de estadística básica que las tablas de resultados llevan "media ± desviación estándar" y aplica esa plantilla sin pensar. Si los datos no cumplieron supuestos de normalidad y se analizaron con Kruskal-Wallis, la medida descriptiva coherente es la mediana con el rango intercuartílico. Presentar la media engaña al lector: sugiere una distribución simétrica que no existe y puede dar una impresión completamente distinta de la ubicación central de los datos.

Excepción legítima: algunas revistas piden que se reporten ambas medidas (media ± DE y mediana con RIQ). Si es el caso, la tabla debe dejar claro cuál es la medida principal y cuál la complementaria.

8.5.2. Omitir el estadístico de prueba

Escribir “la diferencia fue significativa (p = 0.03, prueba de Mann-Whitney)” sin reportar el valor de *U* impide que el lector recalcule, verifique o realice un meta-análisis posterior con esos datos. El estadístico de prueba es obligatorio. La única excepción parcial es la prueba exacta de Fisher, donde el valor p se calcula directamente sin un estadístico intermedio.

8.5.3. No especificar si el valor p es exacto, asintótico o con corrección de continuidad

La función *wilcox.test()* de R, por defecto, aplica una corrección de continuidad al cálculo asintótico del valor p. Esta corrección puede cambiar un valor p de 0.048 a 0.052, convirtiendo un resultado “significativo” en “no significativo” para quienes usan el umbral de 0.05 como frontera binaria. El investigador debe reportar qué tipo de valor p obtuvo: exacto (preferible con muestras pequeñas), asintótico con corrección de continuidad, o asintótico sin corrección. En el paquete *coin*, la función *wilcox_test()* calcula valores p exactos con el argumento *distribution = "exact"*, y el valor resultante puede diferir del que da *wilcox.test()*.

8.5.4. No reportar el tamaño del efecto

Un valor p de 0.002 con $n = 200$ puede corresponder a una diferencia clínicamente trivial. Un valor p de 0.08 con $n = 12$ puede esconder un efecto clínicamente relevante que el estudio no tuvo potencia suficiente para detectar. Sin el tamaño del efecto, el lector no tiene forma de distinguir estos escenarios.

8.5.5. Usar “pruebas no paramétricas” como nombre genérico

“Los datos se analizaron con pruebas no paramétricas” no es información. Es como escribir “el animal fue tratado con un fármaco”. El lector necesita saber cuál prueba se aplicó a cuál variable.

8.5.6. Confundir el reporte de Kruskal-Wallis con el de chi-cuadrado

Ambas pruebas producen un estadístico que sigue una distribución chi-cuadrado bajo la hipótesis nula, pero son pruebas distintas. Kruskal-Wallis se reporta como $H(\text{gl})$ o como $\chi^2\_\text{KW}(\text{gl})$, nunca como $\chi^2(\text{gl})$ a secas, para no confundirlo con la prueba de chi-cuadrado de independencia. Si la revista usa la notación χ^2, debe aclararse en la nota de la tabla: “χ^2 corresponde al estadístico de Kruskal-Wallis”.

8.5.7. Gráficos de barras para datos ordinales

El puntaje de condición corporal (1 a 5), el grado de claudicación (0 a 4), la escala de evaluación de heridas quirúrgicas: todas son variables ordinales. Un gráfico de barras con media y barras de error implica que el promedio tiene sentido y que la distribución es simétrica alrededor de él. Para estas variables, un gráfico de frecuencias apiladas, un gráfico de mosaico o un gráfico de violín son más apropiados.

8.5.8. Reportar comparaciones post hoc sin ajuste de multiplicidad

Cuando se hacen 6 comparaciones de pares después de Kruskal-Wallis con 4 grupos, la probabilidad de obtener al menos un falso positivo al 5 % por comparación ya no es 5 %; es considerablemente mayor. Si no se aplica un ajuste (Bonferroni, Holm, Benjamini-Hochberg), el investigador debe al menos declararlo y justificarlo. Reportar valores p de comparaciones múltiples sin mencionar si fueron ajustados o no es un error que dificulta la interpretación.

8.6. Introducción a métodos bayesianos no paramétricos

8.6.1. ¿Por qué métodos bayesianos en un libro de no paramétricos?

Los capítulos anteriores operaron dentro del marco frecuentista: se formuló una hipótesis nula, se calculó un estadístico de prueba, se obtuvo un valor p, y se rechazó o no se rechazó la hipótesis nula. El marco bayesiano ofrece una alternativa donde, en lugar de preguntar "¿Cuál es la probabilidad de obtener estos datos si la hipótesis nula fuera cierta?", se pregunta "Dada la evidencia que tengo, ¿cuál es la probabilidad de que la diferencia entre tratamientos tenga cierta magnitud?". La respuesta no es un valor p sino una distribución posterior que describe la incertidumbre completa sobre el parámetro de interés.

Lo bayesiano y lo no paramétrico se cruzan de dos maneras: primero, en modelos bayesianos que usan priors no paramétricos (como los procesos de Dirichlet) para evitar especificar una forma funcional rígida; segundo, en la estimación bayesiana de los mismos estadísticos que los métodos no paramétricos clásicos evalúan (medianas, diferencias de distribución, funciones de sobrevivencia).

8.6.2. Procesos de Dirichlet: la idea conceptual

Un proceso de Dirichlet (Ferguson, 1973) es una distribución de probabilidad sobre distribuciones de probabilidad. En lugar de asumir que los datos provienen de una distribución normal (o de cualquier otra forma paramétrica), se asume que provienen de una distribución desconocida *G*, y se coloca un prior sobre el espacio de todas las posibles distribuciones *G*. Ese prior es un proceso de Dirichlet.

La idea se entiende mejor con una analogía veterinaria. Supongamos que se quiere modelar la distribución de pesos al nacimiento de terneros en una región, y se sospecha que no es normal: tal vez tiene dos modas porque hay mezcla de razas, o tiene una cola pesada por condiciones nutricionales heterogéneas. En lugar de decidir a priori que es normal, o lognormal, o bimodal, se le dice al modelo: "Aprende la forma de la distribución directamente de los datos, partiendo de una idea general (la distribución base) pero dejando que los datos la modifiquen". El parámetro de concentración α controla cuánto peso se le da a la distribución base frente a los datos: con α grande, la posterior se parece mucho a la distribución base; con α pequeño, los datos dominan y la posterior puede adoptar formas completamente distintas.

En la práctica, los modelos de mezcla con procesos de Dirichlet (*Dirichlet Process Mixture Models*, DPMM) agrupan observaciones en clústeres sin especificar de antemano cuántos clústeres hay. Esto tiene aplicaciones directas en veterinaria:

- Agrupar perfiles metabólicos de vacas lecheras sin decidir a priori cuántos fenotipos metabólicos existen.
- Clasificar patrones de actividad animal medidos por acelerómetros, donde el número de comportamientos distintos (pastoreo, descanso, rumia, locomoción) puede ser desconocido o variable entre individuos.
- Modelar la heterogeneidad en la respuesta farmacológica cuando se sospecha que hay subpoblaciones con respuestas cualitativamente distintas (respondedores vs. no respondedores, por ejemplo).

8.6.3. Ventajas y limitaciones en el contexto veterinario

Los métodos bayesianos no paramétricos tienen ventajas concretas para la investigación veterinaria: manejan bien muestras pequeñas (porque incorporan información previa), producen

intervalos de credibilidad que son más intuitivos que los intervalos de confianza frecuentistas ("hay un 95 % de probabilidad de que el parámetro esté en este intervalo"), y permiten una flexibilidad distribucional que los métodos paramétricos no tienen.

Las limitaciones son igualmente concretas: la especificación del prior es subjetiva y puede influir en los resultados, especialmente con muestras pequeñas; la implementación computacional requiere muestreo de Monte Carlo por cadenas de Markov (MCMC) que es lento con datos grandes; la interpretación y el reporte requieren familiaridad con conceptos bayesianos (distribución posterior, intervalos de credibilidad, factores de Bayes) que no forman parte de la formación estadística estándar en pregrado veterinario; y la revisión por pares en revistas veterinarias todavía está dominada por el marco frecuentista, lo que puede generar resistencia.

8.6.4. Conexión con los métodos clásicos

Los métodos bayesianos no paramétricos no reemplazan las pruebas de este libro; las complementan. Una prueba de Mann-Whitney sigue siendo la herramienta más directa para responder "¿difieren las distribuciones de estos dos grupos?" cuando se quiere un análisis rápido, transparente y que cualquier revisor pueda verificar. Un modelo bayesiano no paramétrico es la herramienta cuando se quiere estimar la forma completa de cada distribución, cuantificar la probabilidad de que la diferencia sea clínicamente relevante (no solo estadísticamente significativa), o incorporar información de estudios previos. Son preguntas distintas que requieren herramientas distintas.

8.7. Machine learning como extensión no paramétrica

8.7.1. ¿Qué tiene que ver el machine learning con la estadística no paramétrica?

La conexión es más directa de lo que parece. Un modelo de regresión lineal es paramétrico: asume que la relación entre las variables es una línea recta y estima dos parámetros (pendiente e intercepto). Un spline o un GAM son no paramétricos: no asumen una forma funcional específica y dejan que los datos determinen la relación. Un random forest lleva esta lógica un paso más allá: construye cientos de árboles de decisión, cada uno de los cuales particiona el espacio de las variables predictoras sin asumir ninguna forma funcional. Es, en esencia, un método no paramétrico con esteroides computacionales.

