Matemáticas para la Estadística Médica con R

Bioestadística — Universidad de Granada

Miguel Ángel Luque Fernández

2026

Contenidos

Prefacio **1**
Objetivos del Curso . . . 1
Estructura del Libro . . . 2
Sobre el Autor . . . 2
Software y Reproducibilidad . . . 3
Nota sobre el Material . . . 4

I Parte I: Análisis Exploratorio de Datos **5**

1. Semana 1 — Análisis Exploratorio de Datos (Parte I) **6**
1.1. Introducción . . . 6
1.2. Conceptos Fundamentales . . . 6
1.2.1. Unidades Estadísticas, Variables y Valores . . . 6
1.2.2. Observaciones Atípicas (Outliers) . . . 7
1.3. Escalas de Medición . . . 7
1.3.1. Variables Categóricas . . . 8
1.3.2. Variables Ordinales . . . 8
1.3.3. Variables Numéricas . . . 8
1.4. Agrupación de Datos en Clases (Binning) . . . 9
1.5. Tablas de Frecuencias y Distribución Empírica . . . 10
1.5.1. Frecuencias Absolutas y Relativas . . . 10
1.5.2. Tabla de Frecuencias . . . 10
1.5.3. Frecuencias Acumuladas . . . 11
1.5.4. Función de Distribución Empírica (ECDF) . . . 13
1.5.5. Interpolación Lineal para Datos Agrupados . . . 16
1.6. Histogramas y Gráficos de Densidad . . . 17
1.6.1. Construcción de Histogramas . . . 17
1.6.2. Histograma vs Gráfico de Densidad . . . 18
1.7. Clasificación de Medidas Estadísticas . . . 20
1.8. Medidas de Posición Central . . . 20
1.8.1. Media Aritmética . . . 20
1.8.2. Media para Datos Agrupados . . . 21
1.8.3. Moda . . . 22
1.8.4. Mediana . . . 23
1.8.5. Relación entre Media, Mediana y Moda . . . 24

1.9. Medidas de Forma: Sesgo y Curtosis . . . 24
1.9.1. Cálculo con R . . . 25
1.9.2. Cuantiles . . . 26
1.10. Comparación de Medidas de Posición . . . 29
1.11. Resumen de Conceptos Clave . . . 29
1.11.1. Tabla Resumen: Escalas de Medición y Estadísticos Apropiados . . . 29
1.11.2. Tabla Resumen: Fórmulas Importantes . . . 29
1.12. Ejercicios . . . 30
1.13. Respuestas a los Ejercicios . . . 32

2. Semana 2 — Análisis Exploratorio de Datos (Parte II) 34
2.1. Medidas de Dispersión . . . 34
2.1.1. ¿Por qué necesitamos medidas de dispersión? . . . 34
2.1.2. Rango . . . 36
2.1.3. Rango Intercuartílico (IQR) . . . 36
2.1.4. Varianza y Desviación Estándar . . . 37
2.1.5. Cálculo de Estadísticos con R Base . . . 38
2.2. Transformaciones Lineales: Propiedades . . . 39
2.2.1. Transformación Lineal Estándar . . . 39
2.2.2. Interpretación de los Parámetros . . . 40
2.3. Estandarización y Normalización . . . 40
2.3.1. Propiedades de los Z-Scores . . . 41
2.3.2. Interpretación . . . 41
2.3.3. Z-scores con Parámetros Poblacionales . . . 42
2.3.4. Regla Empírica (68-95-99.7) . . . 43
2.4. Diagramas de Caja (Boxplots) . . . 44
2.4.1. Construcción de un Boxplot . . . 44
2.4.2. Interpretación Visual . . . 44
2.5. Análisis Bivariante: Distribuciones Conjuntas . . . 46
2.5.1. Distribución Conjunta . . . 46
2.5.2. Variables Discretas: Tablas de Frecuencia . . . 46
2.5.3. Distribución Marginal . . . 46
2.5.4. Distribución Condicional . . . 47
2.5.5. Covarianza y Correlacion: Advertencia Visual . . . 55
2.5.6. Matriz de Correlaciones . . . 57
2.6. Función de Distribución Empírica (ECDF) . . . 59
2.7. Ejemplo 2.11: ECDF de Alturas de Estudiantes . . . 59
2.7.1. Interpretación . . . 60
2.8. Resumen . . . 61
2.8.1. Puntos Clave . . . 61
2.9. Ejercicios . . . 61
2.10. Respuestas a los Ejercicios . . . 63

II Parte II: Probabilidad y Variables Aleatorias 65

3. Semana 3 — Fundamentos de la Teoría de la Probabilidad 66
3.1. Introducción . . . 66

3.2. Experimentos Aleatorios y Eventos . . . 66
3.2.1. Experimento Aleatorio . . . 66
3.2.2. Evento Elemental y Partición . . . 67
3.3. Álgebra de Conjuntos . . . 67
3.3.1. Relaciones y Operaciones de Eventos . . . 68
3.3.2. Diagramas de Venn . . . 68
3.3.3. Leyes de De Morgan . . . 68
3.4. Definiciones de Probabilidad . . . 69
3.4.1. Probabilidad Clásica (Laplace) . . . 69
3.4.2. Probabilidad Frecuentista (von Mises) . . . 70
3.4.3. Probabilidad Axiomática (Kolmogorov) . . . 71
3.5. Propiedades de la Probabilidad . . . 72
3.6. Teorema de la Suma (Inclusión-Exclusión) . . . 73
3.7. Probabilidad Condicional . . . 74
3.7.1. Teorema de la Multiplicación . . . 75
3.8. Independencia de Eventos . . . 75
3.8.1. Independencia de Múltiples Eventos . . . 76
3.9. Ley de la Probabilidad Total . . . 76
3.10. Teorema de Bayes . . . 77
3.10.1. Forma Simple del Teorema de Bayes . . . 77
3.10.2. Forma General con Partición . . . 77
3.10.3. Ejemplo: Prueba Diagnóstica . . . 78
3.10.4. Ejemplo: Clasificación de pacientse . . . 79
3.11. Resumen . . . 80
3.11.1. Conceptos Clave . . . 80
3.11.2. Definiciones de Probabilidad . . . 80
3.11.3. Fórmulas Importantes . . . 80
3.12. Ejercicios . . . 81
3.13. Respuestas a los Ejercicios . . . 82

4. Semana 4 — Variables Aleatorias y Distribuciones de Probabilidad 84
4.1. Introducción . . . 84
4.2. Variables Aleatorias . . . 84
4.2.1. Tipos de Variables Aleatorias . . . 84
4.2.2. Función de Probabilidad y Función de Distribución . . . 85
4.3. Funciones de Probabilidad en R . . . 85
4.4. Fundamentos de Cálculo de Probabilidades: Discreta vs. Continua . . . 86
4.4.1. La Singularidad del Punto en Variables Continuas . . . 86
4.4.2. El Cuidado de las Desigualdades en Variables Discretas . . . 86
4.5. Varianza y Desviación Estándar . . . 87
4.5.1. Desviación Estándar poblacional . . . 88
4.5.2. Propiedades de la Varianza . . . 88
4.6. Estandarización (Transformación Z) . . . 89
4.7. Variables Aleatorias Bivariantes . . . 89
4.7.1. Distribuciones Conjunta, Marginal y Condicional (Caso Discreto) . . . 89
4.7.2. Distribuciones Conjunta, Marginal y Condicional (Caso Continuo) . . . 90
4.7.3. Covarianza y Correlación . . . 90
4.8. Distribuciones Discretas Importantes . . . 91

4.8.1. Distribución Uniforme Discreta $U(n)$ . . . 91
4.8.2. Distribución de Bernoulli $B(p)$ . . . 93
4.8.3. Distribución Binomial $B(n, p)$ . . . 94
4.8.4. Distribución de Poisson $\text{Po}(\lambda)$ . . . 95
4.9. Distribuciones Continuas Importantes . . . 97
4.9.1. Distribución Uniforme Continua $U(a, b)$ . . . 97
4.9.2. Distribución Exponencial $\text{Exp}(\lambda)$ . . . 99
4.9.3. Distribución Normal $N(\mu, \sigma^2)$ . . . 100
4.9.4. Aproximación Normal a la Binomial y Corrección de Continuidad . . . 102
4.10. Distribuciones para Muestreo . . . 107
4.10.1. Distribución Chi-Cuadrado χ^2_k . . . 107
4.10.2. Distribución t de Student t_k . . . 109
4.10.3. Distribución F de Snedecor F_{k_1,k_2} . . . 112
4.11. Teorema Central del Límite . . . 114
4.12. Aproximaciones entre distribuciones vistas como aplicaciones del Teorema Central del Límite . . . 116
4.12.1. Binomial → Normal . . . 116
4.12.2. Poisson → Normal . . . 118
4.12.3. Exponencial Poisson (dualidad tiempo-conteo) . . . 118
4.13. Tabla de Referencia de Distribuciones . . . 120
4.14. Resumen de Conceptos Clave . . . 121
4.15. Ejercicios . . . 121
4.15.1. Ejercicio 1: Distribución Binomial . . . 121
4.15.2. Ejercicio 2: Distribución Normal . . . 121
4.15.3. Ejercicio 3: Distribución de Poisson . . . 122
4.15.4. Ejercicio 4: Teorema Central del Límite . . . 122
4.15.5. Ejercicio 5: Transformación Z . . . 122
4.16. Respuestas a los Ejercicios . . . 122
4.17. Lecturas Complementarias . . . 123

5. Semana 5 — Muestreo y Distribuciones Muestrales **125**
5.1. Conceptos Fundamentales de Muestreo . . . 125
5.1.1. ¿Qué es el Muestreo? . . . 125
5.1.2. Parámetros Poblacionales . . . 126
5.1.3. Parámetros vs. Estadísticos . . . 126
5.1.4. El Problema del Sesgo de Muestreo: El Estudio Nurses' Health (NHS) y la Terapia Hormonal Sustitutiva . . . 126
5.2. Tipos de Muestreo . . . 127
5.2.1. Muestreo Aleatorio Simple . . . 127
5.2.2. Muestreo Estratificado . . . 128
5.2.3. Muestreo por Conglomerados . . . 128
5.2.4. Muestreo por Criterio (Judgment Sampling) . . . 128
5.3. Estadísticos Muestrales como Variables Aleatorias . . . 129
5.3.1. El Concepto Clave . . . 129
5.3.2. Estadísticos Importantes . . . 129
5.4. Distribución Muestral de la Media . . . 129
5.4.1. Caso: Muestreo con Reemplazo . . . 129
5.4.2. Caso: Muestreo sin Reemplazo . . . 130

5.4.3. Distribución Exacta Cuando $X \sim N(\mu, \sigma^2)$ . . . 130
5.5. Distribución Muestral de la Proporción . . . 130
5.5.1. Con Reemplazo . . . 131
5.5.2. Normalidad Asintótica . . . 131
5.6. Distribución Muestral de la Varianza . . . 131
5.6.1. Caso: μ Conocida . . . 131
5.6.2. Caso: μ Desconocida . . . 132
5.7. Teorema Central del Límite (Versión Formal) . . . 132
5.7.1. Interpretación . . . 133
5.8. La Distribución t de Student . . . 133
5.8.1. Propiedades de la Distribución t . . . 133
5.8.2. Cuándo Usar Cada Distribución . . . 134
5.9. Simulaciones del Teorema Central del Límite . . . 134
5.10. Resumen de Distribuciones Muestrales . . . 135
5.11. Corrección por Población Finita . . . 136
5.12. Ejercicios . . . 137
5.13. Respuestas a los Ejercicios . . . 138

III Parte III: Inferencia Estadística 140

6. Semana 6 — Estimación de Parámetros 141
6.1. El Modelo Estadístico Paramétrico . . . 141
6.2. Estimadores Puntuales . . . 142
6.3. Propiedades de los Estimadores . . . 142
6.3.1. Insesgadez . . . 142
6.3.2. Eficiencia . . . 143
6.3.3. Consistencia . . . 143
6.3.4. Error Cuadrático Medio (ECM) . . . 144
6.4. Tabla Comparativa de Estimadores Comunes . . . 144
6.5. Método de los Momentos (MoM) . . . 144
6.5.1. Procedimiento del Método de Momentos . . . 145
6.5.2. Ejemplo 6.5.1: Distribución Normal . . . 145
6.5.3. Ejemplo 6.5.2: Distribución de Poisson . . . 147
6.5.4. Ejemplo 6.5.3: Distribución Gamma . . . 147
6.6. Estimación por Máxima Verosimilitud (MLE) . . . 147
6.6.1. La Función de Verosimilitud . . . 148
6.6.2. El Estimador de Máxima Verosimilitud . . . 148
6.6.3. La Función Log-Verosimilitud . . . 148
6.6.4. Ejemplo 6.6.1: MLE para Distribución Binomial . . . 149
6.6.5. Ejemplo 6.6.2: MLE para Distribución Normal . . . 149
6.6.6. Propiedades del EMV . . . 150
6.7. Optimización Numérica para MLE . . . 151
6.8. Intervalos de Confianza . . . 153
6.8.1. Definición de Intervalo de Confianza . . . 153
6.8.2. Interpretación Frecuentista . . . 153
6.8.3. Intervalo de Confianza para la Media (Varianza Conocida) . . . 153
6.8.4. Intervalo de Confianza para la Media (Varianza Desconocida) . . . 154

6.8.5. Intervalo de Confianza para una Proporción . . . 155
6.9. Interpretación de Intervalos de Confianza . . . 157
6.10. Trabajando con Ejemplos Completos . . . 157
6.11. Resumen . . . 162
6.11.1. Fórmulas Clave para Intervalos de Confianza . . . 162
6.12. Cálculo del Tamaño Muestral: `nm()` y `np()` . . . 164
6.12.1. Fórmulas del Tamaño Muestral . . . 164
6.13. Ejercicios . . . 166
6.14. Respuestas a los Ejercicios . . . 167

7. Semana 7 — Contrastes de Hipótesis (Parte I) 169
7.1. Introducción . . . 169
7.2. Hipótesis Nula y Alternativa . . . 169
7.2.1. Tipos de Hipótesis . . . 170
7.2.2. Ejemplos . . . 170
7.3. Errores Tipo I y Tipo II . . . 170
7.3.1. Tabla de Decisión . . . 171
7.3.2. Analogía: El Pronóstico Médico . . . 171
7.3.3. Relación entre α y β . . . 171
7.4. Estadístico de Prueba y Región Crítica . . . 171
7.4.1. Procedimiento de Decisión . . . 172
7.5. Función de Potencia . . . 172
7.6. Tamaño y Nivel de una Prueba . . . 172
7.6.1. Relación con la Significación . . . 173
7.7. Contraste Z para la Media (Varianza Conocida) . . . 173
7.7.1. Condiciones . . . 173
7.7.2. Estadístico de Prueba . . . 173
7.7.3. Regiones de Rechazo y Valores Críticos . . . 173
7.7.4. Valores Críticos Comunes (Normal Estándar) . . . 174
7.7.5. Ejemplo: Productor de Harina . . . 174
7.8. Contraste t para la Media (Varianza Desconocida) . . . 174
7.8.1. Condiciones . . . 174
7.8.2. Estimador de la Varianza . . . 175
7.8.3. Estadístico de Prueba y Distribución . . . 175
7.8.4. Derivación de la Distribución . . . 175
7.8.5. Regiones de Rechazo y Valores Críticos . . . 175
7.8.6. Tabla de Cuantiles t de Student . . . 176
7.8.7. Ejemplo: Una Muestra t (Productor de Harina Revisado) . . . 176
7.9. Contraste para una Proporción . . . 180
7.9.1. Función Muestral . . . 180
7.9.2. Distribución bajo H_0 . . . 181
7.9.3. Aproximación Normal (Muestra Grande) . . . 181
7.9.4. Ejemplo: Moneda Justa . . . 181
7.10. Valor-p (p-value) . . . 183
7.10.1. Interpretación . . . 183
7.10.2. Regla de Decisión . . . 184
7.10.3. Ejemplo: Cálculo de p-valores . . . 184
7.11. Tabla Resumen: Regiones de Aceptación y Rechazo . . . 184

7.12. Ejemplos Completos . . . 185
7.12.1. Ejemplo 1: Contraste Z Bilateral (Conocida la Varianza) . . . 185
7.12.2. Ejemplo 2: Contraste t Unilateral (Desconocida la Varianza) . . . 185
7.12.3. Ejemplo 3: Contraste Binomial Exacto . . . 186
7.13. Relación entre Hipótesis Unilaterales y Bilaterales . . . 188
7.13.1. Elección de la Dirección . . . 188
7.13.2. Implicaciones para Errores . . . 188
7.14. Resumen . . . 189
7.14.1. Tabla de Decisión Rápida . . . 189
7.15. Ejercicios . . . 189
7.15.1. Ejercicio 1: Conceptos Básicos . . . 189
7.15.2. Ejercicio 2: Contraste Z . . . 190
7.15.3. Ejercicio 3: Contraste t . . . 190
7.15.4. Ejercicio 4: Contraste Binomial . . . 190
7.15.5. Ejercicio 5: Decisión con p-valor . . . 191
7.15.6. Ejercicio 6: Interpretación de Resultados . . . 191
7.16. Respuestas a los Ejercicios . . . 191

8. Semana 8 — Contrastes de Hipótesis (Parte II) 193
8.1. Prueba Z para Diferencia de Medias (Varianzas Conocidas) . . . 193
8.1.1. Preparación: Distribución de la Diferencia de Medias . . . 193
8.1.2. Hipótesis e Estadîstico de Prueba . . . 194
8.2. Prueba t para Dos Muestras Independientes (Varianzas Desconocidas) . . . 194
8.2.1. Caso 1: Varianzas Iguales (Homogéneas) . . . 194
8.2.2. Caso 2: Varianzas Desiguales (Heterogéneas) — Prueba de Welch . . . 195
8.3. Prueba t Pareada . . . 198
8.3.1. Motivación y Formulación . . . 198
8.4. Prueba F para Igualdad de Varianzas . . . 198
8.4.1. Hipótesis y Estadîstico . . . 198
8.5. Potencia de una Prueba . . . 199
8.5.1. Conceptos Fundamentales . . . 199
8.5.2. Relación con Errores de Tipo I y II . . . 199
8.5.3. Derivación: Potencia para Prueba Bilateral (Varianza Conocida) . . . 200
8.5.4. Potencia: Prueba Unilateral Derecha . . . 200
8.5.5. Potencia: Prueba Unilateral Izquierda . . . 201
8.5.6. Factores que Afectan la Potencia . . . 201
8.6. Potencia para Pruebas de Proporción . . . 201
8.6.1. Formulación . . . 201
8.7. Cálculo de Tamaño Muestral . . . 202
8.7.1. Principio General . . . 202
8.7.2. Fórmula para Prueba t de Dos Muestras (Varianzas Iguales) . . . 202
8.7.3. Cálculo con BioEstatR: `testt()` y `testp()` . . . 203
8.8. Comparación Múltiple y Corrección de Bonferroni . . . 206
8.8.1. Problema: Inflación de Tasa de Error Tipo I . . . 206
8.8.2. Corrección de Bonferroni . . . 206
8.9. Ejemplo Completo: Análisis de Datos de Mosquitos . . . 207
8.9.1. Contexto del Problema . . . 207
8.9.2. Datos y Exploración . . . 207

8.9.3. Test t: Varianzas Homogéneas . . . 208
8.9.4. Test t: Varianzas Heterogéneas (Welch) . . . 209
8.9.5. Test F: Igualdad de Varianzas . . . 210
8.9.6. Cálculo de Potencia . . . 210
8.10. Tabla Resumen: Contrastes de Dos Muestras . . . 210
8.11. Resumen . . . 211
8.11.1. Conceptos Clave . . . 211
8.11.2. Potencia Estadística . . . 211
8.11.3. Tamaño Muestral . . . 211
8.11.4. Testing Múltiple . . . 212
8.12. Verificación de Supuestos: Normalidad y Homogeneidad . . . 212
8.12.1. Test de Shapiro-Wilk: Normalidad . . . 213
8.12.2. Test de Bartlett: Homogeneidad de Varianzas . . . 214
8.13. Bootstrap y Métodos Robustos como Alternativas . . . 216
8.14. Pruebas No Paramétricas Basadas en Rangos . . . 221
8.14.1. Prueba de Wilcoxon-Mann-Whitney (Muestras Independientes) . . . 221
8.14.2. Prueba de Wilcoxon de los Rangos con Signo (Muestras Pareadas) . . . 223
8.14.3. Prueba de Kruskal-Wallis (K Muestras Independientes) . . . 224
8.15. Ejercicios . . . 225
8.16. Respuestas a los Ejercicios . . . 226
8.17. Recursos Adicionales . . . 226

IV Parte IV: Regresión Lineal 227

9. Semana 9 — Regresión Lineal Simple 228
9.1. El Problema de Regresión . . . 228
9.1.1. Variables en un modelo de regresión . . . 228
9.1.2. Observaciones en una muestra . . . 228
9.2. Supuestos del Modelo Lineal . . . 229
9.2.1. Implicaciones de los supuestos . . . 229
9.3. El Modelo de Regresión Lineal Simple . . . 229
9.3.1. Modelo poblacional . . . 229
9.3.2. Interpretación de los parámetros . . . 230
9.3.3. Ejemplo: Presión Arterial Sistólica vs. Edad . . . 230
9.4. Estimación por Mínimos Cuadrados Ordinarios . . . 231
9.4.1. El criterio de MCO . . . 231
9.4.2. Derivación: Las ecuaciones normales . . . 231
9.4.3. Fórmulas cerradas para los estimadores MCO . . . 232
9.4.4. Propiedad importante . . . 232
9.5. Relación entre Regresión y Correlación . . . 232
9.6. Descomposición de Varianza . . . 233
9.6.1. Suma de cuadrados totales, explicada y residual . . . 233
9.7. Coeficiente de Determinación (R^2) . . . 233
9.7.1. Ejemplo: Presión Arterial Sistólica . . . 234
9.8. Distribución de los Estimadores MCO . . . 234
9.8.1. Estimación de la varianza del error . . . 234
9.9. Test t para $\beta_1 = 0$ . . . 235

9.9.1. Hipótesis . . . 235
9.9.2. Estadístico de prueba . . . 235
9.9.3. Regla de decisión . . . 235
9.9.4. Ejemplo ilustrativo: PAS vs. Edad . . . 235
9.10. Intervalos de Confianza . . . 236
9.10.1. Para β_1 . . . 236
9.10.2. Para β_0 . . . 236
9.10.3. Para el valor predicho $E(Y|X = x_0)$ . . . 236
9.10.4. Ejemplo ilustrativo: PAS vs. Edad . . . 236
9.11. Diagnóstico de Residuos . . . 237
9.11.1. 1. Gráfico de Residuos vs. Valores Predichos . . . 237
9.11.2. 2. Gráfico Q-Q (Quantile-Quantile) . . . 237
9.11.3. 3. Gráfico de Escala-Ubicación . . . 238
9.11.4. 4. Residuos vs. Orden de Observación . . . 238
9.12. Ejemplo Completo: Colesterol Total e IMC . . . 238
9.12.1. Contexto clínico . . . 238
9.12.2. Estadísticas descriptivas . . . 238
9.12.3. Cálculos . . . 239
9.12.4. Ecuación estimada . . . 239
9.12.5. Bondad del ajuste . . . 239
9.12.6. Inferencia estadística . . . 239
9.13. Código R . . . 240
9.13.1. Datos clínicos simulados y ajuste del modelo . . . 240
9.13.2. Gráfico de dispersión con la recta de regresión . . . 241
9.13.3. Gráficos de diagnóstico de residuos . . . 242
9.13.4. Intervalos de confianza y predicción clínica . . . 242
9.13.5. Interpretación de la salida de `summary(lm())` . . . 243
9.13.6. Ejemplo completo: Colesterol Total ~ IMC . . . 244
9.14. Resumen de Fórmulas Clave . . . 246
9.15. Ejercicios . . . 247
9.16. Respuestas a los Ejercicios . . . 249

10. Semana 10 — Regresión Lineal Múltiple **251**
10.1. El Modelo de Regresión Lineal Múltiple . . . 251
10.2. Formulación Matricial del Modelo . . . 252
10.3. Estimación de Parámetros: Mínimos Cuadrados Ordinarios . . . 253
10.3.1. Derivación del Estimador OLS . . . 253
10.3.2. Condiciones para Invertibilidad . . . 254
10.4. Matriz de Varianza-Covarianza de los Coeficientes . . . 254
10.4.1. Estimador Insesgado de σ_U^2 . . . 254
10.5. Variables Indicadoras (Dummy) . . . 255
10.6. Bondad de Ajuste: R^2 y R^2 Ajustado . . . 255
10.6.1. Coeficiente de Determinación Múltiple . . . 255
10.6.2. R^2 Ajustado . . . 256
10.7. Criterios de Información para Selección de Modelos . . . 256
10.8. Pruebas de Hipótesis para Coeficientes Individuales . . . 257
10.8.1. Test t para β_j . . . 257
10.9. Prueba F para Comparación de Modelos Anidados . . . 258

10.10 ANOVA como Modelo Lineal . . . 258
10.11 Gráficos de Diagnóstico . . . 259
10.12 Transformaciones para Violaciones de Supuestos . . . 260
10.13 Intervalos de Confianza para Coeficientes . . . 260
10.14 Intervalos de Confianza y Predicción . . . 261
10.14.1 Intervalo de Confianza para $E(Y|\mathbf{x}_0)$ . . . 261
10.14.2 Intervalo de Predicción para una Nueva Observación . . . 261
10.15 Multicolinealidad . . . 261
10.16 Selección de Variables . . . 262
10.17 Ejemplo Completo en R . . . 263
10.17.1 Modelo de regresión múltiple: PAS ~ Edad + IMC . . . 263
10.17.2 Comparación de modelos anidados con prueba F . . . 265
10.17.3 Diagnóstico de residuos del modelo múltiple . . . 266
10.17.4 Selección de variables por AIC (stepwise) . . . 266
10.17.5 Predicción para perfiles clínicos específicos . . . 267
10.17.6 Ejemplo con variables indicadoras: Tratamientos antihipertensivos . . . 268
10.17.7 Ejemplo con `BioEstatR`: regresión lineal múltiple . . . 270
10.18 Resumen . . . 275
10.19 Ejercicios . . . 276
10.20 Respuestas a los Ejercicios . . . 277

V Parte V: Análisis de Datos Categóricos 279

11. Semana 11 — Análisis de Datos Categóricos 280
11.1. Variables Categóricas y Tablas de Contingencia . . . 280
11.1.1. Tabla 2×2: Estructura y Notación Epidemiológica . . . 280
11.1.2. Tablas r×c Generalizadas . . . 281
11.2. Prueba ^{2} de Independencia . . . 281
11.2.1. Hipótesis . . . 281
11.2.2. Frecuencias Esperadas Bajo H_0 . . . 282
11.2.3. Cálculo Manual para Tabla 2×2 . . . 282
11.2.4. Ejemplo Médico: Tabaquismo y Osteoporosis . . . 282
11.3. Prueba ^{2} de Homogeneidad . . . 284
11.3.1. Independencia vs. Homogeneidad . . . 284
11.3.2. Procedimiento . . . 284
11.4. Condiciones de Aplicación y Alternativas . . . 285
11.4.1. Condición de Frecuencias Esperadas . . . 285
11.4.2. Corrección de Continuidad de Yates (Tabla 2×2) . . . 286
11.4.3. Prueba Exacta de Fisher . . . 288
11.5. Prueba de McNemar para Datos Apareados . . . 290
11.5.1. ¿Cuándo usar McNemar? . . . 290
11.5.2. Estructura de la Tabla Apareada . . . 290
11.5.3. Estadístico de McNemar . . . 291
11.5.4. Ejemplo: Intervención para el Control de la Hipertensión . . . 292
11.5.5. Implementación: BioEstatR vs. Base R . . . 293
11.6. Diseños de Estudio y Medidas de Asociación . . . 295
11.6.1. Estructura Epidemiológica de la Tabla 2×2 . . . 295