La diferencia central entre la estadística no paramétrica clásica y el machine learning no está en la matemática subyacente (ambos evitan supuestos distribucionales rígidos), sino en el objetivo: la estadística clásica busca inferencia (¿hay una diferencia? ¿cuál es su magnitud? ¿es generalizable?), mientras que el machine learning busca predicción (¿puedo predecir correctamente la categoría o el valor de una observación nueva?). La misma investigación veterinaria puede necesitar ambas cosas: un ensayo clínico necesita inferencia para decidir si un tratamiento es eficaz; un sistema de detección temprana de mastitis necesita predicción para alertar al ganadero antes de que la vaca muestre signos clínicos evidentes.

8.7.2. Random forests

Un random forest (Breiman, 2001) construye un gran número de árboles de decisión (típicamente 500 a 2000), cada uno entrenado con una muestra bootstrap del conjunto de datos original y un subconjunto aleatorio de las variables predictoras en cada nodo de decisión. La predicción final es el promedio (para regresión) o el voto mayoritario (para clasificación) de todos los árboles.

¿Por qué funciona tan bien? Porque cada árbol individual es un modelo flexible que captura patrones no lineales, interacciones y umbrales sin que el investigador tenga que especificarlos, pero también es un modelo inestable: cambiando unos pocos datos, el árbol puede cambiar drásticamente. Al promediar cientos de árboles inestables pero diversos (gracias a la aleatorización), la varianza de la predicción se reduce sin aumentar el sesgo. Es un trueque de interpretabilidad por rendimiento predictivo.

En veterinaria, random forests se han aplicado para:

- Predecir la fertilidad de vacas lecheras a partir de variables productivas, reproductivas y metabólicas (Fenlon et al., 2017).
- Clasificar el estado sanitario de granjas porcinas a partir de indicadores de matadero (lesiones pulmonares, artritis, decomiso parcial o total).
- Identificar factores de riesgo para mastitis clínica integrando variables de manejo, instalaciones, clima y producción.
- Detectar bovinos enfermos a partir de datos de sensores de actividad y rumia (Dittrich et al., 2019).

La variable de importancia (*variable importance*) es la contribución de cada predictor al rendimiento del modelo. Se calcula de dos maneras: importancia por permutación (cuánto aumenta el error cuando se permutan aleatoriamente los valores de esa variable) e importancia por impureza (cuánto reduce la variable la impureza Gini o entropía promediada sobre todos los árboles). La primera es más confiable, especialmente cuando hay variables correlacionadas.

8.7.3. Gradient boosting

Gradient boosting (Friedman, 2001) toma un camino distinto al random forest. En lugar de construir muchos árboles independientes y promediarlos, construye árboles secuencialmente: el primer árbol ajusta los datos, el segundo ajusta los residuos del primero, el tercero ajusta los residuos del segundo, y así sucesivamente. Cada árbol nuevo corrige los errores de los anteriores. El proceso es más lento, pero a menudo produce predicciones más precisas que random forest, especialmente con conjuntos de datos grandes.

Las implementaciones más usadas son XGBoost (Chen & Guestrin, 2016) y LightGBM (Ke et al., 2017). En veterinaria, gradient boosting se ha utilizado para predecir la mortalidad en unidades de cuidados intensivos de hospitales veterinarios, clasificar razas a partir de medidas morfométricas, y modelar la distribución geográfica de vectores de enfermedades en función de variables climáticas y de uso de suelo.

8.7.4. Validación cruzada: evaluar la predicción con honestidad

Un error grave y frecuente en la aplicación de machine learning a datos veterinarios es evaluar el rendimiento del modelo con los mismos datos que se usaron para entrenarlo. Esto produce estimaciones de precisión infladas: el modelo "memoriza" los datos en lugar de aprender patrones generalizables. La validación cruzada resuelve este problema dividiendo los datos en k subconjuntos (*folds*): el modelo se entrena con $k-1$ subconjuntos y se evalúa con el subconjunto restante, rotando hasta que cada observación haya sido parte del conjunto de evaluación una vez.

Para datos veterinarios con muestras pequeñas (lo habitual), la validación cruzada con $k = 5$ o $k = 10$ subconjuntos es la práctica estándar. Con muestras muy pequeñas (< 30), la validación cruzada tipo *leave-one-out* (LOOCV) es una alternativa que maximiza el tamaño del conjunto de entrenamiento en cada iteración, aunque tiene mayor varianza en la estimación del error.

Las métricas de evaluación más relevantes para clasificación en veterinaria son:

- **Precisión global (*accuracy*):** proporción de predicciones correctas. Engañosa cuando las clases están desbalanceadas (si el 95 % de las granjas son sanas, un modelo que clasifique todas como sanas tendrá 95 % de precisión y será completamente inútil).
- **Sensibilidad y especificidad:** proporción de verdaderos positivos y verdaderos negativos, respectivamente. Relevantes en diagnóstico.
- **Área bajo la curva ROC (AUC):** medida global de la capacidad discriminativa, independiente del umbral de clasificación. Un AUC de 0.5 = azar; 0.7–0.8 = aceptable; 0.8–0.9 = bueno; > 0.9 = excelente.
- **Kappa de Cohen:** acuerdo entre la clasificación predicha y la observada, corregido por azar. Particularmente útil en clasificaciones ordinales (grados de lesión, estadios de enfermedad).

8.8. Datos masivos en veterinaria

8.8.1. Sensores y monitoreo continuo

La ganadería de precisión ha transformado la cantidad de datos disponibles por animal y por unidad de tiempo. Un collar con acelerómetro triaxial genera datos a frecuencias de 1 a 10 Hz: eso son entre 86 400 y 864 000 registros por animal por día. En un hato de 500 vacas lecheras, un mes de monitoreo produce cientos de millones de registros. Estos datos se usan para detectar estros (cambios en patrones de actividad), enfermedades (reducción de tiempo de rumia), y bienestar animal (tiempo de descanso, comportamiento de locomoción).

El análisis de estos datos no se resuelve con una prueba de Kruskal-Wallis. Se requieren métodos de series temporales, clasificación automática de comportamientos (random forests, redes neuronales), y detección de anomalías. Sin embargo, cuando el investigador quiere comparar el tiempo de rumia entre dos grupos de tratamiento, necesita primero extraer un resumen por animal (mediana diaria de minutos de rumia, por ejemplo) y luego puede aplicar las pruebas no paramétricas clásicas a esos resúmenes. La combinación de machine learning para procesar datos crudos y estadística clásica para la inferencia final es un flujo de trabajo cada vez más común.

8.8.2. Metagenómica y datos composicionales

La secuenciación del gen 16S rRNA o la metagenómica shotgun producen tablas de abundancia de miles de taxones (OTUs o ASVs) en cada muestra. Estos datos tienen propiedades que los hacen difíciles de analizar con métodos paramétricos: son composicionales (las abundancias relativas suman 1, de modo que un aumento en un taxón implica una disminución aparente en otros), contienen muchos ceros (la mayoría de los taxones están ausentes en la mayoría de las muestras), y la distribución de abundancias es extremadamente sesgada.

Los métodos no paramétricos multivariados del Capítulo 7 (PERMANOVA, ANOSIM, NMDS) son las herramientas estándar para analizar estos datos a nivel de comunidad: ¿difiere la composición global del microbiota entre grupos? Para análisis a nivel de taxón individual (¿cuáles taxones difieren entre grupos?), se usan pruebas como ANCOM-BC (Lin & Peddada, 2020) o DESeq2 adaptado a datos de metagenómica, que incorporan correcciones específicas para la composicionalidad y la multiplicidad de pruebas.

En el contexto veterinario, la metagenómica se aplica al microbioma ruminal (eficiencia alimenticia, producción de metano), al microbioma intestinal de aves de corral (relación con productividad y resistencia a patógenos), al microbioma de heridas crónicas, y al microbioma del tracto reproductivo en bovinos.

8.8.3. Registros clínicos electrónicos

Los hospitales veterinarios universitarios y las redes de clínicas privadas generan registros clínicos electrónicos que, cuando se agregan, constituyen bases de datos con miles o decenas de miles de casos. Estos registros incluyen diagnósticos codificados (o texto libre que puede codificarse), resultados de laboratorio, tratamientos, costos, desenlaces clínicos y datos demográficos del paciente (especie, raza, edad, sexo, estado reproductivo).

El análisis de estas bases de datos presenta desafíos específicos: datos faltantes (no a todos los pacientes se les hizo el mismo panel de laboratorio), codificación inconsistente (un mismo diagnóstico puede aparecer con nombres distintos en registros de diferentes clínicos), selección de casos (los pacientes que llegan al hospital universitario no son representativos de la población

general de mascotas), y dependencia temporal (los protocolos de diagnóstico y tratamiento cambian a lo largo de los años).

Los métodos no paramétricos son particularmente útiles en estos contextos porque los datos de registros clínicos rara vez cumplen supuestos paramétricos: las distribuciones de valores de laboratorio suelen ser sesgadas, las variables categóricas son frecuentes (grado de enfermedad, tipo de lesión, respuesta al tratamiento), y los tamaños de muestra por categoría son muy desiguales. Pero cuando el objetivo es predictivo (¿qué factores predicen la supervivencia de un perro con linfoma multicentrico?), los métodos de machine learning permiten explorar interacciones y no linealidades que los modelos clásicos no capturan.

8.9. La estadística no paramétrica clásica en la era del machine learning: complemento, no obsolescencia

8.9.1. Inferencia vs. predicción

La distinción entre inferencia y predicción no es solo filosófica; tiene consecuencias prácticas directas en la investigación veterinaria:

Cuando el objetivo es determinar si un tratamiento antihelmíntico nuevo reduce la carga parasitaria más que el tratamiento estándar, se necesita una prueba de hipótesis con control del error tipo I, un estimado del tamaño del efecto y un intervalo de confianza. La pregunta es causal e inferencial. Mann-Whitney, Wilcoxon, Kruskal-Wallis con sus tamaños del efecto y sus intervalos de confianza son las herramientas diseñadas para esto.