11.6.2. Medidas de Asociación . . . 296
11.6.3. Diferencia de Riesgos (RD) y NNT . . . 302
11.6.4. Estudio Transversal: Razón de Prevalencias (RP) . . . 303
11.6.5. Estudio Caso-Control: Odds Ratio (OR) . . . 303
11.6.6. Resumen de Intervalos de Confianza . . . 304
11.7. Cálculo en R base de las medidas de asociación y sus intervalos de confianza . . . 305
11.8. Ajuste por Confusión: El Estimador de Mantel-Haenszel . . . 306
11.8.1. El Estimador de Mantel-Haenszel (MH) . . . 306
11.8.2. Ejemplo Numérico en R: Evaluación de Confounding . . . 306
11.8.3. Control multivariable de la confusión: regresión logística múltiple con `rlogitm()` 308
11.9. Ejercicios del tema . . . 312
11.10Respuestas a los Ejercicios . . . 314
11.11Recomendaciones de Métodos Avanzados . . . 315
11.12Lecturas Recomendadas Adicionales . . . 315

Referencias **317**
Sobre esta sección . . . 317
Libros de Referencia General . . . 317
Análisis Exploratorio de Datos . . . 317
Probabilidad y Variables Aleatorias . . . 317
Teoría del Muestreo e Inferencia Estadística . . . 318
Regresión Lineal . . . 318
Análisis de Datos Categóricos y Epidemiología . . . 318
Recursos en R . . . 318
Perspectivas Bayesianas . . . 319
Tópicos Especializados . . . 319
Diccionarios y Referencias Rápidas . . . 319
Cómo citar este curso . . . 319
Acceso Abierto . . . 319

Appendices **323**

A. Apéndice A: Repaso Matemático Riguroso **323**
A.1. A.1 Funciones y Gráficas . . . 323
A.2. A.2 Exponenciales y Logaritmos . . . 324
A.3. A.3 Sumatorias y Productos . . . 325
A.4. A.4 Teoría de Conjuntos . . . 326
A.5. A.5 Cálculo: Derivadas e Integrales . . . 327
A.5.1. A.5.1 Derivadas . . . 327
A.5.2. A.5.2 Integrales . . . 328
A.6. A.6 Álgebra Matricial . . . 328
A.7. A.7 Referencias . . . 330

B. Apéndice B: El Paquete `BioEstatR` — Guía de Referencia **334**
B.1. B.1 Instalación y el conjunto de datos `osteo` . . . 334
B.2. B.2 Análisis descriptivo: `freq()` y `grps()` . . . 336
B.2.1. B.2.1 Tablas de frecuencias: `freq()` . . . 336

B.2.2. B.2.2 Estadísticas por grupos: grps() . . . 337
B.3. B.3 Intervalos de confianza: icm(), icp(), icl() . . . 338
B.3.1. B.3.1 IC para la media: icm() . . . 338
B.3.2. B.3.2 IC para proporción: icp() . . . 339
B.3.3. B.3.3 IC para tasa Poisson: icl() . . . 340
B.4. B.4 Tamaño muestral: nm(), np() y nl() . . . 341
B.4.1. B.4.1 Para la media: nm() . . . 341
B.4.2. B.4.2 Para proporciones: np() . . . 342
B.4.3. B.4.3 Para tasa Poisson: nl() . . . 342
B.5. B.5 Contrastes de hipótesis . . . 343
B.5.1. B.5.1 Test de normalidad: testnormal() . . . 343
B.5.2. B.5.2 Test t: testt() . . . 344
B.5.3. B.5.3 Test F de varianzas: testf() . . . 346
B.5.4. B.5.4 Test de proporciones: testp() . . . 347
B.5.5. B.5.5 Test de Wilcoxon: testwx() . . . 348
B.5.6. B.5.6 Test de McNemar: testmcnemar() . . . 349
B.6. B.6 Regresión lineal: rls() y rlm() . . . 350
B.6.1. B.6.1 Regresión lineal simple: rls() . . . 350
B.6.2. B.6.2 Regresión lineal múltiple: rlm() . . . 352
B.7. B.7 Regresión logística: rlogits() y rlogitm() . . . 354
B.7.1. B.7.1 Regresión logística simple: rlogits() . . . 354
B.7.2. B.7.2 Regresión logística múltiple: rlogitm() . . . 356
B.8. B.8 Tablas de contingencia: tabla2x2() y tablarxc() . . . 358
B.8.1. B.8.1 Tabla 2×2: tabla2x2() . . . 358
B.8.2. B.8.2 Tabla r×c: tablarxc() . . . 359
B.9. B.9 Tabla resumen de funciones . . . 360

C. Apéndice C: Guía de R para Estadística Médica 363
C.1. C.1 Conceptos Fundamentales . . . 363
C.1.1. Asignación, Tipos y Vectores . . . 363
C.1.2. Indexación: posicional y lógica . . . 363
C.1.3. Valores Perdidos: NA . . . 364
C.2. C.2 Estructuras de Datos . . . 365
C.2.1. Data Frames y Tibbles . . . 365
C.2.2. Listas . . . 365
C.3. C.3 Factores y Variables Categóricas . . . 365
C.3.1. Recodificación con cut() y dplyr::case_when() . . . 366
C.4. C.4 Importación de Datos Clínicos . . . 366
C.5. C.5 Manipulación de Datos con dplyr . . . 366
C.5.1. Uniones (*_join) y reestructuración (pivot_*) . . . 367
C.6. C.6 Estadísticos Descriptivos . . . 367
C.6.1. Variables Continuas . . . 367
C.6.2. Variables Categóricas . . . 367
C.6.3. Resumen Estratificado . . . 368
C.7. C.7 Funciones Propias y Control de Flujo . . . 368
C.7.1. Definir funciones . . . 368
C.7.2. Control de flujo: if, for, sapply . . . 368
C.8. C.8 Visualización con ggplot2 . . . 368

C.9. C.9 Tests Estadísticos en R Base . . . 371
C.10. C.10 Paquetes Esenciales en Bioestadística . . . 373
C.11. C.11 Reproducibilidad en Investigación Médica . . . 374
C.12. C.12 Resolución de Problemas Comunes . . . 375
C.12.1. Comandos de ayuda . . . 375
C.13. Resumen . . . 375

D. Apéndice D: Investigación Reproducible con Git, GitHub y RStudio 382
D.1. D.1 ¿Por qué investigación reproducible? . . . 382
D.1.1. D.1.1 Niveles de reproducibilidad . . . 383
D.1.2. D.1.2 Principios FAIR . . . 383
D.1.3. D.1.3 ¿Qué herramientas mínimas necesitamos? . . . 383
D.2. D.2 Control de versiones con Git . . . 383
D.2.1. D.2.1 Conceptos básicos . . . 383
D.2.2. D.2.2 Instalación . . . 384
D.2.3. D.2.3 Configuración inicial (una sola vez por máquina) . . . 384
D.2.4. D.2.4 Flujo de trabajo básico (comandos esenciales) . . . 385
D.2.5. D.2.5 Buenas prácticas para mensajes de commit . . . 385
D.2.6. D.2.6 Ramas y fusiones (introducción) . . . 385
D.3. D.3 GitHub como plataforma colaborativa . . . 386
D.3.1. D.3.1 Crear una cuenta y un repositorio remoto . . . 386
D.3.2. D.3.2 Autenticación: HTTPS + PAT o SSH . . . 386
D.3.3. D.3.3 Conectar un repositorio local con un remoto . . . 386
D.3.4. D.3.4 Pull requests: revisión por pares del código . . . 387
D.4. D.4 Integración con RStudio . . . 387
D.4.1. D.4.1 Crear un proyecto RStudio con Git desde cero . . . 387
D.4.2. D.4.2 Conectar un proyecto existente a Git . . . 387
D.4.3. D.4.3 El panel Git de RStudio . . . 387
D.4.4. D.4.4 Visualización de cambios y resolución de conflictos . . . 388
D.5. D.5 Flujos de trabajo reproducibles para Bioestadística . . . 388
D.5.1. D.5.1 Estructura recomendada de un proyecto bioestadístico . . . 388
D.5.2. D.5.2 `.gitignore` típico para un proyecto en R . . . 389
D.5.3. D.5.3 Entornos reproducibles con `renv` . . . 389
D.5.4. D.5.4 Publicar un libro o análisis con GitHub Pages . . . 390
D.6. D.6 Ejercicio guiado: tu primer repositorio reproducible . . . 390
D.7. D.7 Recursos adicionales . . . 391

E. Examen Final 393
E.1. Instrucciones . . . 393
E.2. Parte I: Análisis Exploratorio de Datos (25 puntos) . . . 393
E.2.1. Pregunta 1.1 (12 puntos) . . . 393
E.2.2. Pregunta 1.2 (13 puntos) . . . 394
E.3. Parte II: Probabilidad y Variables Aleatorias (25 puntos) . . . 395
E.3.1. Pregunta 2.1 (12 puntos) . . . 395
E.3.2. Pregunta 2.2 (13 puntos) . . . 396
E.4. Parte III: Inferencia Estadística (25 puntos) . . . 397
E.4.1. Pregunta 3.1 (12 puntos) . . . 397
E.4.2. Pregunta 3.2 (13 puntos) . . . 397

E.5. Parte IV: Regresión Lineal (25 puntos) 398
E.5.1. Pregunta 4.1 (25 puntos) 398
E.6. Parte V: Análisis de Datos Categóricos (25 puntos) 399
E.6.1. Pregunta 5.1 (12 puntos) 399
E.6.2. Pregunta 5.2 (13 puntos) 400
E.7. Instrucciones Finales 401

Prefacio

Este libro contiene el material docente del curso **Matemáticas para la Estadística Médica con R** desarrollado por **Miguel Ángel Luque Fernández** en el Departamento de Estadística e Investigación Operativa de la Universidad de Granada.

Objetivos del Curso

El objetivo principal de este curso y material docente es proporcionar al estudiante de ciencias de la salud los fundamentos matemáticos y estadísticos necesarios para el análisis cuantitativo de datos médicos con un alto grado de rigor académico y con ejemplos contextualizados en la práctica clínica. El libro se sustenta de la filosofía de ciencia reproducible, abierta y computacionalmente aplicada usando el software libre R. Al final del libro se presentan cuatro apéndices reforzando a los alumnos que quieran repasar sobre: i) los contenidos matemáticos básicos necesarios para el buen aprovechamiento del curso, ii) una introducción básica a la programación con R, iii) una introducción al paquete **BioEstatR** producido para la docencia específica de este curso cuyo interés radica en la rica interpretación de los resultados de las funciones del paquete y iv) una introducción a la **investigación reproducible** mediante el control de versiones con Git, GitHub y RStudio. El curso no se limita al uso del paquete `BioEstatR`, ya que se complementa con el uso de las funciones clásicas de R base para todos los ejemplos presentados. Además, se invita al estudiante a explorar y utilizar JAMOVI, una plataforma de estadística gratuita y de código abierto diseñada para ser fácil de usar. Al finalizar el curso, el estudiante será capaz de:

1. **Describir y visualizar** conjuntos de datos mediante técnicas de Análisis Exploratorio de Datos.
2. **Comprender los fundamentos** de la teoría de la probabilidad como base del razonamiento estadístico.
3. **Manejar las distribuciones** de probabilidad más importantes, tanto discretas como continuas.
4. **Aplicar métodos de inferencia estadística**: estimación puntual, estimación por intervalos y contraste de hipótesis.
5. **Construir e interpretar modelos de regresión lineal** simple y múltiple.

Estructura del Libro

El material está organizado en cuatro partes, correspondientes a las diez semanas lectivas del curso:

Parte I — Análisis Exploratorio de Datos (Semanas 1–2) Introducción a los tipos de variables, estadísticos descriptivos univariantes y bivariantes, medidas de dispersión, correlación y representaciones gráficas básicas.

Parte II — Probabilidad y Variables Aleatorias (Semanas 3–5) Fundamentos de la teoría de la probabilidad: espacio muestral, axiomas de Kolmogórov, probabilidad condicional y teorema de Bayes. Variables aleatorias discretas y continuas, distribuciones de probabilidad más importantes y el teorema del límite central.

Parte III — Inferencia Estadística (Semanas 7–9) Conceptos de muestreo y estadísticos muestrales. Estimación puntual: propiedades de los estimadores, método de momentos y máxima verosimilitud. Estimación por intervalos de confianza. Contrastes de hipótesis: errores de tipo I y II, potencia del contraste, pruebas t, z y F.

Parte IV — Regresión Lineal (Semanas 10–11) Regresión lineal simple: mínimos cuadrados ordinarios, interpretación de coeficientes, R^2 y análisis de residuos. Regresión lineal múltiple: formulación matricial, selección de variables y ANOVA de regresión.

Parte V — Métodos Avanzados en Bioestadística Para profundizar en técnicas estadísticas avanzadas aplicadas a la medicina, se recomienda consultar el recurso complementario: **Bioestadística Avanzada**.

Este material avanzado cubre temas críticos para la investigación médica moderna, incluyendo: - **Pruebas Diagnósticas:** Evaluación de precisión, sensibilidad, especificidad y curvas ROC. - **Regresión Logística:** Modelado de variables dependientes binarias y cálculo de Odds Ratios. - **Análisis de Supervivencia:** Estimadores de Kaplan-Meier y modelos de riesgos proporcionales de Cox para datos de tiempo a evento. - **Inferencia Causal y Métodos Robustos:** Técnicas avanzadas para el control de la confusión y el análisis de datos longitudinale (en preparación).

Sobre el Autor

Miguel Ángel Luque Fernández es Catedrático del Departamento de Estadística e Investigación Operativa de la Universidad de Granada. Sus líneas de investigación se centran en la epidemiología cuantitativa, la bioestadística y la inferencia causal.

- Correo: mluquefe@ugr.es
- Web: migariane.github.io
- GitHub: github.com/migariane

Software y Reproducibilidad

Todos los ejemplos de código en este libro están escritos en R. Para reproducir los análisis se recomienda instalar R (versión 4.3) y los siguientes paquetes:

```
install.packages(c("ggplot2", "dplyr", "tidyr", "knitr", "kableExtra", "MASS"))
```

Además, para facilitar el aprendizaje y la aplicación práctica, contamos con:

- **BioEstatR:** Paquete de R desarrollado en la Unidad de Bioestadística de la Facultad de Medicina de la Universidad de Granada por **Pedro Femia Marzo** y **Miguel Ángel Luque Fernández**. Proporciona funciones de alto nivel que integran descripción, inferencia, modelización (lineal y logística) y visualización en una sola llamada, diseñadas específicamente para la docencia de bioestadística en ciencias de la salud. El paquete incluye también el conjunto de datos `osteo` (94 pacientes diabéticos, 27 variables), utilizado como hilo conductor de los ejemplos del libro.
  - **Sitio web del paquete:** https://migariane.github.io/BioEstatR/
  - **Repositorio GitHub:** https://github.com/migariane/BioEstatR
  - **Documentación del Apéndice B** del libro: guía de referencia completa de las 22 funciones del paquete.

Para instalarlo basta con ejecutar (requiere el paquete `remotes`):

```
# Instala el paquete remotes si no lo tienes
install.packages("remotes")

# Instala BioEstatR directamente desde GitHub (compila para tu sistema)
remotes::install_github("migariane/BioEstatR")

# Cargar el paquete y el conjunto de datos osteo
library(BioEstatR)
data(osteo)
```

Para obtener la referencia bibliográfica del paquete:

```
citation("BioEstatR")
```

- **jamovi:** Una plataforma de estadística gratuita y de código abierto diseñada para ser fácil de usar. jamovi proporciona una interfaz intuitiva basada en el motor estadístico de R, ideal para quienes desean realizar análisis complejos sin necesidad de escribir código directamente, facilitando la reproducibilidad de los resultados.

El código fuente completo de este libro está disponible en: https://github.com/migariane/CursoMatematicaEstadistica

Nota sobre el Material

Este material docente ha sido elaborado por el autor para uso docente en la Universidad de Granada. El material tiene carácter introductorio y no pretende ser un tratado exhaustivo de ninguno de los temas cubiertos. Para cada tema se proporcionan referencias para una lectura más profunda.

> **Aviso:** Estos apuntes se encuentran en continua revisión. Si detecta algún error o imprecisión, por favor comuníquelo al autor a través del repositorio de GitHub o por correo electrónico.

Versión correspondiente al curso académico 2025–2026.

Parte I

Parte I: Análisis Exploratorio de Datos

Capítulo 1

Semana 1 — Análisis Exploratorio de Datos (Parte I)

1.1. Introducción

El análisis exploratorio de datos (AED) es el primer paso en cualquier análisis estadístico. A través de visualizaciones, tablas y estadísticos descriptivos, aprendemos a entender las características principales de un conjunto de datos antes de realizar inferencia estadística más compleja.

1.2. Conceptos Fundamentales

1.2.1. Unidades Estadísticas, Variables y Valores

En estadística trabajamos con conceptos fundamentales que debemos dominar desde el principio.

Definición: Conceptos Fundamentales

- **Unidad estadística** (o individuo): cada elemento sobre el que se realiza la observación. Puede ser una persona, animal, planta, empresa, etc.
- **Variable**: una característica de una unidad estadística que puede variar de una unidad a otra. Se denota con letras mayúsculas (X, Y, etc.)
- **Valor**: el resultado específico de observar una variable en una unidad estadística. Se denota con letras minúsculas (x_i, y_i, etc.)
- **Población**: el conjunto completo de todas las unidades estadísticas de interés en un estudio.
- **Muestra**: un subconjunto de la población observado en la práctica. La muestra se utiliza para extraer conclusiones sobre la población.

Ejemplo 1.1: Presión Arterial en Pacientes Diabéticos

Supongamos que deseamos estudiar la presión arterial sistólica (PAS) en pacientes diabéticos de Granada.

- **Unidad estadística**: cada paciente diabético individual
- **Variable**: presión arterial sistólica (X)
- **Valores**: mediciones específicas en mmHg (ej. $x_1 = 125$, $x_2 = 130$, $x_3 = 145$, …)
- **Población**: todos los pacientes diabéticos de Granada
- **Muestra**: los 200 pacientes diabéticos seleccionados para el estudio

1.2.2. Observaciones Atípicas (Outliers)

Definición: Outlier

Un **outlier** (observación atípica) es un valor que:

- Parece desviarse marcadamente de los otros miembros de la muestra
- Aparece o está ausente de forma inesperada
- Puede reflejar una anomalía en la característica medida o un error en la medición/registro

Ejemplo: En el conjunto de presiones arteriales sistólicas $\{125, 128, 130, 131, 198\}$, el valor 198 podría considerarse un outlier, ya que es sustancialmente mayor que los otros valores.

1.3. Escalas de Medición

El tipo de escala utilizada para medir una variable determina qué técnicas analíticas son apropiadas y qué conclusiones se pueden extraer.

Definición: Escalas de Medición

Las variables se clasifican en cuatro tipos según la escala en que se miden:

1. **Variables nominales (categóricas sin orden)**
2. **Variables ordinales (categóricas con orden)**
3. **Variables discretas (numéricas, valores aislados)**
4. **Variables continuas (numéricas, infinitos valores posibles)**

1.3.1. Variables Categóricas

Definición: Variables Categóricas (Nominales)

Una **variable categórica** es aquella que puede tomar uno de un conjunto limitado de valores posibles, asignando cada unidad a una categoría particular basada en una propiedad cualitativa.
Ejemplos:

- Grupo sanguíneo: A, B, AB, O
- Estado civil: soltero, casado, divorciado, viudo
- Género: masculino, femenino
- Región: Andalucía, Cataluña, Madrid, etc.

Los valores de una variable categórica se llaman **niveles**.
Las variables categóricas pueden ser:

- **Binarias (dicotómicas)**: solo dos niveles (ej. éxito/fracaso, enfermo/sano)
- **Politómicas**: más de dos niveles (ej. grupo sanguíneo)

1.3.2. Variables Ordinales

Definición: Variables Ordinales

Una **variable ordinal** es una variable categórica en la que los niveles tienen un orden natural, pero las diferencias entre ellos no son cuantificables.
Relaciones válidas: “mayor que” y “menor que” (o “mejor que” y “peor que”).
Ejemplos:

- Estadio de un tumor: I, II, III, IV
- Grado de dependencia: leve, moderada, severa, total
- Escala de dolor: nada, poco, moderado, intenso, insoportable

1.3.3. Variables Numéricas

Definición: Variables Discretas

Una **variable discreta** es una variable numérica que solo puede tomar un número finito o contablemente infinito de valores. Cada valor tiene vecinos claramente identificables.
Ejemplos:

- Número de nacimientos en un mes
- Número de episodios de hipoglucemia
- Número de dientes extraídos
- Cantidad de admisiones previas en el hospital

Definición: Variables Continuas

Una **variable continua** es una variable numérica que puede tomar infinitos valores no contables. Entre dos valores cualesquiera siempre existen otros valores posibles.

Ejemplos:

- Presión arterial sistólica (en mmHg)
- Peso corporal (en kg)
- Altura (en cm)
- Temperatura (en °C)

En la práctica: variables con muchas unidades contables (como el recuento de glóbulos rojos o blancos) se tratan como continuas y a veces se llaman "cuasi-continuas".

1.4. Agrupación de Datos en Clases (Binning)

Cuando disponemos de muchas observaciones de una variable continua, frecuentemente es útil agruparlas en intervalos de clase para mejorar la claridad.

Definición: Binning (Agrupación en Clases)

El **binning** es la partición de los valores de una variable continua en varias clases (típicamente intervalos). Esto mejora la claridad cuando se tiene gran cantidad de datos.

Para cada intervalo o clase j definimos:

- **Límite inferior de la clase**: x_j^l
- **Límite superior de la clase**: x_j^u
- **Amplitud de la clase**: $\Delta x_j = x_j^u - x_j^l$

Propiedades importantes:

- Los intervalos son no superpuestos (disjuntos) y adyacentes
- $x_j^u = x_{j+1}^l$ (el límite superior de una clase es el límite inferior de la siguiente)
- Convencionalmente: $x_j^l < X \leq x_j^u$ (incluimos el límite superior, excluimos el inferior)

Ejemplo 1.2: Niveles de Colesterol Total

En un estudio sobre salud cardiovascular tenemos:

- **Unidad estadística**: paciente
- **Variable**: nivel de colesterol total (en mg/dL)

Intervalo de Colesterol (mg/dL)	Límites	Amplitud
150 – 200 (Deseable)	$x_1^l = 150$, $x_1^u = 200$	$\Delta x_1 = 50$
200 – 240 (Límite alto)	$x_2^l = 200$, $x_2^u = 240$	$\Delta x_2 = 40$

240 – 300 (Alto)	$x_3^l = 240,\ x_3^u = 300$	$\Delta x_3 = 60$

1.5. Tablas de Frecuencias y Distribución Empírica

Una forma eficaz de resumir un conjunto de datos es construir una tabla de frecuencias.

1.5.1. Frecuencias Absolutas y Relativas

i Definición: Frecuencia Absoluta

La **frecuencia absoluta** de un valor x_j es el número de unidades estadísticas con ese valor característico:

$$h(x_j) = h_j$$

Propiedades:

$$0 \leq h(x_j) \leq n, \quad \sum_{j=1}^{k} h(x_j) = n$$

donde n es el número total de observaciones y k es el número de valores distintos.

i Definición: Frecuencia Relativa

La **frecuencia relativa** de un valor x_j es la proporción de unidades estadísticas con ese valor:

$$f(x_j) = \frac{h(x_j)}{n}$$

Propiedades:

$$0 \leq f(x_j) \leq 1, \quad \sum_{j=1}^{k} f(x_j) = 1$$

La frecuencia relativa es útil para comparar distribuciones de diferentes tamaños.

1.5.2. Tabla de Frecuencias

i Definición: Distribución de Frecuencias Empírica

Una **distribución de frecuencias empírica** es una tabla que lista los valores (o intervalos de valores) de una variable junto con sus frecuencias (absolutas o relativas).
Estructura de una tabla de frecuencias:

Valor (x_j)	Frecuencia Absoluta (h_j)	Frecuencia Relativa (f_j)
x_1	h_1	f_1
x_2	h_2	f_2
...	...	...
x_k	h_k	f_k
Suma	**n**	**1**

1.5.3. Frecuencias Acumuladas

Definición: Frecuencia Acumulada

La **frecuencia acumulada** es la suma de las frecuencias de todos los valores hasta un punto determinado.

Frecuencia acumulada absoluta:

$$H(x_j) = \sum_{i=1}^{j} h(x_i), \quad j = 1, \dots, k$$

Frecuencia acumulada relativa:

$$F(x_j) = \frac{H(x_j)}{n} = \sum_{i=1}^{j} f(x_i), \quad j = 1, \dots, k$$

Propiedades:

- $H(x_1) = h(x_1)$ y $F(x_1) = f(x_1)$
- $H(x_k) = n$ y $F(x_k) = 1$
- Las frecuencias acumuladas son secuencias no decrecientes

Ejemplo 1.3: Tabla de Frecuencias

Consideremos los datos de 10 pacientes diabéticos. Sus presiones arteriales sistólicas (en mmHg) son: 120, 125, 118, 130, 122, 128, 115, 135, 124, 129

PAS (mmHg)	h_j	f_j	H_j	F_j
115	1	0.10	1	0.10
118	1	0.10	2	0.20
120	1	0.10	3	0.30
122	1	0.10	4	0.40
124	1	0.10	5	0.50
125	1	0.10	6	0.60
128	1	0.10	7	0.70
129	1	0.10	8	0.80

130	1	0.10	9	0.90
135	1	0.10	10	1.00

Podemos observar que todas las presiones son diferentes en este ejemplo, por lo que cada una aparece una sola vez.

Cuadro 1.1 Código R para tabla de frecuencias de PAS

```r
# Datos de presión arterial sistólica (PAS) en pacientes diabéticos
pas_diabeticos <- c(120, 125, 118, 130, 122, 128, 115, 135, 124, 129)

# Crear tabla de frecuencias
tabla_freq <- table(pas_diabeticos)
freq_rel <- as.numeric(tabla_freq) / length(pas_diabeticos)
freq_acum <- cumsum(as.numeric(tabla_freq))
freq_acum_rel <- cumsum(freq_rel)

# Crear data frame
result <- data.frame(
  PAS = as.numeric(names(tabla_freq)),
  h = as.numeric(tabla_freq),
  f = round(freq_rel, 2),
  H = freq_acum,
  F = round(freq_acum_rel, 2)
)
print(result, row.names = FALSE)
```

```
 PAS h   f  H   F
 115 1 0.1  1 0.1
 118 1 0.1  2 0.2
 120 1 0.1  3 0.3
 122 1 0.1  4 0.4
 124 1 0.1  5 0.5
 125 1 0.1  6 0.6
 128 1 0.1  7 0.7
 129 1 0.1  8 0.8
 130 1 0.1  9 0.9
 135 1 0.1 10 1.0
```

Ejemplo 1.4: BioEstatR: freq()

La función `freq()` genera automáticamente la tabla de frecuencias con proporciones y acumuladas (documentación completa en Sección B.2):

Cuadro 1.2 Código R

```
library(BioEstatR)

# Tabla de frecuencias de la edad - pacientes diabéticos, UGR
freq(osteo$edad, grf = FALSE)
```

```
Distribución de frecuencias
--------------------------------
Variable:  osteo$edad
n = 94

        x Freq  Prop Prop.Acum
1 (18,23]   27 0.287     0.287
2 (23,28]   22 0.234     0.521
3 (28,33]   14 0.149     0.670
4 (33,38]   15 0.160     0.830
5 (38,43]    4 0.043     0.873
6 (43,48]    5 0.053     0.926
7 (48,53]    3 0.032     0.958
8 (53,58]    3 0.032     0.990
```

El 52.1 % de los pacientes tiene 28 años o menos (la columna `Prop.Acum` del intervalo `(23,28]` muestra `0.521`). Con `grf = TRUE` (por defecto) también se genera el histograma. Con `cuts = 5` agrupa en 5 clases de igual amplitud.