Cuando el objetivo es construir un sistema que prediga si una vaca desarrollará cetosis subclínica en los próximos 7 días a partir de datos de sensores, producción y laboratorio, lo que importa es la capacidad predictiva del modelo: sensibilidad, especificidad, valor predictivo positivo. Random forest o gradient boosting son las herramientas diseñadas para esto.

Confundir los objetivos lleva a errores costosos. Aplicar un random forest a un ensayo clínico controlado y reportar que “las variables más importantes fueron el tratamiento y la edad” no es inferencia; no hay valor p, no hay control del error tipo I, no hay prueba de que la asociación sea causal. Aplicar Mann-Whitney a un problema de predicción y reportar que “la diferencia en el

nivel de beta-hidroxibutirato fue significativa entre vacas sanas y enfermas" no dice nada sobre si ese nivel permite predecir la enfermedad con precisión útil en la granja.

8.9.2. Lo que la estadística clásica aporta y el machine learning no

La estadística no paramétrica clásica ofrece:

- **Control explícito del error tipo I.** Cuando se rechaza la hipótesis nula con $p < 0.05$, se sabe que, si la hipótesis nula fuera cierta, la probabilidad de haber obtenido ese resultado (o uno más extremo) por azar es menor al 5 %. Los modelos de machine learning no tienen un mecanismo equivalente integrado.
- **Tamaños del efecto estandarizados** que permiten comparar la magnitud de los efectos entre estudios y realizar meta-análisis.
- **Transparencia y reproducibilidad.** Una prueba de Mann-Whitney aplicada a los mismos datos siempre dará el mismo resultado. Un random forest con semilla aleatoria diferente puede dar importancias de variables ligeramente distintas.
- **Parsimonia.** Con 15 animales por grupo, una prueba de rangos es todo lo que se necesita y todo lo que los datos soportan. Ajustar un random forest con 500 árboles a 30 observaciones es como usar un tomógrafo para diagnosticar una fractura de dedo.

8.9.3. Lo que el machine learning aporta y la estadística clásica no

El machine learning ofrece:

- **Capacidad para manejar muchas variables predictoras simultáneamente**, incluyendo interacciones y no linealidades, sin que el investigador tenga que especificar cada término.
- **Rendimiento predictivo superior** en problemas complejos con muchas variables y relaciones no lineales.
- **Descubrimiento de patrones** en datos de alta dimensionalidad (ómicas, sensores, imágenes) que los métodos clásicos no pueden explorar.
- **Escalabilidad** a grandes volúmenes de datos.

8.9.4. La síntesis práctica

En la investigación veterinaria contemporánea, el flujo de trabajo más sólido combina ambos enfoques:

1. Machine learning (o métodos exploratorios no paramétricos como NMDS) para explorar datos de alta dimensionalidad, identificar variables candidatas y generar hipótesis.
2. Estadística clásica (incluyendo métodos no paramétricos) para probar esas hipótesis con control del error, estimar tamaños del efecto y producir la evidencia que respalde decisiones clínicas o productivas.
3. Machine learning para construir herramientas predictivas operacionales (sistemas de alerta temprana, clasificación automática) una vez que la evidencia causal está establecida.

Los tres niveles no son secuenciales en sentido estricto; pueden operar en paralelo. Pero confundirlos usar herramientas del nivel 3 para responder preguntas del nivel 2, o herramientas del nivel 2 para resolver problemas del nivel 3 lleva a resultados técnicamente correctos, pero científicamente inadecuados.

8.10. Implementación en R

8.10.1. Paquetes requeridos

```
# ——— Paquetes para reporte y tablas ——————————————————————————
# install.packages(c("gtsummary", "flextable", "ggpubr", "broom",
#                    "knitr", "kableExtra", "rmarkdown"))
library(gtsummary)
library(flextable)
library(ggpubr)
library(broom)
library(dplyr)
library(ggplot2)

# ——— Paquetes para machine learning ——————————————————————————
# install.packages(c("randomForest", "ranger", "tidymodels", "vip"))
library(randomForest)
library(ranger)
library(tidymodels)
library(vip)

# ——— Paquetes para métodos bayesianos (solo demostración conceptual)
```

```r
# install.packages("brms")   # requiere rstan o cmdstanr
# library(brms)

# ——— Paquetes para gráficos de sobrevivencia ————————————————————
# install.packages(c("survival", "survminer"))
library(survival)
library(survminer)

# ——— Paquetes para análisis no paramétricos previos ——————————————
library(coin)
library(rcompanion)
library(PMCMRplus)
```

8.10.2. Tablas de resultados automatizadas con *gtsummary*

El paquete *gtsummary* (Sjoberg et al., 2021) genera tablas de estadísticos descriptivos y comparaciones entre grupos con una sintaxis concisa. Lo especialmente útil para datos no paramétricos es que detecta automáticamente variables no normales (aplicando un test de Shapiro-Wilk interno) y reporta mediana con RIQ en lugar de media ± DE. Pero siempre conviene controlar este comportamiento explícitamente.

```r
# ——— Ejemplo: Tabla comparativa de dos grupos ————————————————————
# Datos simulados de HPG en bovinos Brahman vs. Holstein
set.seed(42)
datos_hpg <- data.frame(
  raza = rep(c("Brahman", "Holstein"), each = 20),
  hpg = c(
    # Brahman: distribución sesgada con mediana baja
    rnbinom(20, size = 2, mu = 500),
    # Holstein: distribución sesgada con mediana alta
    rnbinom(20, size = 2, mu = 1300)
  ),
  hematocrito = c(
    rnorm(20, mean = 32, sd = 3),
    rnorm(20, mean = 30, sd = 3.5)
  ),
  eosinofilos = c(
    rlnorm(20, meanlog = log(0.8), sdlog = 0.5),
    rlnorm(20, meanlog = log(1.5), sdlog = 0.5)
  )
)

# Tabla con gtsummary: mediana (RIQ) y prueba de Mann-Whitney
tabla_hpg <- datos_hpg %>%
```

```
  tbl_summary(
    by = raza,                  # variable de agrupación
    statistic = list(
      all_continuous() ~ "{median} ({p25}–{p75})"
      # Fuerza mediana y RIQ para todas las variables continuas
    ),
    digits = list(
      hpg ~ 0,                # HPG sin decimales
      hematocrito ~ 1,        # hematocrito con un decimal
      eosinofilos ~ 2         # eosinófilos con dos decimales
    ),
    label = list(
      hpg ~ "HPG (McMaster)",
      hematocrito ~ "Hematocrito (%)",
      eosinofilos ~ "Eosinófilos (×10³/µL)"
    )
  ) %>%
  add_p(
    test = list(all_continuous() ~ "wilcox.test"),
    # Fuerza Mann-Whitney para todas las variables continuas
    pvalue_fun = ~ style_pvalue(.x, digits = 3)
    # Valores p con tres decimales
  ) %>%
  add_overall() %>%             # columna de totales
  modify_header(label ~ "**Variable**") %>%
  bold_labels()

# Imprimir la tabla en consola
tabla_hpg

# ——— Exportar a Word con flextable ———————————————————————————
# gtsummary se convierte a flextable para exportación
tabla_flex <- tabla_hpg %>%
  as_flex_table()

# Guardar como documento Word
# library(officer)
# doc <- read_docx() %>%
#   body_add_flextable(tabla_flex)
# print(doc, target = "tabla_resultados.docx")
```

Interpretación de la salida: *gtsummary* produce una tabla formateada que incluye, para cada variable, la mediana con el rango intercuartílico entre paréntesis, separada por grupo. La columna de valor p corresponde a la prueba de Mann-Whitney (bilateral). La función *add_p()* permite especificar la prueba; si no se especifica, *gtsummary* selecciona automáticamente en función de la

distribución de los datos, pero esa selección automática no siempre coincide con la intención del investigador, por lo que es preferible ser explícito.

8.10.3. Tabla para Kruskal-Wallis con comparaciones post hoc

```
# ——— Datos simulados: tiempo de recuperación en tres protocolos ———
set.seed(123)
datos_anest <- data.frame(
  protocolo = factor(rep(c("Propofol+Iso", "Til-Zol", "Propofol+Sevo"),
                         each = 12),
                     levels = c("Propofol+Iso", "Til-Zol", "Propofol+Sevo")),
  tiempo_rec = c(
    rnorm(12, mean = 18, sd = 5),    # Propofol + isoflurano
    rnorm(12, mean = 28, sd = 6),    # Tiletamina-zolazepam
    rnorm(12, mean = 16, sd = 4)     # Propofol + sevoflurano
  ) %>% pmax(5)  # asegurar valores positivos
)

# Tabla con gtsummary para tres grupos
tabla_anest <- datos_anest %>%
  tbl_summary(
    by = protocolo,
    statistic = list(
      all_continuous() ~ "{median} ({p25}–{p75})"
    ),
    label = list(
      tiempo_rec ~ "Tiempo de recuperación (min)"
    ),
    digits = list(tiempo_rec ~ 1)
  ) %>%
  add_p(
    test = list(all_continuous() ~ "kruskal.test"),
    pvalue_fun = ~ style_pvalue(.x, digits = 3)
  )

tabla_anest

# ——— Comparaciones post hoc de Dunn ———————————————————————————
—
library(PMCMRplus)
dunn_result <- kwAllPairsDunnTest(
  tiempo_rec ~ protocolo,
  data = datos_anest,
  p.adjust.method = "bonferroni"
)
```

```r
summary(dunn_result)

# Tamaño del efecto: epsilon-cuadrado
library(rcompanion)
epsilonSquared(
  x = datos_anest$tiempo_rec,
  g = datos_anest$protocolo
)
```