1.5.4. Función de Distribución Empírica (ECDF)

Definición: Función de Distribución Empírica (ECDF)

La **Función de Distribución Empírica** (ECDF) asigna a cada valor x la proporción de observaciones en la muestra que son menores o iguales a x.
Para datos no agrupados, la ECDF es una **función escalonada** que salta en cada valor observado:

$$F(x) = \begin{cases} 0 & \text{si } x < x_{(1)} \\ \frac{i}{n} & \text{si } x_{(i)} \leq x < x_{(i+1)} \\ 1 & \text{si } x \geq x_{(n)} \end{cases}$$

donde $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ son los valores ordenados de la muestra.

Ejemplo 1.5: Visualización: ECDF Escalonada (Datos de PAS)

Usando los datos de PAS del Ejemplo 1.3, observamos cómo la función acumula probabilidad en saltos de $1/n$ en cada valor observado:

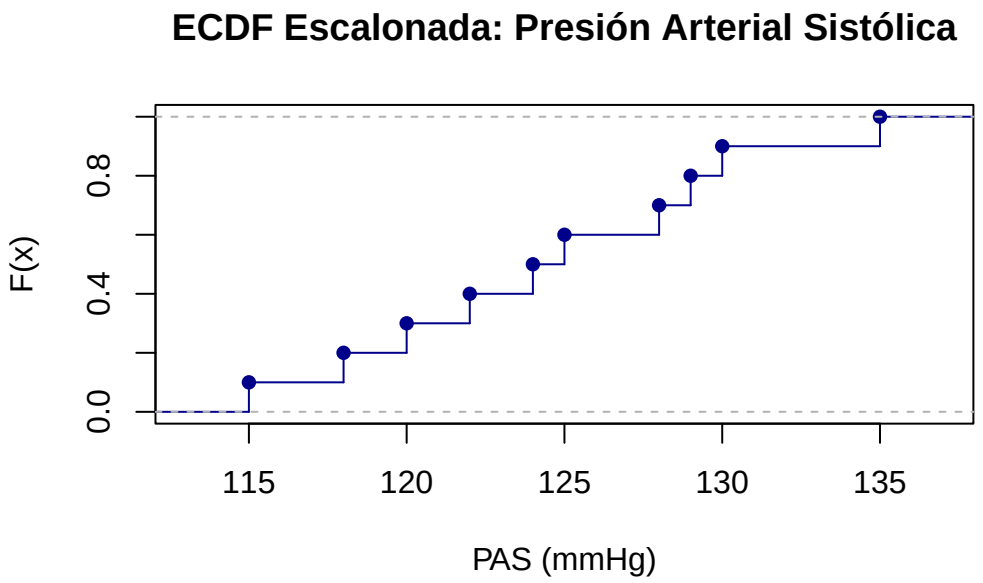


Figura 1.1: ECDF Escalonada para datos de PAS

Propiedad Importante: ECDF

De la ECDF podemos recuperar las frecuencias relativas mediante:

$$f(x_j) = F(x_j) - F(x_{j-1}) \quad \text{para } j = 1, \ldots, k$$

con la convención $F(x_0) = 0$.

Ejemplo 1.6: Tiempo de Supervivencia de Pacientes

Se estudió el tiempo de supervivencia (en días) tras el diagnóstico de una patología agresiva en una cohorte de 100 pacientes. Los datos se agrupan en clases:

Intervalo (días)	h_j	f_j	H_j	F_j
0 – 100	1	0.01	1	0.01
100 – 500	24	0.24	25	0.25
500 – 1000	45	0.45	70	0.70
1000 – 4000	30	0.30	100	1.00

Total	**100**	**1.00**

1.5.5. Interpolación Lineal para Datos Agrupados

Para datos agrupados, no conocemos los valores exactos dentro de cada clase. Por ello, aproximamos la ECDF mediante una **interpolación lineal** entre los límites de las clases. Esto genera una línea quebrada llamada **polígono de frecuencias acumuladas** (u ojiva).
La fórmula general de interpolación para un valor x dentro de la clase $j = [x_j^l, x_j^u]$ es:

$$F(x) = F(x_j^l) + \frac{x - x_j^l}{x_j^u - x_j^l} \cdot f(x_j)$$

Donde:

- x_j^l y x_j^u son los límites inferior y superior de la clase.
- $F(x_j^l)$ es la frecuencia relativa acumulada hasta la clase anterior.
- $f(x_j)$ es la frecuencia relativa de la clase actual.

Nota: Esta fórmula es la base para estimar **cuantiles** en datos agrupados. Si despejamos x para un valor de $F(x) = p$, obtenemos el valor del cuantil correspondiente.
Aplicando esto a nuestros datos:

$$F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \frac{x-0}{100} \cdot 0.01 & \text{si } 0 < x \leq 100 \\ 0.01 + \frac{x-100}{400} \cdot 0.24 & \text{si } 100 < x \leq 500 \\ 0.25 + \frac{x-500}{500} \cdot 0.45 & \text{si } 500 < x \leq 1000 \\ 0.70 + \frac{x-1000}{3000} \cdot 0.30 & \text{si } 1000 < x \leq 4000 \\ 1 & \text{si } x > 4000 \end{cases}$$

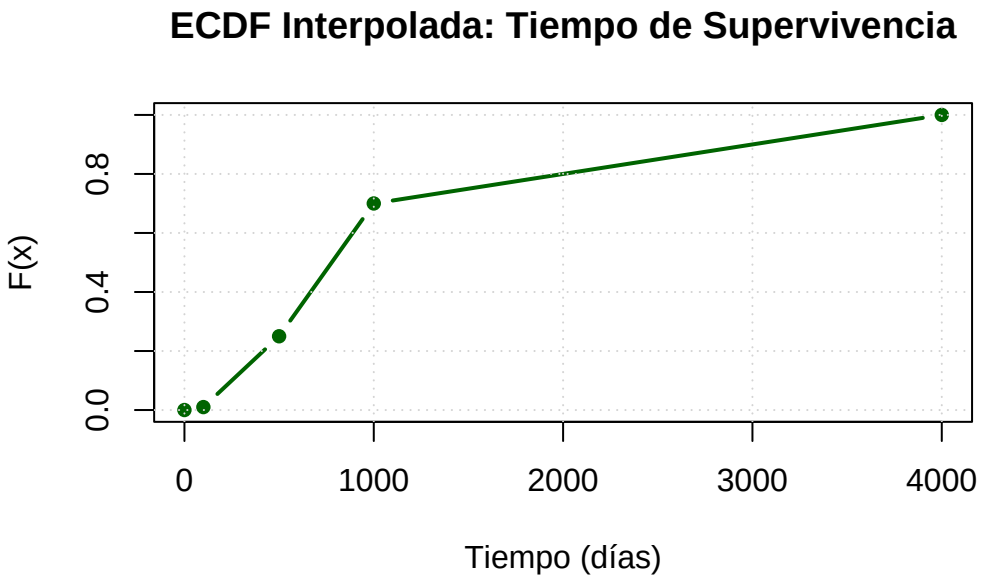


Figura 1.2: Polígono de frecuencias acumuladas (ECDF interpolada)

1.6. Histogramas y Gráficos de Densidad

Para variables continuas agrupadas en clases, el histograma es la forma estándar de visualización.

1.6.1. Construcción de Histogramas

Definición: Histograma

Un **histograma** es una representación gráfica de una distribución de frecuencias para datos agrupados. Se construye de la siguiente manera:

- **Eje X**: límites de las clases x_j^l, x_j^u
- **Eje Y**: densidad de frecuencia (frecuencia dividida por la amplitud de la clase)

Densidad de frecuencia absoluta:

$$\hat{h}(x_j) = \frac{h(x_j)}{\Delta x_j} = \frac{h(x_j)}{x_j^u - x_j^l}$$

Densidad de frecuencia relativa:

$$\hat{f}(x_j) = \frac{f(x_j)}{\Delta x_j} = \frac{f(x_j)}{x_j^u - x_j^l}$$

Propiedad fundamental: El **área** de cada rectángulo representa la frecuencia (relativa o absoluta), y el área total del histograma es igual a n (si usamos densidades absolutas) o 1 (si usamos densidades relativas).

$$\sum_{j=1}^{k} \hat{f}(x_j) \cdot \Delta x_j = \sum_{j=1}^{k} f(x_j) = 1$$

Ejemplo 1.7: Índice de Masa Corporal (IMC)

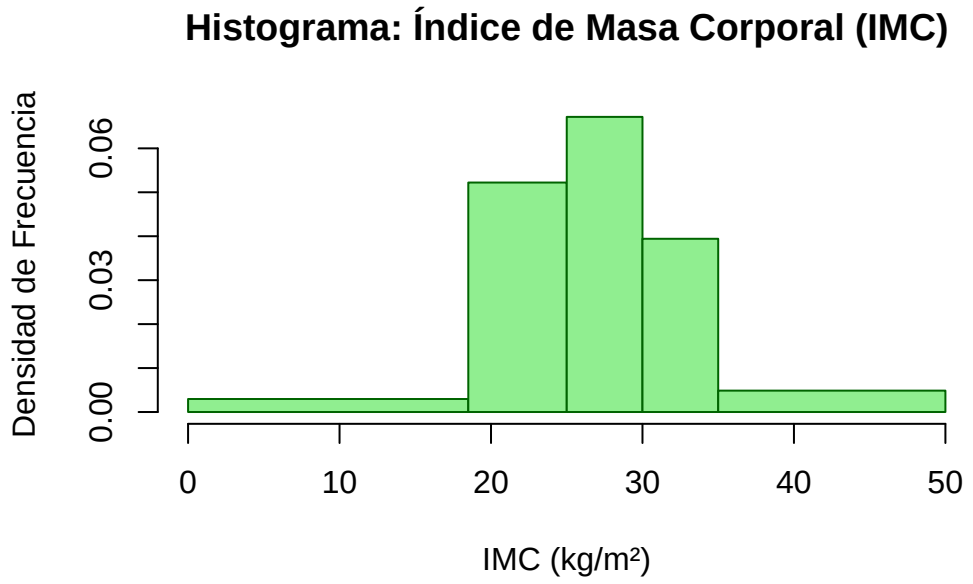
Se recopilaron los IMC de 274 pacientes. Los datos se clasifican en intervalos según los criterios de la OMS:

IMC (kg/m²)	h_j	f_j	Δx_j	$\hat{f}(x_j)$
0 – 18.5 (Bajo peso)	15	0.055	18.5	0.0030
18.5 – 25 (Normal)	93	0.339	6.5	0.0522
25 – 30 (Sobrepeso)	92	0.336	5	0.0672
30 – 35 (Obesidad I)	54	0.197	5	0.0394
35 – 50 (Obesidad II/III)	20	0.073	15	0.0049
Total	**274**	**1.000**		

Interpretación: La densidad más alta está en la clase de "Sobrepeso", indicando que ese intervalo tiene la mayor concentración de pacientes en esta muestra.

Cuadro 1.3 Código R para histograma de IMC

```r
# Simular datos de IMC basados en la tabla
imc <- c(runif(15, 15, 18.5), runif(93, 18.5, 25), runif(92, 25, 30),
         runif(54, 30, 35), runif(20, 35, 50))

# Crear histograma con densidad
hist(imc, breaks = c(0, 18.5, 25, 30, 35, 50),
     freq = FALSE,
     main = "Histograma: Índice de Masa Corporal (IMC)",
     xlab = "IMC (kg/m²)",
     ylab = "Densidad de Frecuencia",
     col = "lightgreen",
     border = "darkgreen")
```



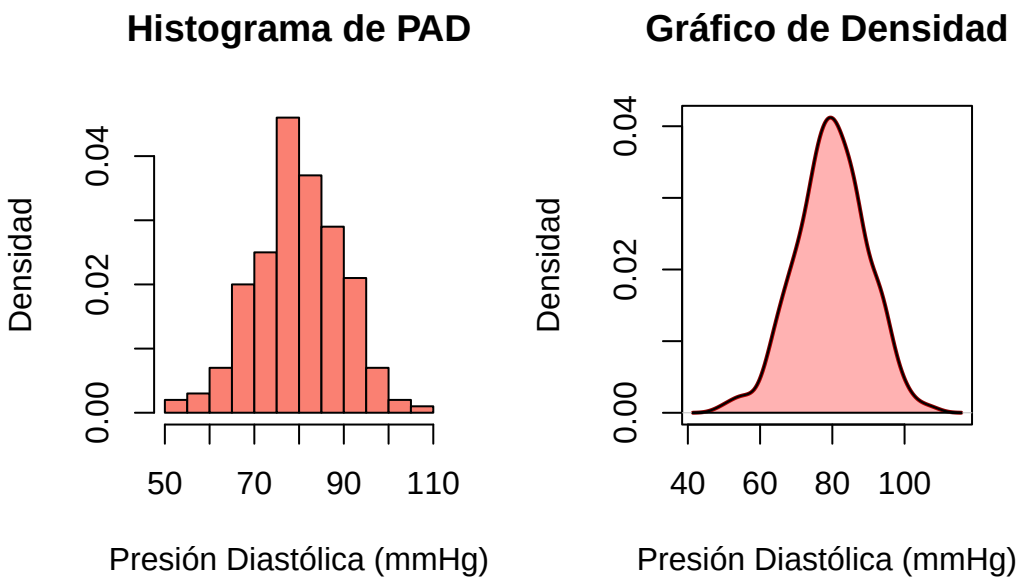
:::

1.6.2. Histograma vs Gráfico de Densidad

Ejemplo 1.8: Visualización en R

Cuadro 1.4 Comparación de gráficos

```r
# Simular datos de Presión Arterial Diastólica (en mmHg)
set.seed(42)
pad <- rnorm(200, mean = 80, sd = 10)

# Configurar área de trazado para 2 gráficos
par(mfrow = c(1, 2))

# Histograma
hist(pad, breaks = 20, freq = FALSE,
     main = "Histograma de PAD",
     xlab = "Presión Diastólica (mmHg)",
     ylab = "Densidad",
     col = "salmon")

# Gráfico de densidad suavizada
plot(density(pad),
     main = "Gráfico de Densidad",
     xlab = "Presión Diastólica (mmHg)",
     ylab = "Densidad",
     lwd = 2, col = "darkred")
polygon(density(pad), col = rgb(1, 0, 0, 0.3))
```



Cuadro 1.5 Comparación de gráficos

```
# Restaurar configuración
par(mfrow = c(1, 1))
```

1.7. Clasificación de Medidas Estadísticas

Para describir una distribución, utilizamos diferentes medidas que resumen la información.