8.10.4. Figuras con *ggpubr*

El paquete *ggpubr* (Kassambara, 2023) simplifica la creación de gráficos listos para publicación, con anotaciones estadísticas integradas.

```r
# ——— Boxplot con puntos individuales y comparaciones estadísticas ——
# Comparación de HPG entre razas
fig_hpg <- ggboxplot(
  datos_hpg,
  x = "raza",
  y = "hpg",
  color = "raza",
  palette = c("#4E79A7", "#E15759"),    # colores distinguibles
  add = "jitter",                       # puntos individuales
  add.params = list(size = 2, alpha = 0.6),
  xlab = "Raza",
  ylab = "HPG (McMaster)",
  title = ""
) +
  stat_compare_means(
    method = "wilcox.test",
    label = "p.format",           # muestra el valor p formateado
    label.x = 1.5,                # posición horizontal del texto
    label.y = max(datos_hpg$hpg) * 1.1  # posición vertical
  ) +
  theme(legend.position = "none")   # la leyenda es redundante aquí

fig_hpg

# ——— Boxplot para tres grupos con comparaciones post hoc ——————————
# Necesitamos definir las comparaciones de interés
comparaciones <- list(
  c("Propofol+Iso", "Til-Zol"),
  c("Til-Zol", "Propofol+Sevo"),
  c("Propofol+Iso", "Propofol+Sevo")
)
```

```r
fig_anest <- ggboxplot(
  datos_anest,
  x = "protocolo",
  y = "tiempo_rec",
  color = "protocolo",
  palette = c("#4E79A7", "#E15759", "#59A14F"),
  add = "jitter",
  add.params = list(size = 2, alpha = 0.6),
  xlab = "Protocolo anestésico",
  ylab = "Tiempo de recuperación (min)"
) +
  stat_compare_means(
    comparisons = comparaciones,
    method = "wilcox.test",
    p.adjust.method = "bonferroni",
    label = "p.signif"            # asteriscos: ns, *, **, ***
  ) +
  stat_compare_means(
    method = "kruskal.test",
    label.y = max(datos_anest$tiempo_rec) * 1.3
    # Resultado global de Kruskal-Wallis arriba
  ) +
  theme(legend.position = "none")

fig_anest

# ——— Guardar figuras ———————————————————————————————————————
ggsave("fig_hpg_boxplot.tiff", fig_hpg,
       width = 5, height = 4, dpi = 300, compression = "lzw")
ggsave("fig_anest_boxplot.tiff", fig_anest,
       width = 7, height = 5, dpi = 300, compression = "lzw")
```

8.10.5. Curva de Kaplan-Meier con *survminer*

```r
# ——— Datos simulados: sobrevivencia post-mastocitoma ———————————————
set.seed(456)
n_quimio <- 26
n_observ <- 22

datos_surv <- data.frame(
  grupo = c(rep("Vinblastina-Pred", n_quimio),
            rep("Observación", n_observ)),
  tiempo = c(
    rweibull(n_quimio, shape = 1.2, scale = 300),
    rweibull(n_observ, shape = 1.2, scale = 180)
```

```
  ) %>% round(0),
  evento = c(
    rbinom(n_quimio, 1, 0.65),   # 65 % recidiva en grupo quimio
    rbinom(n_observ, 1, 0.82)    # 82 % recidiva en observación
  )
)

# Ajustar objeto de sobrevivencia
surv_obj <- survfit(Surv(tiempo, evento) ~ grupo, data = datos_surv)

# Gráfico de Kaplan-Meier con survminer
fig_km <- ggsurvplot(
  surv_obj,
  data = datos_surv,
  pval = TRUE,                  # valor p del log-rank
  pval.method = TRUE,           # nombre de la prueba
  conf.int = TRUE,              # bandas de confianza
  risk.table = TRUE,            # tabla de riesgo debajo
  risk.table.col = "strata",    # colores por grupo
  ggtheme = theme_bw(),
  palette = c("#4E79A7", "#E15759"),
  xlab = "Tiempo (días)",
  ylab = "Probabilidad de sobrevivencia libre de recidiva",
  legend.title = "Grupo",
  legend.labs = c("Observación", "Vinblastina-Pred"),
  surv.median.line = "hv"       # líneas de mediana horizontal/vertical
)

fig_km

# Mediana de sobrevivencia con IC 95 %
summary(surv_obj)$table
```

8.10.6. Random forest para clasificación con *tidymodels*

```
# ——— Ejemplo: clasificar estado sanitario de granjas porcinas ————————
# Datos simulados basados en indicadores de matadero

set.seed(789)
n_granjas <- 200

datos_granjas <- data.frame(
  # Variables predictoras (indicadores de matadero)
  pct_neumonia = rbeta(n_granjas, 2, 5) * 100,
  pct_pleuritis = rbeta(n_granjas, 1.5, 8) * 100,
  pct_artritis = rbeta(n_granjas, 1, 10) * 100,
  pct_decomiso = rbeta(n_granjas, 1.5, 20) * 100,
```

```
  indice_cola = rbeta(n_granjas, 2, 6) * 100,  # lesiones de cola
  peso_canal = rnorm(n_granjas, mean = 85, sd = 8),
  espesor_grasa = rnorm(n_granjas, mean = 14, sd = 3),
  densidad = rpois(n_granjas, lambda = 15),     # animales/m²
  n_lotes_anio = rpois(n_granjas, lambda = 6)
)

# Variable respuesta: estado sanitario (bueno vs. deficiente)
# Basada en una combinación no lineal de predictoras
prob_deficiente <- with(datos_granjas, plogis(
  -3 + 0.05 * pct_neumonia + 0.08 * pct_pleuritis +
    0.1 * pct_artritis + 0.15 * pct_decomiso +
    0.03 * indice_cola - 0.02 * peso_canal +
    0.05 * densidad
))
datos_granjas$estado <- factor(
  ifelse(rbinom(n_granjas, 1, prob_deficiente) == 1,
         "Deficiente", "Bueno"),
  levels = c("Bueno", "Deficiente")
)

table(datos_granjas$estado)

# ——— Flujo con tidymodels ————————————————————————————————————————
———

# 1. Dividir datos en entrenamiento (75 %) y prueba (25 %)
set.seed(101)
division <- initial_split(datos_granjas, prop = 0.75, strata = estado)
entrenamiento <- training(division)
prueba <- testing(division)

# 2. Receta de preprocesamiento (mínimo necesario)
receta <- recipe(estado ~ ., data = entrenamiento)

# 3. Especificación del modelo: random forest con ranger
modelo_rf <- rand_forest(
  mtry = tune(),       # número de variables por nodo (a optimizar)
  trees = 500,         # número de árboles
  min_n = tune()       # mínimo de observaciones por nodo terminal
) %>%
  set_engine("ranger", importance = "permutation") %>%
  set_mode("classification")

# 4. Flujo de trabajo
flujo <- workflow() %>%
  add_recipe(receta) %>%
```

```r
  add_model(modelo_rf)

# 5. Validación cruzada para optimizar hiperparámetros
set.seed(202)
cv_folds <- vfold_cv(entrenamiento, v = 5, strata = estado)

# Grilla de búsqueda
grilla <- grid_regular(
  mtry(range = c(2, 6)),     # entre 2 y 6 variables por nodo
  min_n(range = c(3, 15)),   # entre 3 y 15 observaciones mínimas
  levels = 4                 # 4 niveles por hiperparámetro
)

# Optimización
resultados_cv <- tune_grid(
  flujo,
  resamples = cv_folds,
  grid = grilla,
  metrics = metric_set(roc_auc, accuracy, sensitivity, specificity)
)

# Mejores hiperparámetros
show_best(resultados_cv, metric = "roc_auc")
mejores_hp <- select_best(resultados_cv, metric = "roc_auc")
mejores_hp

# 6. Finalizar el modelo con los mejores hiperparámetros
flujo_final <- flujo %>%
  finalize_workflow(mejores_hp)

# 7. Ajustar al conjunto completo de entrenamiento
modelo_final <- flujo_final %>%
  fit(entrenamiento)

# 8. Evaluar en el conjunto de prueba
predicciones <- modelo_final %>%
  predict(prueba, type = "prob") %>%
  bind_cols(predict(modelo_final, prueba)) %>%
  bind_cols(prueba %>% select(estado))

# Métricas
metricas <- predicciones %>%
  metrics(truth = estado, estimate = .pred_class, .pred_Deficiente)
metricas

# Matriz de confusión
predicciones %>%
```

```
  conf_mat(truth = estado, estimate = .pred_class)

# Curva ROC
predicciones %>%
  roc_curve(truth = estado, .pred_Deficiente) %>%
  autoplot()

# 9. Importancia de variables
modelo_ajustado <- modelo_final %>%
  extract_fit_parsnip()

# Gráfico de importancia
vip(modelo_ajustado, num_features = 9, aesthetics = list(fill = "#4E79A7")) +
  labs(title = "Importancia de variables (permutación)",
       x = "Variable", y = "Importancia") +
  theme_minimal()
```

Interpretación: El flujo de *tidymodels* divide los datos, define la receta de preprocesamiento, especifica el modelo, optimiza hiperparámetros mediante validación cruzada, y evalúa el rendimiento en datos no vistos. La importancia de variables por permutación indica cuáles indicadores de matadero contribuyen más a la clasificación del estado sanitario. Un AUC alto sugiere que los indicadores de matadero contienen información suficiente para discriminar granjas con problemas sanitarios, lo que tiene implicaciones prácticas para programas de vigilancia epidemiológica basados en datos de planta de faena.

8.10.7. Modelo bayesiano no paramétrico con *brms* (demostración conceptual)

```
# ——— Regresión bayesiana con splines (semi-no-paramétrica) ——————————
# Nota: requiere que rstan o cmdstanr estén instalados y funcionales
# Este código es demostrativo; la ejecución completa puede tomar
# varios minutos dependiendo del hardware

# library(brms)
#
# # Datos: producción de leche vs. temperatura ambiental
# set.seed(303)
# datos_leche <- data.frame(
#   temp = runif(80, min = 18, max = 38),
#   leche = NA
# )
# # Relación no lineal: producción estable hasta ~28°C, luego descenso
# datos_leche$leche <- with(datos_leche, {
#   base <- 22 - 0.1 * (temp - 28)^2 * (temp > 28)
```

```
#   base + rnorm(80, 0, 2)
# })
#
# # Modelo bayesiano con spline
# modelo_bayes <- brm(
#   leche ~ s(temp, k = 8),       # spline con 8 nudos
#   data = datos_leche,
#   family = gaussian(),
#   prior = c(
#     prior(normal(20, 5), class = "Intercept"),
#     prior(student_t(3, 0, 2), class = "sds")
#   ),
#   chains = 4,
#   iter = 2000,
#   warmup = 1000,
#   seed = 404
# )
#
# # Resumen del modelo
# summary(modelo_bayes)
#
# # Gráfico del efecto suavizado con intervalo de credibilidad
# conditional_effects(modelo_bayes, effects = "temp")
#
# # Verificación de convergencia: R-hat y tamaño efectivo de muestra
# # Todos los R-hat deben ser < 1.01; todos los ESS > 400
```

Nota para el lector: El código anterior está comentado porque requiere la instalación de Stan (un lenguaje de programación probabilística) a través de *rstan* o *cmdstanr*, lo cual puede ser complejo en algunas plataformas. El propósito aquí es mostrar la sintaxis y la lógica del enfoque, no exigir su ejecución en clase. Para quienes deseen profundizar, la documentación de *brms* (Bürkner, 2017) ofrece guías de instalación detalladas.

8.10.8. Generación de un reporte reproducible con R Markdown

```
# ——— Estructura de un archivo .Rmd para reporte veterinario ————————
# Este bloque no se ejecuta; muestra la estructura del archivo

# ---
# title: "Comparación de protocolos anestésicos en perras:
#         análisis no paramétrico"
# author: "Dr. [Nombre]"
# date: "`r Sys.Date()`"
# output:
```

```
#   word_document:
#     reference_docx: plantilla_revista.docx
# bibliography: referencias.bib
# csl: apa-7th.csl
# ---
#
# ```{r setup, include=FALSE}
# knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)
# library(gtsummary)
# library(ggpubr)
# library(flextable)
# ```
#
# ## Resultados
#
# Las características demográficas y quirúrgicas de las 36 perras
# incluidas en el estudio se presentan en la Tabla 1.
#
# ```{r tabla1}
# datos %>%
#   tbl_summary(by = protocolo) %>%
#   add_p() %>%
#   as_flex_table()
# ```
#
# El tiempo de recuperación posanestésica difirió significativamente
# entre los tres protocolos (Tabla 2, Figura 1).
#
# ```{r figura1, fig.cap="Tiempo de recuperación según protocolo..."}
# fig_anest
# ```
```

8.11. Estudio de caso integrador

Parte A: Sección de resultados completa para un artículo sobre protocolos de anestesia en perros

Contexto del problema

Se diseñó un ensayo clínico controlado aleatorizado para comparar tres protocolos anestésicos en perras sometidas a ovariohisterectomía electiva en el Hospital Veterinario de la Universidad San Francisco de Quito (USFQ), Ecuador, durante el período de marzo a agosto de 2024. Los tres protocolos fueron: (1) inducción con propofol y mantenimiento con isoflurano (PI, n = 12), (2)

inducción y mantenimiento con tiletamina-zolazepam intramuscular (TZ, n = 12), y (3) inducción con propofol y mantenimiento con sevoflurano (PS, n = 12). Se incluyeron perras clínicamente sanas (ASA I-II), de entre 1 y 5 años de edad, con pesos entre 5 y 25 kg, de razas diversas. Las variables respuesta fueron: tiempo de recuperación posanestésica (minutos desde la interrupción del anestésico hasta la bipedestación sostenida), puntaje de dolor postoperatorio en escala analógica visual modificada (0–4) a las 2, 6 y 12 horas, e incidencia de vómito en las primeras 24 horas.

Justificación del enfoque no paramétrico

El tiempo de recuperación presentó una distribución sesgada a la derecha (prueba de Shapiro-Wilk: grupo PI, W = 0.89, p = 0.112; grupo TZ, W = 0.86, p = 0.046; grupo PS, W = 0.91, p = 0.196). Dado que al menos un grupo no cumplió el supuesto de normalidad y las muestras fueron pequeñas (n = 12 por grupo), se optó por la prueba de Kruskal-Wallis en lugar de ANOVA de una vía. El puntaje de dolor es una variable ordinal medida en escala discreta de 0 a 4, lo que excluye el uso de pruebas paramétricas por naturaleza de la variable, no solo por distribución. La comparación entre tiempos (2, 6 y 12 h) dentro de cada grupo se evaluó con la prueba de Friedman. La incidencia de vómito (variable dicotómica) se analizó con la prueba exacta de Fisher-Freeman-Halton, dado que tres de las seis celdas presentaron frecuencias esperadas inferiores a 5.

Datos

```
# ——— Datos del caso integrador ——————————————————————————————
set.seed(2024)

# Tiempo de recuperación (minutos)
caso_datos <- data.frame(
  id_perra = 1:36,
  protocolo = factor(rep(c("PI", "TZ", "PS"), each = 12),
                     levels = c("PI", "TZ", "PS")),
  peso_kg = round(runif(36, 5, 25), 1),
  edad_anios = sample(1:5, 36, replace = TRUE),
  tiempo_rec = c(
    round(rlnorm(12, meanlog = log(18), sdlog = 0.3)),
    round(rlnorm(12, meanlog = log(28), sdlog = 0.25)),
    round(rlnorm(12, meanlog = log(16), sdlog = 0.28))
  ),
```

```
  dolor_2h = c(
    sample(2:4, 12, replace = TRUE, prob = c(0.2, 0.5, 0.3)),
    sample(2:4, 12, replace = TRUE, prob = c(0.1, 0.3, 0.6)),
    sample(1:3, 12, replace = TRUE, prob = c(0.2, 0.5, 0.3))
  ),
  dolor_6h = c(
    sample(1:3, 12, replace = TRUE, prob = c(0.3, 0.5, 0.2)),
    sample(1:4, 12, replace = TRUE, prob = c(0.1, 0.2, 0.4, 0.3)),
    sample(0:2, 12, replace = TRUE, prob = c(0.3, 0.4, 0.3))
  ),
  dolor_12h = c(
    sample(0:2, 12, replace = TRUE, prob = c(0.4, 0.4, 0.2)),
    sample(1:3, 12, replace = TRUE, prob = c(0.2, 0.4, 0.4)),
    sample(0:1, 12, replace = TRUE, prob = c(0.6, 0.4))
  ),
  vomito = c(
    rbinom(12, 1, 0.17),    # PI: ~17 % vómito
    rbinom(12, 1, 0.58),    # TZ: ~58 % vómito
    rbinom(12, 1, 0.08)     # PS: ~8 % vómito
  )
)

# ——— Análisis completo ———————————————————————————————————————
———

# 1. Descriptivos por grupo
caso_datos %>%
  group_by(protocolo) %>%
  summarise(
    n = n(),
    md_tiempo = median(tiempo_rec),
    q1_tiempo = quantile(tiempo_rec, 0.25),
    q3_tiempo = quantile(tiempo_rec, 0.75),
    md_dolor2h = median(dolor_2h),
    md_dolor6h = median(dolor_6h),
    md_dolor12h = median(dolor_12h),
    pct_vomito = mean(vomito) * 100
  )

# 2. Kruskal-Wallis para tiempo de recuperación
kw_tiempo <- kruskal.test(tiempo_rec ~ protocolo, data = caso_datos)
kw_tiempo
# Estadístico H, grados de libertad y valor p

# Tamaño del efecto: epsilon-cuadrado
library(rcompanion)
epsilonSquared(
```

```
  x = caso_datos$tiempo_rec,
  g = caso_datos$protocolo
)

# Comparaciones post hoc de Dunn
library(PMCMRplus)
dunn_tiempo <- kwAllPairsDunnTest(
  tiempo_rec ~ protocolo,
  data = caso_datos,
  p.adjust.method = "holm"
)
summary(dunn_tiempo)

# 3. Friedman para puntaje de dolor por protocolo
#    Reorganizar datos en formato largo para cada grupo

# Ejemplo para el grupo PI
dolor_pi <- caso_datos %>%
  filter(protocolo == "PI") %>%
  select(id_perra, dolor_2h, dolor_6h, dolor_12h) %>%
  tidyr::pivot_longer(
    cols = starts_with("dolor"),
    names_to = "tiempo",
    values_to = "puntaje"
  )

# Friedman dentro del grupo PI
friedman_pi <- friedman.test(
  puntaje ~ tiempo | id_perra,
  data = dolor_pi
)
friedman_pi

# W de Kendall como tamaño del efecto
kendallW <- friedman_pi$statistic / (12 * (3 - 1))
# W = chi²_F / (n * (k-1)) donde n = sujetos, k = condiciones
kendallW

# Post hoc: Conover-Iman
frdAllPairsConoverTest(
  puntaje ~ tiempo | id_perra,
  data = dolor_pi,
  p.adjust.method = "holm"
)

# Repetir para TZ y PS (mismo procedimiento)
```

```r
# 4. Fisher-Freeman-Halton para vómito
tabla_vomito <- table(caso_datos$protocolo, caso_datos$vomito)
colnames(tabla_vomito) <- c("No", "Sí")
tabla_vomito

fisher.test(tabla_vomito, simulate.p.value = FALSE)
# Para tablas 3x2, R calcula automáticamente la prueba exacta
# de Fisher-Freeman-Halton

# V de Cramér
library(rcompanion)
cramerV(tabla_vomito)

# ——— Tabla de resultados final con gtsummary ———————————————————

tabla_final <- caso_datos %>%
  select(protocolo, tiempo_rec, dolor_2h, dolor_6h, dolor_12h, vomito) %>%
  mutate(
    vomito = factor(vomito, levels = c(0, 1), labels = c("No", "Sí"))
  ) %>%
  tbl_summary(
    by = protocolo,
    statistic = list(
      all_continuous() ~ "{median} ({p25}–{p75})",
      all_categorical() ~ "{n} ({p}%)"
    ),
    label = list(
      tiempo_rec ~ "Tiempo de recuperación (min)",
      dolor_2h ~ "Dolor 2 h (0–4)",
      dolor_6h ~ "Dolor 6 h (0–4)",
      dolor_12h ~ "Dolor 12 h (0–4)",
      vomito ~ "Vómito 24 h"
    ),
    digits = list(
      tiempo_rec ~ 0,
      starts_with("dolor") ~ 0
    )
  ) %>%
  add_p(
    test = list(
      tiempo_rec ~ "kruskal.test",
      starts_with("dolor") ~ "kruskal.test",
      vomito ~ "fisher.test"
    ),
    pvalue_fun = ~ style_pvalue(.x, digits = 3)
  )
```

```
tabla_final

# ——— Boxplot con anotaciones ———————————————————————————————
fig_caso <- ggboxplot(
  caso_datos,
  x = "protocolo",
  y = "tiempo_rec",
  color = "protocolo",
  palette = c("#4E79A7", "#E15759", "#59A14F"),
  add = "jitter",
  add.params = list(size = 2.5, alpha = 0.6),
  xlab = "Protocolo anestésico",
  ylab = "Tiempo de recuperación (min)",
  legend = "none"
) +
  stat_compare_means(
    method = "kruskal.test",
    label.y = max(caso_datos$tiempo_rec) * 1.2
  ) +
  stat_compare_means(
    comparisons = list(c("PI", "TZ"), c("TZ", "PS"), c("PI", "PS")),
    method = "wilcox.test",
    p.adjust.method = "holm",
    label = "p.signif",
    tip.length = 0.02
  )

fig_caso
```