Clasificación: Tendencia Central vs. Posición

- **Medidas de Tendencia Central**: Buscan identificar el "valor representativo" o centro de la distribución.
  - **Media** ($\bar{x}$): Promedio aritmético.
  - **Mediana** ($\tilde{x}$): Valor central (posición central).
  - **Moda** (Mo): Valor más frecuente.
- **Medidas de Posición (No central)**: Indican la posición de un valor respecto al conjunto de datos, permitiendo dividir la distribución en partes iguales.
  - **Cuartiles** (Q_1, Q_2, Q_3): Dividen los datos en 4 partes iguales.
  - **Percentiles** (P_k): Dividen los datos en 100 partes iguales.
- **Sesgo (Skewness)**: Medida de asimetría.
  - **Simétrica**: Media $\approx$ Mediana.
  - **Sesgo Derecha (Positivo)**: Media $>$ Mediana.
  - **Sesgo Izquierda (Negativo)**: Media $<$ Mediana.
- **Curtosis (Kurtosis)**: Medida del grado de apuntamiento o concentración en las colas.

1.8. Medidas de Posición Central

Las medidas de posición central nos indican dónde se concentran los datos. Son fundamentales para resumir la tendencia central de una distribución.

1.8.1. Media Aritmética

Definición: Media Aritmética

La **media aritmética** de una muestra $x_1, x_2, \ldots, x_n$ es:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Representa el "centro de gravedad" de los datos. Es la suma de todos los valores dividida por el número total de observaciones.
Para datos en una distribución de frecuencias:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} x_j h(x_j) = \sum_{j=1}^{k} x_j f(x_j)$$

donde x_j es el valor distintivo y $h(x_j)$ es su frecuencia absoluta.

⚠ Propiedades Algebraicas de la Media

1. **Propiedad de desviación nula**:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

   Las desviaciones positivas y negativas de la media se cancelan.
2. **Propiedad de suma**: Si $z_i = x_i + y_i$, entonces:

$$\bar{z} = \bar{x} + \bar{y}$$

1.8.2. Media para Datos Agrupados

ℹ Definición: Media Aproximada para Datos Binned

Cuando solo disponemos de datos agrupados en clases, aproximamos la media usando la marca de clase. Equivalentemente, en términos de frecuencias absolutas n_j o relativas $f_j = n_j/n$:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} x_j \cdot n_j \;=\; \sum_{j=1}^{k} x_j \cdot f_j$$

donde x_j es la **marca de clase** (punto medio del intervalo j):

$$x_j = \frac{x_j^l + x_j^u}{2}$$

y n_j es el número de observaciones en la clase j (f_j la frecuencia relativa correspondiente). El Ejemplo 1.9 utiliza la versión con frecuencias relativas.
Nota: Esta es una aproximación porque asumimos que todas las observaciones en una clase están concentradas en su punto medio.

Ejemplo 1.9: Tiempo de Estancia Hospitalaria

Se estudia el tiempo de estancia (en días) de una muestra de pacientes en una unidad de cuidados intensivos. La distribución es:

Intervalo (días)	Marca de Clase (x_j)	$f(x_j)$	Contribución
1 – 3	2	0.20	$2 \times 0.20 = 0.40$
3 – 7	5	0.45	$5 \times 0.45 = 2.25$
7 – 14	10.5	0.25	$10.5 \times 0.25 = 2.625$
14 – 30	22	0.08	$22 \times 0.08 = 1.76$
30 – 60	45	0.02	$45 \times 0.02 = 0.90$

Media aproximada:

$$\bar{x} = 0.40 + 2.25 + 2.625 + 1.76 + 0.90 = 7.935 \text{ días}$$

Cuadro 1.6 Cálculo de estancia media en R

```
# Datos
marca_clase <- c(2, 5, 10.5, 22, 45)
frecuencia_rel <- c(0.20, 0.45, 0.25, 0.08, 0.02)

# Media
media_estancia <- sum(marca_clase * frecuencia_rel)
print(paste("Media de estancia aproximada:",
            round(media_estancia, 3), "días"))
```

```
[1] "Media de estancia aproximada: 7.935 días"
```

1.8.3. Moda

Definición: Moda

La **moda** x_D es el valor más frecuente en un conjunto de observaciones:

$$x_D = \{x_j \mid h(x_j) = \max_k h(x_k) \text{ o } f(x_j) = \max_k f(x_k)\}$$

Es útil para variables nominales, ordinales, discretas o clasificadas, pero **no es apropiada para variables continuas**.
Para datos agrupados, hablamos de la **clase modal** (la clase con la frecuencia más alta) y estimamos la moda como la marca de clase.

Ejemplo 1.10: Moda en Tiempo de Supervivencia

Usando la tabla de supervivencia del Ejemplo 1.6:

Intervalo (días)	h_j
0 – 100	1
100 – 500	24
500 – 1000	**45** (máximo)
1000 – 4000	30

- **Clase modal**: 500 – 1000 días
- **Moda estimada**: $x_D = 750$ días (marca de clase)

1.8.4. Mediana

Definición: Mediana

La **mediana** $x_{0.5}$ es el valor que divide los datos ordenados en dos partes de igual tamaño: 50 % de las observaciones están por debajo y 50 % están por encima.

Para datos no agrupados:

- Si n es **impar**:

$$x_{0.5} = x_{(\frac{n+1}{2})}$$

- Si n es **par**:

$$x_{0.5} = \frac{1}{2}\left\{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right\}$$

donde $x_{(i)}$ es la i-ésima observación cuando los datos están **ordenados de menor a mayor**.

Ventaja: La mediana es **robusta** (insensible a outliers), a diferencia de la media.

Ejemplo 1.11: Cálculo de Mediana

Conjunto (n par): 115, 118, 120, 122, 124, 125, 128, 129, 130, 135

Datos ya ordenados. Con $n = 10$ (par):

$$x_{0.5} = \frac{1}{2}(x_{(5)} + x_{(6)}) = \frac{1}{2}(124 + 125) = 124.5 \text{ mmHg}$$

Cuadro 1.7 Cálculo de mediana en R

```
pas <- c(120, 125, 118, 130, 122, 128, 115, 135, 124, 129)
mediana_pas <- median(pas)
print(paste("Mediana de PAS:", mediana_pas, "mmHg"))
```

```
[1] "Mediana de PAS: 124.5 mmHg"
```

1.8.5. Relación entre Media, Mediana y Moda

La siguiente figura ilustra la posición relativa de la media, mediana y moda según el sesgo de la distribución:

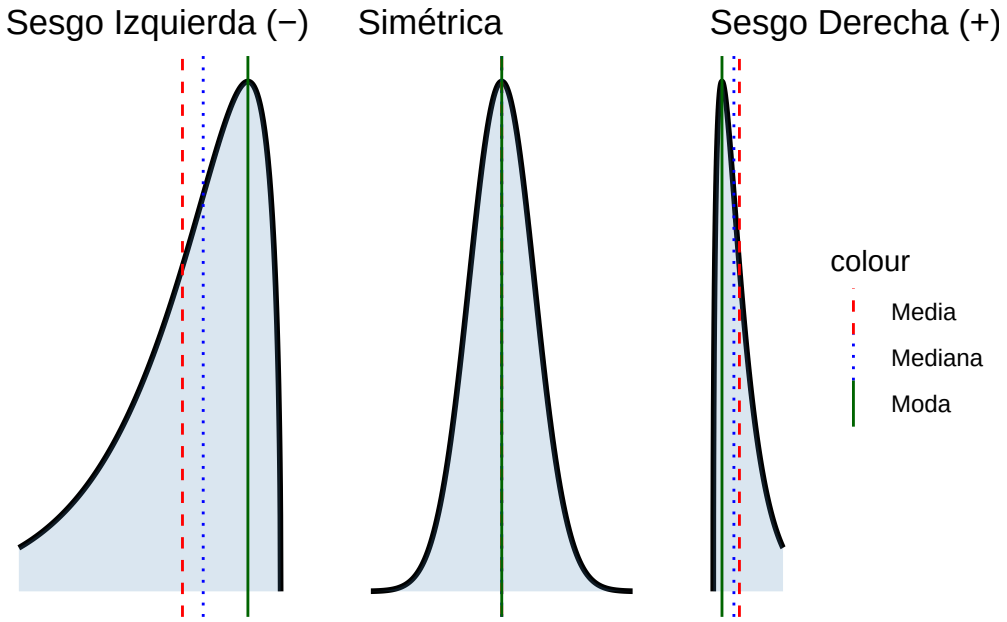


Figura 1.3: Relación entre media, mediana y moda según el sesgo

1.9. Medidas de Forma: Sesgo y Curtosis

Además de la tendencia central y la dispersión, la **forma** de la distribución es crucial para entender el comportamiento de los datos.

1.9.0.1. Coeficiente de Asimetría (Sesgo)

Mide el grado de simetría de la distribución respecto a su media. El coeficiente de asimetría de Fisher se define como:

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

donde s es la desviación típica.

Interpretación:

- $g_1 = 0$: La distribución es **simétrica**.
- $g_1 > 0$: **Asimetría positiva** (sesgo a la derecha). La cola derecha es más larga.

- $g_1 < 0$: **Asimetría negativa** (sesgo a la izquierda). La cola izquierda es más larga.

1.9.0.2. Curtosis o Apuntamiento

Mide el grado de concentración de los valores alrededor de la zona central de la distribución y el "peso" de las colas. La **curtosis de Pearson** se define como:

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4}$$

En muchos paquetes estadísticos (incluido R) se utiliza también la **curtosis en exceso** o **curtosis de Fisher**, $g_2 - 3$, que centra la referencia en cero para la distribución normal.

Interpretación (Curtosis en exceso):

- $g_2 - 3 = 0$: **Mesocúrtica**. El apuntamiento es similar al de una distribución normal.
- $g_2 - 3 > 0$: **Leptocúrtica**. La distribución es más apuntada y tiene colas más gruesas ("pesadas") que la normal.
- $g_2 - 3 < 0$: **Platicúrtica**. La distribución es más plana y tiene colas más delgadas que la normal.

1.9.1. Cálculo con R

Cuadro 1.8 Cálculo de estadísticos básicos

```r
library(moments)
library(dplyr)

# Datos de ejemplo
datos <- c(120, 125, 118, 130, 122, 128, 115, 135, 124, 129, 150)

# Resumen básico
resumen <- data.frame(
  Media = mean(datos),
  Mediana = median(datos),
  Sesgo = skewness(datos), # g1
  Curtosis = kurtosis(datos), # g2 (Pearson, restar 3 para exceso)
  Curtosis_Exceso = kurtosis(datos) - 3
)
print(resumen)
```

```
     Media Mediana    Sesgo Curtosis Curtosis_Exceso
1 126.9091     125 1.191328  4.14324         1.14324
```

1.9.2. Cuantiles

Definición: Cuantiles

El **cuantil** x_p es el valor que divide los datos ordenados en la proporción p a $(1-p)$, donde $0 \leq p \leq 1$.

- Debajo del cuantil x_p se encuentran aproximadamente $p \times 100\,\%$ de los datos
- Arriba del cuantil x_p se encuentran aproximadamente $(1-p) \times 100\,\%$ de los datos

Cuantiles especiales:

- **Cuartiles**: $p = 0.25, 0.50, 0.75$ (dividen en 4 partes)
  - $x_{0.25}$ = primer cuartil (Q1)
  - $x_{0.50}$ = mediana
  - $x_{0.75}$ = tercer cuartil (Q3)
- **Deciles**: $p = 0.1, 0.2, \ldots, 0.9$ (dividen en 10 partes)
- **Percentiles**: $p = 0.01, 0.02, \ldots, 0.99$ (dividen en 100 partes)

Ejemplo 1.12: Cuartiles de Supervivencia

Con la **distribución agrupada** de supervivencia del Ejemplo 1.6, podemos aproximar los cuartiles usando interpolación en la ECDF. *Nota:* el chunk de R a continuación simula datos **inspirados** en esa distribución (no idénticos), por lo que los valores observados son cercanos pero no exactamente iguales a los obtenidos por la fórmula analítica.

Para el primer cuartil ($p = 0.25$):

- $F(x_j^l) = 0.01$ (límite inferior de clase 100–500)
- $f(x_j) = 0.24$ (frecuencia relativa de clase 100–500)
- Fórmula: $x_{0.25} = x_j^l + \frac{p - F(x_j^l)}{f(x_j)} \cdot \Delta x_j$

$$x_{0.25} = 100 + \frac{0.25 - 0.01}{0.24} \times 400 = 100 + 400 = 500 \text{ días}$$

```
[1] "Q1 (25%): 500.37"
```

```
[1] "Q2 (50%, mediana): 828.9"
```

```
[1] "Q3 (75%): 1086.7"
```

Visualización: Diagrama de Caja (Boxplot)

El **diagrama de caja** (boxplot) es una herramienta robusta para visualizar la distribución, la mediana, los cuartiles y detectar valores atípicos (outliers) de forma intuitiva.

Cuadro 1.9 Cálculo de cuantiles

```
# Generamos los mismos datos que en el Ejemplo 1.4.2
set.seed(42)
supervivencia <- c(
  runif(1, 0, 100),
  runif(24, 100, 500),
  runif(45, 500, 1000),
  runif(25, 1000, 2000),
  3200, 3500, 3700, 3800, 3950
)

Q1 <- quantile(supervivencia, probs = 0.25)
Q2 <- quantile(supervivencia, probs = 0.50)
Q3 <- quantile(supervivencia, probs = 0.75)

print(paste("Q1 (25%):", round(Q1, 2)))
```

Cuadro 1.10 Cálculo de cuantiles

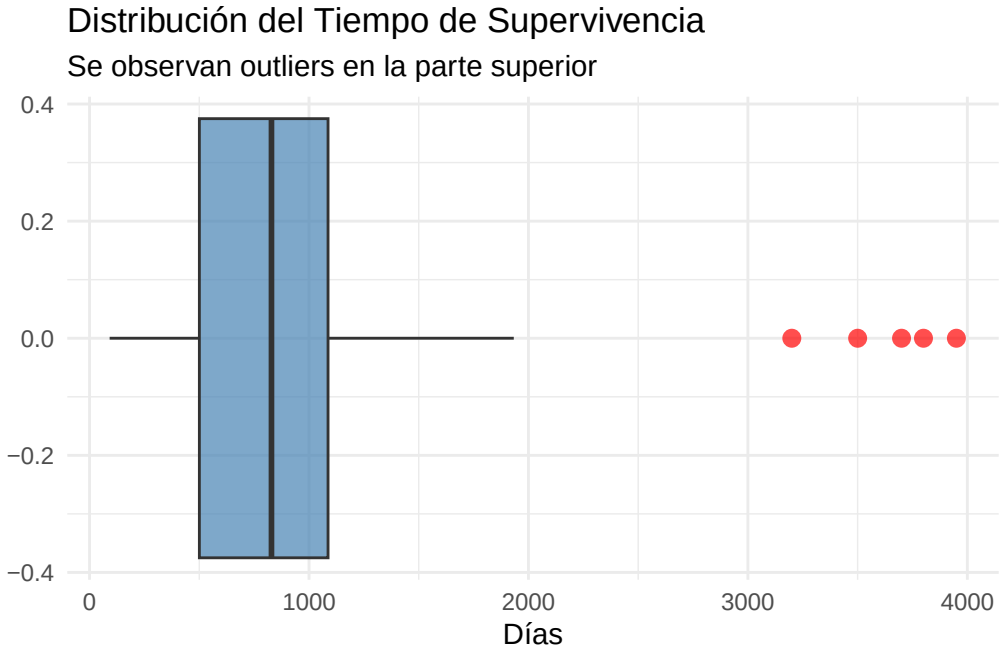
```
print(paste("Q2 (50%, mediana):", round(Q2, 2)))
```

Cuadro 1.12 Boxplot mejorado con ggplot2

```
library(ggplot2)

# Crear un data frame con los datos de supervivencia
df <- data.frame(supervivencia = supervivencia)

# Visualización con ggplot2 y labels para cuartiles
ggplot(df, aes(y = supervivencia)) +
  geom_boxplot(fill = "steelblue", alpha = 0.7,
               outlier.colour = "red", outlier.shape = 16, outlier.size = 3) +
  labs(title = "Distribución del Tiempo de Supervivencia",
       subtitle = "Se observan outliers en la parte superior",
       y = "Días",
       x = "") +
  theme_minimal() +
  coord_flip() # Para orientación horizontal
```

Cuadro 1.11 Cálculo de cuantiles

```
print(paste("Q3 (75%):", round(Q3, 2)))
```



Lectura del Boxplot

Un boxplot divide los datos en:

- **Caja**: Contiene el 50 % central (IQR: de Q_1 a Q_3).
- **Línea central**: Mediana.
- **Bigotes**: Extienden hasta 1.5 veces el IQR.
- **Puntos (outliers)**: Valores fuera del rango de los bigotes.

1.10. Comparación de Medidas de Posición

Robustez de las Medidas de Tendencia Central

La **media** es muy sensible a valores extremos (outliers):

- Fácil de calcular algebraicamente
- Propiedad de suma útil para combinación de datos
- Se ve afectada por datos atípicos

La **mediana** es robusta:

- Insensible a outliers
- Más representativa en distribuciones asimétricas
- Requiere ordenar los datos

La **moda** es útil para datos categóricos pero limitado para continuos.

Recomendación: En presencia de outliers, la mediana es preferible a la media como medida de tendencia central.

1.11. Resumen de Conceptos Clave

1.11.1. Tabla Resumen: Escalas de Medición y Estadísticos Apropiados

Tipo de Variable	Ejemplos	Estadísticos Apropiados
Nominal	Género, grupo sanguíneo	Moda, frecuencias
Ordinal	Satisfacción, grado	Mediana, moda, cuantiles
Discreta	Número de eventos	Media, mediana, varianza
Continua	Peso, presión arterial	Media, mediana, varianza, cuantiles

1.11.2. Tabla Resumen: Fórmulas Importantes

Estadístico	Datos No Agrupados	Datos Agrupados (Intervalos)
Media ($\bar{x}$)	$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\sum_{j=1}^{k} x_j f_j$ (x_j: marca clase)
Mediana/Cuantil (x_p)	Valor en pos. ordenado	$x_j^l + \frac{p-F(x_j^l)}{f(x_j)} \cdot \Delta x_j$
Moda (Mo)	Valor con máx. h_j	Marca de la clase modal
ECDF ($F(x)$)	Función escalonada	Interpolación lineal (ojiva)
Sesgo (g_1)	$\frac{\frac{1}{n}\sum(x_i-\bar{x})^3}{s^3}$	(Aprox. marcas de clase)
Curtosis (g_2)	$\frac{\frac{1}{n}\sum(x_i-\bar{x})^4}{s^4}$	(Aprox. marcas de clase)

1.12. Ejercicios

Ejercicio 1.1: Conceptos Fundamentales

Define con tus propias palabras (desde una perspectiva clínica): a) La diferencia entre una variable y un valor (ej. Glucemia) b) Por qué necesitamos distinguir entre población y muestra en un ensayo clínico c) Qué es un outlier y cómo podría afectar a la media de una serie de mediciones de tensión arterial

Ejercicio 1.2: Identificación de Escalas

Para cada una de las siguientes variables clínicas, identifica la escala de medición (nominal, ordinal, discreta, continua) y justifica tu respuesta:

a) Número de episodios asmáticos al año
b) Gravedad de una quemadura (Primer, Segundo, Tercer grado)
c) Concentración de hemoglobina en sangre (g/dL)
d) Factor Rh (Positivo, Negativo)
e) Escala de coma de Glasgow (puntuación de 3 a 15)

Ejercicio 1.3: Cálculo de la Media

Se registraron los tiempos de espera (en minutos) en urgencias de 8 pacientes: 12, 8, 15, 10, 7, 14, 9, 11

a) Calcula la media aritmética
b) Ordena los datos y encuentra la mediana
c) ¿Cuál es más representativa en este caso? ¿Por qué?

Cuadro 1.13 Solución en R

```r
tiempos <- c(12, 8, 15, 10, 7, 14, 9, 11)
media <- mean(tiempos)
mediana <- median(tiempos)
cat("Media:", media, "\n")
```

```
Media: 10.75
```

Cuadro 1.14 Solución en R

```r
cat("Mediana:", mediana, "\n")
```

```
Mediana: 10.5
```

Ejercicio 1.4: Tabla de Frecuencias

Los siguientes datos representan el número de admisiones hospitalarias previas para 20 pacientes crónicos: 3, 2, 5, 2, 4, 3, 6, 2, 3, 4, 5, 2, 3, 4, 2, 5, 3, 6, 4, 2

a) Construye una tabla de frecuencias (absoluta, relativa, acumulada)
b) ¿Cuál es la moda (número más frecuente de admisiones)?
c) ¿Qué proporción de pacientes tuvieron 3 o menos admisiones?

Ejercicio 1.5: Datos Agrupados

Se agrupan los niveles de glucosa en ayunas (mg/dL) de 100 pacientes en riesgo de diabetes:

Intervalo (mg/dL)	Frecuencia (h_j)
70 – 90	15
90 – 100	32
100 – 110	35
110 – 126	15
126 – 150	3

a) Calcula la media aproximada de la glucemia
b) Identifica la clase modal
c) Estima el percentil 50 (mediana) usando interpolación

Tip

Cuadro 1.15 Solución en R

```
# Datos
marca_clase <- c(80, 95, 105, 118, 138)
frecuencia <- c(15, 32, 35, 15, 3)
n <- sum(frecuencia)

# Media
media_aprox <- sum(marca_clase * frecuencia) / n
print(paste("Media aproximada:", round(media_aprox, 2), "mg/dL"))
```

```
[1] "Media aproximada: 100.99 mg/dL"
```

Cuadro 1.16 Solución en R

```r
# Frecuencia relativa acumulada
freq_rel <- frecuencia / n
freq_acum <- cumsum(freq_rel)

# Clase que contiene la mediana (F >= 0.5)
print(data.frame(Clase = c("70-90", "90-100", "100-110", "110-126", "126-150"),
                 Frecuencia = frecuencia,
                 F_acum = freq_acum))
```

```
    Clase Frecuencia F_acum
1   70-90         15   0.15
2  90-100         32   0.47
3 100-110         35   0.82
4 110-126         15   0.97
5 126-150          3   1.00
```

Ejercicio 1.6: Interpretación de Histograma

Considera el histograma de IMC del Ejemplo 1.7:

a) ¿Qué clase tiene la mayor densidad de pacientes?

b) ¿Qué porcentaje de pacientes tienen un IMC inferior a 25 (Normal o bajo peso)?

c) ¿En qué rango se encuentra el 25 % de pacientes con menor IMC?

1.13. Respuestas a los Ejercicios

Ejercicio 1.1: Conceptos a) Variable: Característica (Glucemia). Valor: Resultado medido (105 mg/dL). b) Población: Todos los pacientes con la patología. Muestra: Los que participan en el estudio. c) Un outlier (ej. una medición de 220/120 por error de manguito) elevaría artificialmente la media, haciendo que la tensión promedio parezca más alta de lo que realmente es.

Ejercicio 1.2: Escalas a) Discreta (conteo de eventos). b) Ordinal (hay un orden de gravedad). c) Continua (puede tomar cualquier valor decimal). d) Nominal (categorías sin orden). e) Ordinal (puntuación con orden intrínseco).

Ejercicio 1.3: Tiempos de espera a) Media = 86 / 8 = 10.75 min. b) Ordenados: 7, 8, 9, 10, 11, 12, 14, 15. Mediana = (10+11)/2 = 10.5 min. c) Ambas son similares porque no hay outliers extremos.

Ejercicio 1.4: Admisiones a)

Admisiones	h_j	f_j	H_j	F_j
2	6	0.30	6	0.30
3	5	0.25	11	0.55
4	4	0.20	15	0.75
5	3	0.15	18	0.90
6	2	0.10	20	1.00
Total	**20**	**1.00**		

b) Moda = 2 admisiones ($h_j = 6$). c) Proporción = $F_3 = 0.55$ (55 %).

Ejercicio 1.5: Glucemia a) Media 100.99 mg/dL (cálculo: (80·15 + 95·32 + 105·35 + 118·15 + 138·3)/100). b) Clase modal: 100-110 mg/dL ($h_j = 35$). c) P 100 + (0.50 - 0.47)/0.35 · 10 100.86 mg/dL (la mediana cae en la clase 100-110, donde F pasa de 0.47 a 0.82).

Ejercicio 1.6: IMC (Ejemplo 1.7) a) Sobrepeso (25-30) con densidad 0.0672. b) $F(25) = 0.055 + 0.339 = 0.394$ (39.4 %). c) En el rango de 0 a 25 (específicamente, el P está en la clase "Normal").

Fin de Semana 1

Para referencias avanzadas sobre análisis exploratorio de datos y estimación robusta, consulta: Robust Estimation of Location

Capítulo 2

Semana 2 — Análisis Exploratorio de Datos (Parte II)

En esta semana continuamos con el Análisis Exploratorio de Datos (AED), enfocándonos en medidas de dispersión, transformaciones lineales, diagramas de caja, análisis bivariante y medidas de dependencia. Estos conceptos son fundamentales para comprender la variabilidad de los datos y las relaciones entre variables.

2.1. Medidas de Dispersión

Mientras que las medidas de tendencia central (como la media y la mediana) describen dónde está el centro de los datos, las **medidas de dispersión** describen cuánta variabilidad existe en los datos. La dispersión es crítica para comprender la calidad de los datos y la precisión de nuestras estimaciones.

2.1.1. ¿Por qué necesitamos medidas de dispersión?

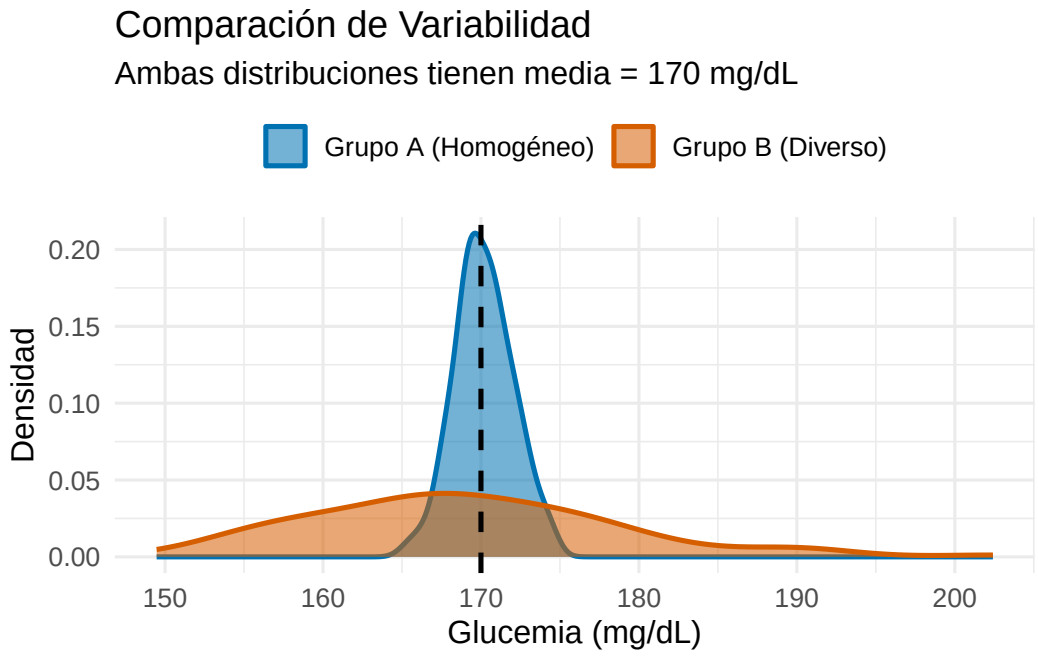
Es fundamental comprender que la media por sí sola no describe una distribución. Dos conjuntos de datos pueden compartir la misma media pero tener estructuras de dispersión radicalmente diferentes.