Redacción modelo de la sección de Resultados

A continuación, se presenta cómo se redactaría la sección de Resultados de este estudio en un manuscrito para una revista veterinaria indexada. Esta redacción debe servir de plantilla para el lector.

Resultados

Se incluyeron 36 perras con una mediana de peso de 14.3 kg (RIQ: 9.1–19.8) y una mediana de edad de 3 años (RIQ: 2–4), sin diferencias significativas entre grupos para ninguna de las dos variables (peso: $H(2) = 0.84$, $p = 0.657$; edad: $H(2) = 1.12$, $p = 0.572$). No se registraron complicaciones intraoperatorias ni muertes durante el periodo de estudio.

Tiempo de recuperación posanestésica. El tiempo de recuperación difirió significativamente entre protocolos (Kruskal-Wallis $H(2) = 14.82$, $p < 0.001$, $\varepsilon^2 = 0.43$). La mediana fue de 18 min (RIQ: 14–23) en PI, 28 min (RIQ: 24–34) en TZ y 16 min (RIQ: 13–20) en PS (Tabla 8.5, Figura 8.6). Las comparaciones post hoc (Dunn, ajuste de Holm) indicaron que TZ presentó tiempos significativamente más prolongados que PI (p_adj = 0.008) y PS (p_adj = 0.002), mientras que PI y PS no difirieron entre sí (p_adj = 0.512).

Puntaje de dolor postoperatorio. En el grupo PI, el puntaje de dolor disminuyó significativamente a lo largo de las tres evaluaciones (Friedman $\chi^2(2) = 18.17$, $p < 0.001$, $W = 0.76$). Las comparaciones post hoc (Conover-Iman, ajuste de Holm) mostraron diferencias entre las 2 h y las 12 h (p_adj < 0.001) y entre las 2 h y las 6 h (p_adj = 0.021), pero no entre las 6 h y las 12 h (p_adj = 0.089). En el grupo TZ, los puntajes también disminuyeron (Friedman $\chi^2(2) = 12.40$, $p = 0.002$, $W = 0.52$), con diferencias entre las 2 h y las 12 h (p_adj = 0.003). En el grupo PS, la disminución fue significativa (Friedman $\chi^2(2) = 15.89$, $p < 0.001$, $W = 0.66$), con la mayor reducción entre las 2 h y las 12 h (p_adj < 0.001).

Incidencia de vómito. La incidencia de vómito en las primeras 24 h fue del 16.7 % (2/12) en PI, 58.3 % (7/12) en TZ y 8.3 % (1/12) en PS. La prueba exacta de Fisher-Freeman-Halton indicó una asociación significativa entre el protocolo y la presencia de vómito ($p = 0.018$). La V de Cramér fue 0.50, indicando una asociación fuerte. TZ mostró una incidencia de vómito significativamente mayor que los otros dos protocolos.

Consideración sobre el análisis paramétrico alternativo. Si se hubiera aplicado ANOVA de una vía al tiempo de recuperación (ignorando el incumplimiento de normalidad en el grupo TZ y la asimetría de los datos), el resultado habría sido cualitativamente similar ($F(2,33) = 12.1$, $p < 0.001$). Sin embargo, la magnitud de las diferencias estimadas por las medias (PI: 19.2 ± 6.1, TZ: 29.4 ± 5.8, PS: 16.8 ± 4.9) habría exagerado las diferencias entre PI y PS porque las medias son más sensibles que las medianas a los valores extremos. Para el puntaje de dolor, un ANOVA de medidas repetidas habría sido inapropiado dado que la variable es ordinal con solo cinco niveles posibles (0–4), violando tanto la normalidad como la equidistancia entre categorías.

Parte B: Clasificación del estado sanitario de granjas porcinas con random forest

Contexto

Un programa de vigilancia sanitaria del Servicio Nacional de Sanidad Agropecuaria de Bolivia (SENASAG) recopiló datos de 200 lotes de cerdos de engorde provenientes de granjas tecnificadas del departamento de Santa Cruz, registrados durante la inspección post mortem en tres plantas faenadoras entre enero y diciembre de 2023. Para cada lote se registraron nueve indicadores de matadero: porcentaje de canales con neumonía, pleuritis, artritis, decomiso parcial o total, lesiones de cola (como indicador de bienestar), peso promedio de canal, espesor de grasa dorsal, densidad de alojamiento (animales/m²) y número de lotes faenados por granja en el año. Una comisión de veterinarios inspectores clasificó cada granja como "Bueno" o "Deficiente" según un protocolo estandarizado que integraba los indicadores de matadero con una visita técnica a la granja.

El objetivo fue evaluar si los indicadores de matadero, por sí solos, permitían predecir la clasificación sanitaria de la granja sin necesidad de la visita técnica, lo que reduciría costos y permitiría priorizar las visitas a las granjas con mayor probabilidad de clasificación deficiente.

Justificación del enfoque

Este es un problema de predicción, no de inferencia. No se pregunta "¿la neumonía está asociada con el estado sanitario?" (eso se respondería con chi-cuadrado o Mann-Whitney). Se pregunta "¿puedo predecir si la granja tiene estado deficiente a partir de los indicadores de matadero?". Random forest es apropiado porque: (a) puede manejar variables predictoras continuas y de conteo simultáneamente, (b) captura interacciones y no linealidades sin especificación previa, (c) proporciona una medida de importancia de variables que orienta la interpretación, y (d) es robusto a valores atípicos y distribuciones sesgadas, propiedades que comparte con la filosofía no paramétrica.

Resultados del modelo

El random forest con hiperparámetros optimizados (mtry = 4, min_n = 5) alcanzó un AUC de 0.83 (IC 95 % por bootstrap: 0.74–0.91) en el conjunto de prueba (n = 50). La precisión global fue del 78 %, con una sensibilidad del 72 % (capacidad de detectar granjas deficientes) y una especificidad del 82 % (capacidad de identificar granjas con buen estado).

Las tres variables con mayor importancia por permutación fueron: porcentaje de decomiso (importancia = 0.048), porcentaje de neumonía (importancia = 0.035) y porcentaje de pleuritis (importancia = 0.028). El peso de canal y el espesor de grasa tuvieron importancias cercanas a cero, sugiriendo que las variables productivas contribuyen poco a la clasificación sanitaria una vez que se consideran los indicadores patológicos.

Integración con métodos clásicos

La importancia de variables del random forest sugiere que el porcentaje de decomiso es el mejor predictor individual. Pero "mejor predictor" en un random forest no equivale a "asociación estadísticamente significativa". Para confirmar la asociación y cuantificar su magnitud, se complementa con análisis clásico:

```
# Mann-Whitney para % decomiso entre estados sanitarios
wilcox.test(pct_decomiso ~ estado, data = datos_granjas)
# Mediana y RIQ por grupo
datos_granjas %>%
  group_by(estado) %>%
  summarise(
    md = median(pct_decomiso),
    q1 = quantile(pct_decomiso, 0.25),
    q3 = quantile(pct_decomiso, 0.75)
  )
# Tamaño del efecto
library(rcompanion)
wilcoxonR(
  x = datos_granjas$pct_decomiso,
  g = datos_granjas$estado
)
```

Si Mann-Whitney confirma la diferencia con un tamaño del efecto grande ($r > 0.50$), se tiene una convergencia de evidencia: el random forest dice que el decomiso es el mejor predictor, y Mann-Whitney dice que la diferencia en decomiso entre granjas buenas y deficientes es grande y estadísticamente significativa. Ambas piezas de información son valiosas y complementarias; ninguna reemplaza a la otra.

8.12. Errores frecuentes y advertencias

Error 1: Reportar p < 0.05 sin el estadístico de prueba. Esto imposibilita la verificación, el meta-análisis y la evaluación del tamaño del efecto. El estadístico (U, H, T, χ^2, Z) debe acompañar siempre al valor p.

Error 2: Presentar medias para datos analizados con pruebas de rangos. Si se eligió Kruskal-Wallis porque los datos no eran normales, la medida descriptiva debe ser la mediana. Presentar medias contradice la lógica del análisis.

Error 3: No distinguir entre prueba exacta y asintótica. Con muestras pequeñas, el valor p exacto puede diferir sustancialmente del asintótico. El investigador debe reportar cuál usó y, cuando sea posible, preferir el exacto.

Error 4: Interpretar importancia de variables de random forest como inferencia causal. La importancia de variables indica contribución a la predicción, no causalidad. Una variable puede ser importante porque es proxy de otra variable no medida.

Error 5: Usar machine learning sin validación cruzada. Evaluar el modelo con los mismos datos de entrenamiento produce estimaciones infladas de rendimiento. La validación cruzada no es opcional; es el estándar mínimo.

Error 6: Abandonar la estadística clásica porque "ahora se usa machine learning". Los objetivos son distintos. Un ensayo clínico requiere pruebas de hipótesis con control del error. Un sistema de detección temprana requiere capacidad predictiva. El método se elige en función de la pregunta, no de la moda.

Error 7: No reportar los hiperparámetros del modelo de machine learning. Número de árboles, variables por nodo (mtry), profundidad máxima, tasa de aprendizaje: estos valores afectan los resultados y deben ser reportados para que el análisis sea reproducible.

Error 8: Gráficos de barras con barras de error para variables ordinales. Las barras de error simétricas asumen distribución simétrica. Las variables ordinales con pocos niveles no tienen distribución simétrica. Usar gráficos de frecuencias apiladas o boxplots.

8.13. Ejercicios de autoevaluación

Nivel básico

Ejercicio 8.1.

Un artículo reporta: “La diferencia fue significativa ($p < 0.05$, prueba no paramétrica)”. Identifique al menos cuatro piezas de información que faltan para que este reporte sea adecuado según APA 7.ª edición.

Respuesta: Falta: (1) el nombre específico de la prueba no paramétrica utilizada, (2) el valor del estadístico de prueba (*U*, *H*, *T*, etc.) con sus grados de libertad, (3) el valor p exacto (no la desigualdad “< 0.05”), (4) el tamaño del efecto, y (5) las medidas descriptivas (medianas y RIQ). Adicionalmente, falta contexto sobre las variables comparadas y los tamaños de muestra.

Ejercicio 8.2.

Un colega le muestra una tabla de resultados donde la variable “puntaje de condición corporal (1–5)” se presenta como “3.2 ± 0.8” y fue analizada con Kruskal-Wallis. ¿Qué está mal?

Respuesta: La medida descriptiva es incoherente con el análisis. Si se usó Kruskal-Wallis (prueba basada en rangos, apropiada para datos ordinales), la medida descriptiva debe ser la mediana con rango intercuartílico (por ejemplo, “Md = 3, RIQ: 2–4”), no la media ± desviación estándar. Además, el puntaje de condición corporal es ordinal con solo 5 niveles, por lo que la media de 3.2 implica una precisión espuria (sugiere que la escala es continua).

Ejercicio 8.3.

Clasifique cada una de las siguientes situaciones como un problema de inferencia o un problema de predicción:

a) Determinar si un antibiótico nuevo reduce la duración de la diarrea neonatal en terneros en comparación con el tratamiento estándar.
b) Construir un modelo que detecte vacas con mastitis subclínica a partir de datos de conductividad eléctrica de la leche.

c) Evaluar si la densidad de alojamiento está asociada con la mortalidad en pollos de engorde Ross 308.
d) Desarrollar una aplicación que prediga el sexo de crías de alpacas a partir de medidas ecográficas gestacionales.

Respuestas: (a) Inferencia. (b) Predicción. (c) Inferencia. (d) Predicción.

Nivel intermedio

Ejercicio 8.4.

La siguiente salida de R corresponde a una prueba de Mann-Whitney:

```
    Wilcoxon rank sum test with continuity correction

data:  hpg by raza
W = 87, p-value = 0.0003128
alternative hypothesis: true location shift is not equal to 0
```

Redacte el reporte de este resultado según APA 7.ª edición, asumiendo que *n*_Brahman = 22, *n*_Holstein = 19, la mediana de HPG en Brahman fue 450 (RIQ: 200–850) y en Holstein fue 1200 (RIQ: 600–2100). Calcule el tamaño del efecto *r* sabiendo que Z puede derivarse del valor p (Z = qnorm(p/2)).

Respuesta:

Z = qnorm(0.0003128 / 2) = −3.60 (en valor absoluto). $r = |Z| / \sqrt{(n_1 + n_2)} = 3.60 / \sqrt{41} = 3.60 / 6.40 = 0.56$.

Reporte: “El recuento de HPG fue significativamente mayor en bovinos Holstein (Md = 1200, RIQ: 600–2100) que en Brahman (Md = 450, RIQ: 200–850), $U = 87$, $p < 0.001$, $r = 0.56$. El tamaño del efecto fue grande según los criterios de Cohen (1988).”

Nota: R reporta el estadístico como *W*, que corresponde a la suma de rangos del primer grupo. Para reportar como *U* de Mann-Whitney, se usa $U = W - n_1(n_1+1)/2$. Algunos autores reportan directamente *W*; otros convierten a *U*. Ambos son aceptables si se identifica claramente cuál se reporta. En este caso se usa $U = 87$ directamente del output.

Ejercicio 8.5.

Observe la siguiente tabla y señale todos los errores de reporte:

Variable	Grupo A (n=10)	Grupo B (n=10)	p
Peso (kg)	25.3 ± 4.1	22.8 ± 3.9	< 0.05*
Condición corporal	3.1 ± 0.7	2.8 ± 0.9	NS
HPG	850 ± 620	1340 ± 890	0.04

*Se aplicó Mann-Whitney U para todas las variables.

Respuesta: Errores identificados:

1. Las tres variables se presentan con media ± DE pero se analizaron con Mann-Whitney. Deberían presentarse como mediana (RIQ).
2. "Condición corporal" es ordinal (escala 1–5); reportar 3.1 ± 0.7 sugiere una precisión decimal que la escala no soporta.
3. HPG con distribución sesgada (la DE es casi tan grande como la media) presentado como media ± DE; la mediana sería más representativa.
4. "< 0.05*" no es un valor p exacto. Debe reportarse el valor específico.
5. "NS" no es un valor p. Debe reportarse el valor p exacto (por ejemplo, $p = 0.348$).
6. Falta el estadístico *U* para cada prueba.
7. Falta el tamaño del efecto.
8. No se indica si el valor p es exacto o asintótico, ni si se aplicó corrección de continuidad.

Ejercicio 8.6.

Con los datos del caso integrador de anestesia (sección 8.11), genere en R la tabla descriptiva con *gtsummary* y el boxplot con *ggpubr* para el tiempo de recuperación. Exporte la tabla como archivo Word.

Respuesta: El código completo se encuentra en las secciones 8.10.2 y 8.10.4. El estudiante debe ejecutarlo, verificar que la tabla y la figura se producen correctamente, y comprobar que la tabla exportada se abre en Word con el formato adecuado.

Nivel avanzado

Ejercicio 8.7. Diseñe un estudio que compare la eficacia de tres desparasitantes (ivermectina, albendazol y levamisol) en ovinos Corriedale de la sierra central de Perú. Especifique: variable respuesta, tamaño de muestra, método estadístico no paramétrico apropiado, y redacte una sección de resultados ficticia pero completa siguiendo todas las normas de reporte discutidas en este capítulo. Incluya una tabla, una figura y un párrafo de texto.

Rúbrica orientadora:

- *Diseño adecuado (20 %):* Variable respuesta claramente definida (HPG pre y post tratamiento, o porcentaje de reducción de HPG), tamaño de muestra justificado (mínimo 15 por grupo para Kruskal-Wallis con potencia razonable), diseño experimental claro (aleatorización, cegamiento si aplica).
- *Método estadístico correcto (20 %):* Kruskal-Wallis para la comparación global (o Friedman si se usan datos pareados pre-post), con post hoc de Dunn y ajuste especificado. Tamaño del efecto calculado.
- *Tabla completa (20 %):* Medianas con RIQ, estadístico de prueba, valor p exacto, tamaño del efecto, nota al pie con toda la información metodológica, letras superíndices para post hoc.
- *Figura apropiada (20 %):* Boxplot con puntos individuales, escala logarítmica si HPG, anotaciones de significancia.
- *Redacción (20 %):* Texto coherente con la tabla y la figura, valores exactos, interpretación en términos veterinarios (eficacia antihelmíntica, posible resistencia), mención de limitaciones.

Ejercicio 8.8.

A partir de un conjunto de datos con al menos 100 observaciones y 5 variables predictoras (puede usar los datos simulados de granjas porcinas de la sección 8.10.6 o crear los propios), ajuste un random forest con *tidymodels*. Realice validación cruzada con 5 *folds*, reporte AUC, sensibilidad y especificidad en el conjunto de prueba, genere el gráfico de importancia de variables, y escriba un párrafo interpretando los resultados en contexto veterinario. Luego, seleccione la variable más

importante según el random forest y aplique una prueba de Mann-Whitney para comparar esa variable entre los dos grupos de clasificación. Discuta en un párrafo la diferencia entre la información que proporciona la importancia de variables del random forest y la que proporciona la prueba de Mann-Whitney.

Rúbrica orientadora:

- *Implementación correcta del random forest (25 %):* División entrenamiento/prueba con estratificación, receta de preprocesamiento, especificación del modelo con *rango*, validación cruzada funcional, hiperparámetros optimizados.
- *Métricas de evaluación (20 %):* AUC, sensibilidad y especificidad reportadas con intervalos de confianza. Matriz de confusion presentada.
- *Gráfico de importancia (15 %):* Gráfico legible con las variables ordenadas por importancia, etiquetas claras.
- *Mann-Whitney complementario (15 %):* Prueba correctamente aplicada, tamaño del efecto calculado, reporte según APA.
- *Discusión comparativa (25 %):* El estudiante debe articular que la importancia de variables mide contribución a la predicción (en un modelo multivariado que captura interacciones), mientras que Mann-Whitney mide la diferencia marginal en la distribución de una sola variable entre dos grupos (sin considerar las otras variables). Ambas pueden coincidir o no, y la discordancia es informativa: una variable puede ser importante en el random forest solo en interacción con otra, y no mostrar diferencia significativa en Mann-Whitney por sí sola.

8.14. Conclusiones del capítulo

El análisis estadístico y su comunicación son dos mitades de un mismo proceso. Un análisis no paramétrico bien ejecutado pierde su valor si se reporta de manera incompleta o confusa, y un reporte impecable no salva un análisis mal planteado. Las plantillas de reporte de este capítulo cubren cada prueba tratada en el libro y siguen las normas APA 7.ª edición adaptadas a las convenciones de las revistas veterinarias. El investigador que las use como guía podrá producir

secciones de resultados que sean informativas, verificables y compatibles con meta-análisis futuros.

Las tendencias en análisis de datos veterinarios apuntan en tres direcciones simultáneas. Los métodos bayesianos no paramétricos ofrecen flexibilidad distribucional y cuantificación directa de la incertidumbre, pero requieren formación estadística avanzada y aún encuentran resistencia en la revisión por pares. El machine learning proporciona capacidad predictiva sin precedentes para problemas con muchas variables y relaciones complejas, pero no reemplaza la inferencia estadística: predecir bien no es lo mismo que entender por qué. Los datos masivos de sensores, ómicas y registros clínicos electrónicos generan oportunidades analíticas que ni la estadística clásica ni el machine learning por separado pueden aprovechar plenamente; la combinación de ambos, con cada herramienta usada para lo que fue diseñada, es el camino que la evidencia acumulada y la lógica metodológica sugieren.

La estadística no paramétrica que se estudió a lo largo de este libro no es una fase transitoria que será reemplazada por algoritmos más sofisticados. Es un conjunto de herramientas de inferencia que responden preguntas fundamentales ¿hay diferencia?, ¿cuál es su magnitud?, ¿es generalizable? sin depender de supuestos distribucionales que los datos veterinarios frecuentemente no cumplen. Esas preguntas seguirán siendo las preguntas centrales de la investigación veterinaria, independientemente de cuántos terabytes de datos generen los sensores del futuro.

Referencias Bibliográficas

Agresti, A. (2010). *Analysis of ordinal categorical data* (2.ª ed.). John Wiley & Sons.

Agresti, A. (2013). *Categorical data analysis* (3.ª ed.). Wiley.

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7.ª ed.). https://doi.org/10.1037/0000165-000

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*(1), 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Anderson, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics, 62*(1), 245–253. https://doi.org/10.1111/j.1541-0420.2005.00440.x

Anderson, M. J., & Walsh, D. C. I. (2013). PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions. *Ecological Monographs, 83*(4), 557–574. *https://doi.org/10.1890/12-2010.1*

Aubry, P., & Geale, D. W. (2011). A review of bovine anaplasmosis. *Transboundary and Emerging Diseases, 58*(1), 1–30.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. *The Annals of Mathematical Statistics, 25*(2), 290–302. https://doi.org/10.1214/aoms/1177728786 *(*

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs, 27*(4), 325–349. *https://doi.org/10.2307/1942268*

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens-Fisher problem. *Biometrical Journal, 42*(1), 17–25. https://doi.org/10.1002/(SICI)1521-4036(200001)42:1 <17::AID-BIMJ17>3.0.CO;2-U

Camilli, G., & Hopkins, K. D. (1978). Applicability of chi-square to 2 × 2 contingency tables. *Psychological Bulletin, 85*(1), 163–167. https://doi.org/10.1037/0033-2909.85.1.163

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology, 18*(1), 117–143. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement. *Psychological Bulletin, 70*(4), 213–220.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2.ª ed.). Lawrence Erlbaum Associates.

Conover, W. J. (1999). *Practical nonparametric statistics* (3.ª ed.). John Wiley & Sons.

Conover, W. J., & Iman, R. L. (1979). On multiple-comparisons procedures. *Technical Report LA-7677-MS*. Los Alamos Scientific Laboratory.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B, 34*(2), 187–220.

Diavastis, S. (2024). Adoption of contemporary management accounting practices.

Divine, G. W., Norton, H. J., Barón, A. E., & Juarez-Colunga, E. (2018). The Wilcoxon-Mann-Whitney procedure fails as a test of medians. *The American Statistician, 72*(3), 278–286. https://doi.org/10.1080/00031305.2017.1305291

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics, 6*(3), 241–252. https://doi.org/10.1080/00401706.1964.10490181

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Fagerland, M. W. (2012). t-tests, non-parametric tests, and large studies. *BMC Medical Research Methodology, 12*(1), 78. https://doi.org/10.1186/1471-2288-12-78

Friedman, M. (1937). The use of ranks to avoid the assumption of normality. *Journal of the American Statistical Association, 32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522

Hart, A. (2001). Mann-Whitney test is not just a test of medians. *BMJ*, *323*(7309), 391–393. https://doi.org/10.1136/bmj.323.7309.391

Hollander, M., Wolfe, D. A., & Chicken, E. (2014). *Nonparametric statistical methods* (3.ª ed.). John Wiley & Sons.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1/2), 81–93.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger. *The Annals of Mathematical Statistics, 18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Masud, M. (2022). Decision-making and systematic evaluation.

Pavlatos, O. (2018). Application of management accounting tools in the hotel sector.

Romney, M., et al. (2021). Integration of accounting systems and decision-making.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2.ª ed.). McGraw-Hill.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80–83. *https://doi.org/10.2307/3001968*

Zar, J. H. (2010). *Biostatistical analysis* (5.ª ed.). Prentice Hall.

ACERCA DE LOS AUTORES



Víctor Alejandro Bosquez Barcenes

Víctor Alejandro Bosquez Barcenes es docente investigador en la Universidad Estatal de Bolívar, ubicada en Guaranda, Ecuador. Posee una Licenciatura en Ciencias de la Educación mención Informática Educativa, una Especialidad en Redes de Comunicación de Datos, una Maestría en Informática Empresarial, una Maestría en Estadística Aplicada y un Doctorado en Educación. Cuenta con más de 15 años de experiencia docente en niveles de pregrado y posgrado, habiendo impartido módulos en diversas universidades ecuatorianas como la UPSE, UNEMI, UNACH y la Universidad Metropolitana, en maestrías relacionadas con educación, tecnología, inteligencia artificial, ciencias veterinarias y salud pública. Ha dirigido y participado en proyectos de investigación sobre accesibilidad web, inteligencia artificial aplicada a la educación y preservación digital, entre otros.



Washington Fernando Carrasco Sangache

Washington Fernando Carrasco Sangache es Médico Veterinario y Zootecnista, Magíster en Clínica y Cirugía Canina y Doctor en Ciencias Veterinarias. Se desempeña como docente e investigador en la Universidad Estatal de Bolívar. Sus principales áreas de interés profesional abarcan la docencia universitaria, el desarrollo estudios en salud animal, la bioestadística, así como la anestesia, la patología clínica y el control de la reproducción en pequeños animales. Cuenta con publicaciones en el área veterinaria, además de participación en proyectos de investigación y congresos nacionales e internacionales.



Verónica Lourdes Carrasco Sangache

Verónica Lourdes Carrasco Sangache Médica Veterinaria y Zootecnista por la Universidad Estatal de Bolívar y Magíster en Clínica y Cirugía Canina por la Universidad Agraria del Ecuador. Con experiencia profesional en medicina veterinaria desde 2009 y trayectoria como docente universitaria en la Universidad Estatal de Bolívar y la ESPOCH. Ha impartido asignaturas en las áreas de anatomía, clínica, cirugía, bioseguridad y medicina de monogastricos. Participación en actividades de investigación, vinculación con la sociedad. Es autora y coautora de publicaciones científicas en revistas nacionales e internacionales. Posee formación complementaria en cirugía, anestesiología, emergentología, y medicina interna de pequeños animales. Mantiene capacitación continua en docencia universitaria, innovación educativa e inteligencia artificial aplicada a la educación e investigación.



Jorge Jagger Segura Ochoa

Jorge Jagger Segura Ochoa es Médico Veterinario y Zootecnista, Doctor en Ciencias Veterinarias por la Universidad Nacional de La Plata (Argentina, 2023) y Magíster en Salud Animal Avanzada por la Universidad Central Martha Abreu de las Villas (Cuba, 2017). Se desempeña como docente investigador en la Universidad Estatal de Bolívar. Es investigador registrado en Ecuador y miembro del Grupo de Estudio de la Biomasa de la UEB. Su producción científica incluye publicaciones en revistas indexadas como Theriogenology y Caspian Journal of Environmental Sciences, con líneas de investigación en inmunocontracepción felina, detección molecular de patógenos y reproducción bovina. Ha participado como evaluador en concursos de méritos en varias universidades ecuatorianas y como ponente en congresos científicos internacionales.