Consideremos un ejemplo clínico: dos grupos de pacientes con niveles de glucosa en sangre (mg/dL) con la misma media (170 mg/dL):

1. **Grupo A**: Niveles muy homogéneos (todos cercanos a la media).
2. **Grupo B**: Niveles muy diversos (gran dispersión).

Cuadro 2.1 Distribuciones con igual media pero distinta varianza

```r
library(ggplot2)

# Generar datos
set.seed(123)
grupo_a <- rnorm(100, mean = 170, sd = 2)   # Varianza baja
grupo_b <- rnorm(100, mean = 170, sd = 10)  # Varianza alta

df <- data.frame(
  valor = c(grupo_a, grupo_b),
  grupo = rep(c("Grupo A (Homogéneo)", "Grupo B (Diverso)"), each = 100)
)

# Gráfico de densidad con líneas de media (paleta Okabe-Ito)
ggplot(df, aes(x = valor, fill = grupo, color = grupo)) +
  geom_density(alpha = 0.55, linewidth = 0.9) +
  scale_fill_manual(values  = c("Grupo A (Homogéneo)" = "#0072B2",
                                "Grupo B (Diverso)"   = "#D55E00")) +
  scale_color_manual(values = c("Grupo A (Homogéneo)" = "#0072B2",
                                "Grupo B (Diverso)"   = "#D55E00")) +
  geom_vline(xintercept = 170, linetype = "dashed", color = "black", linewidth = 0.9) +
  labs(title = "Comparación de Variabilidad",
       subtitle = "Ambas distribuciones tienen media = 170 mg/dL",
       x = "Glucemia (mg/dL)",
       y = "Densidad",
       fill = NULL, color = NULL) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "top")
```



Como se observa en el gráfico (la línea discontinua representa la media compartida), el **Grupo B** presenta una dispersión mucho mayor, lo que demuestra que el promedio no siempre es suficiente para caracterizar la realidad de un conjunto de datos médicos. :::

Resumen: Medidas de Dispersión

Para cuantificar la variabilidad de un conjunto de datos, utilizamos principalmente las siguientes medidas:

- **Rango**: Diferencia entre el valor máximo y mínimo. Muy sensible a outliers.
- **Rango Intercuartílico (IQR)**: Diferencia entre el tercer y primer cuartil. Mide la dispersión del 50 % central de los datos. Robusto a outliers.
- **Varianza** (s^2): Promedio de las desviaciones al cuadrado respecto a la media.
- **Desviación Estándar** (s): Raíz cuadrada de la varianza. Expresada en las mismas unidades que la variable original, lo que facilita su interpretación clínica.
- **Coeficiente de Variación (CV)**: Medida adimensional de dispersión relativa que permite comparar la variabilidad de diferentes variables clínicas.

2.1.2. Rango

Rango

El **rango** es la diferencia entre la observación más grande y la más pequeña:

$$R = x_{\text{máx}} - x_{\text{mín}} = x_{(n)} - x_{(1)}$$

donde $x_{(1)}, \ldots, x_{(n)}$ son las observaciones ordenadas.

Limitaciones del Rango

El rango es fácil de calcular, pero tiene importantes limitaciones:

- Es muy sensible a valores atípicos (outliers)
- Su valor aumenta con el tamaño de la muestra
- No proporciona información sobre cómo se distribuyen los datos entre los extremos

2.1.3. Rango Intercuartílico (IQR)

Rango Intercuartílico

El **rango intercuartílico** (IQR, del inglés Interquartile Range) es la diferencia entre el tercer cuartil ($Q3$) y el primer cuartil ($Q1$):

$$\text{IQR} = Q3 - Q1$$

El IQR contiene el 50 % central de los datos y es robusto a valores atípicos.

2.1.4. Varianza y Desviación Estándar

Varianza y Desviación Estándar

La **varianza muestral** mide la dispersión promedio de los datos respecto a la media:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Alternativamente, puede calcularse como:

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

La **desviación estándar** es la raíz cuadrada positiva de la varianza:

$$s = \sqrt{s^2}$$

Nota: Se utiliza $n-1$ (en lugar de n) para proporcionar un estimador insesgado de la varianza poblacional.

Interpretación de la Desviación Estándar

La desviación estándar está en las mismas unidades que los datos originales, lo que la hace más interpretable que la varianza. Por ejemplo, si los datos representan edades en años, s también está en años. Aproximadamente el 68 % de los datos caen dentro de una desviación estándar de la media, y el 95 % dentro de dos desviaciones estándar (regla empírica para datos aproximadamente normales).

Ejemplo 2.1: BioEstatR grps()

La función `grps()` calcula estadísticos descriptivos comparativos por grupos en una sola llamada (documentación completa en Sección B.2):

Cuadro 2.2 Código R

```
library(BioEstatR)

# IMC según sexo - pacientes diabéticos UGR (n = 94)
grps(osteo$imc, osteo$sexo, grf = FALSE)
```

```
        n  media    dt
Hombre 45 23.514 2.890
```

```
Mujer  49 24.294 4.389
```

Las mujeres presentan mayor IMC medio y mayor variabilidad (dt = 4.39 vs 2.89 en hombres). Con `ic = TRUE` añade intervalos de confianza individuales por grupo.

2.1.5. Cálculo de Estadísticos con R Base

Aunque paquetes como `BioEstatR` son potentes, es fundamental saber usar las funciones básicas de R para un análisis rápido.

Cuadro 2.3 Cálculos con Base R

```r
# Datos: PAS de pacientes (mmHg)
pas <- c(120, 125, 118, 130, 122, 128, 115, 135, 124, 129)

# Cálculo de varianza y desviación estándar
varianza <- var(pas)
desviacion_std <- sd(pas)
iqr_pas <- IQR(pas)

# Cálculo del Coeficiente de Variación (CV)
# El CV es (sd / media) * 100
cv <- (desviacion_std / mean(pas)) * 100

cat("Varianza (s²):", round(varianza, 2), "\n")
```

```
Varianza (s²): 36.93
```

Cuadro 2.4 Cálculos con Base R

```r
cat("Desv. Est. (s):", round(desviacion_std, 2), "\n")
```

```
Desv. Est. (s): 6.08
```

Cuadro 2.5 Cálculos con Base R

```r
cat("IQR:", round(iqr_pas, 2), "\n")
```

```
IQR: 8.25
```

```
Coef. Variación (CV): 4.88 %
```

Coeficiente de Variación (CV) en Contexto Médico

El **Coeficiente de Variación (CV)** es una medida de dispersión **adimensional** que permite comparar la variabilidad de variables con diferentes unidades o magnitudes. Se define como:

$$CV = \left(\frac{s}{\bar{x}}\right) \times 100$$

En medicina, es crucial:

Cuadro 2.6 Cálculos con Base R

```
cat("Coef. Variación (CV):", round(cv, 2), "%\n")
```

- **Interpretación**: Un CV bajo indica alta precisión o consistencia en la medición.
- **Comparación**: Por ejemplo, si comparamos la variabilidad de la glucosa en sangre (mg/dL) con la variabilidad de la frecuencia cardíaca (latidos/min), el CV nos permite saber cuál medida es más "relativamente" dispersa, ignorando las unidades de medida.

2.2. Transformaciones Lineales: Propiedades

Cuando transformamos los datos mediante una transformación lineal, también transformamos sus estadísticos de una manera predecible y matemáticamente bien definida. Generalmente se usan para simplificar la interpretación de datos, cambiar unidades de medida y facilitar la comparación entre variables.

2.2.1. Transformación Lineal Estándar

⚠ Resultado Importante: Transformación Lineal

Si definimos una transformación lineal $y_i = a + bx_i$ donde $b \neq 0$, entonces:
Para la media:

$$\bar{y} = a + b\bar{x}$$

Para la varianza:

$$s_y^2 = b^2 s_x^2$$

Para la desviación estándar:

$$s_y = |b| s_x$$

Para el rango:

$$R_Y = |b| R_X$$

Para el rango intercuartílico:

$$\text{IQR}_Y = |b| \text{IQR}_X$$

2.2.2. Interpretación de los Parámetros

El parámetro a representa un **desplazamiento** de los datos (shift), mientras que el parámetro b controla la escala:

- Si $0 < b < 1$: compresión de los datos (se hacen más próximos)
- Si $b > 1$: dilatación de los datos (se separan más)
- Si $b < 0$: reflexión (inversión) además de escalado
- Si $b = 1$: solo desplazamiento (la dispersión no cambia)

Ejemplo 2.2: Conversión de Temperaturas

Supongamos que tenemos temperaturas en grados Celsius con media $\bar{x} = 25$ y desviación estándar $s_x = 3$. Queremos convertirlas a Fahrenheit usando $y = 32 + 1.8x$.
Entonces:

- $\bar{y} = 32 + 1.8 \times 25 = 77°\text{F}$
- $s_y = |1.8| \times 3 = 5.4°\text{F}$

La media se desplaza y escala, pero la dispersión relativa (en términos de desviaciones estándar) permanece igual.

2.3. Estandarización y Normalización

La estandarización es una transformación lineal específica que convierte los datos a una escala con media 0 y desviación estándar 1. Esto se logra mediante la fórmula:

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

Desarrollando la expresión:

$$z_i = \frac{x_i}{s_x} - \frac{\bar{x}}{s_x}$$

Por lo tanto, en la transformación lineal $y_i = a + bx_i$ para el z-scocre tenemos:

- $$a = -\frac{\bar{x}}{s_x}$$
- $$b = \frac{1}{s_x}$$

2.3.1. Propiedades de los Z-Scores

Propiedades de los Z-Scores

- Media: $\bar{z} = 0$
- Desviación estándar: $s_z = 1$
- Rango típico: aproximadamente $[-3, 3]$ para datos normales
- Un valor $z_i > 3$ o $z_i < -3$ sugiere un posible valor atípico

2.3.2. Interpretación

Ejemplo 2.3: Interpretación de Z-Scores

Si un estudiante obtiene una puntuación de 85 en un examen con media 80 y desviación estándar 5, su z-score es:

$$z = \frac{85 - 80}{5} = 1$$

Esto significa que su puntuación está 1 desviación estándar **por encima** de la media. Si otro estudiante obtiene 70, su z-score sería:

$$z = \frac{70 - 80}{5} = -1$$

Los z-scores permiten comparar puntuaciones de diferentes escalas o distribuciones.

Estandarización automática en R: `scale()`

Aunque podemos calcular z-scores manualmente, R ofrece la función `scale()` que estandariza automáticamente un vector:

```
# Ejemplo numérico
presion <- c(120, 125, 110, 130, 140)

# Estandarización: (x - mean(x)) / sd(x)
presion_z <- scale(presion)

# Resultados
print(presion)
```

```
[1] 120 125 110 130 140
```

```
print(round(presion_z, 2))
```

```
      [,1]
[1,] -0.45
[2,]  0.00
```

```
[3,] -1.34
[4,]  0.45
[5,]  1.34
attr(,"scaled:center")
[1] 125
attr(,"scaled:scale")
[1] 11.18034
```

La función `scale()` resta la media (`mean(presion) = 125`) y divide por la desviación estándar (`sd(presion)  11.18`), transformando directamente los datos a z-scores.

2.3.3. Z-scores con Parámetros Poblacionales

Cuando conocemos la media poblacional (μ) y la desviación típica poblacional (σ), podemos calcular la probabilidad (porcentaje de personas) de que un valor caiga en un rango utilizando la función `pnorm()`.

Ejemplo 2.4: Glucemia (Z-scores con Parámetros Poblacionales)

Supongamos que la glucemia en adultos sanos sigue una distribución normal con $\mu = 90$ mg/dL y $\sigma = 10$ mg/dL. ¿Qué porcentaje de personas tiene una glucemia entre 80 y 100 mg/dL?
Cálculo manual mediante Z-score: 1. Para $x = 80$: $z_1 = \frac{80-90}{10} = -1.0$ 2. Para $x = 100$: $z_2 = \frac{100-90}{10} = +1.0$
Buscamos $P(-1.0 < Z < 1.0)$. Esto es $P(Z < 1.0) - P(Z < -1.0) \approx 0.8413 - 0.1587 = 0.6827$ (68.27 %). Esto significa que aproximadamente el 68.27 % de las personas sanas tienen una glucemia entre 80 y 100 mg/dL. Antes los valores de z se buscaban en la tabla de la distribución normal estándar, pero con R es mucho más sencillo usando `pnorm()`.
Verificación con R:

Cuadro 2.7 Código R

```r
mu <- 90
sigma <- 10

# Probabilidad de P(80 < X < 100)
# Esto es P(X < 100) - P(X < 80)
prob <- pnorm(100, mean = mu, sd = sigma) - pnorm(80, mean = mu, sd = sigma)

cat("Porcentaje de personas entre 80 y 100 mg/dL:", round(prob * 100, 2), "%\n")
```

```
Porcentaje de personas entre 80 y 100 mg/dL: 68.27 %
```

2.3.4. Regla Empírica (68-95-99.7)

Definición: Regla 68-95-99.7

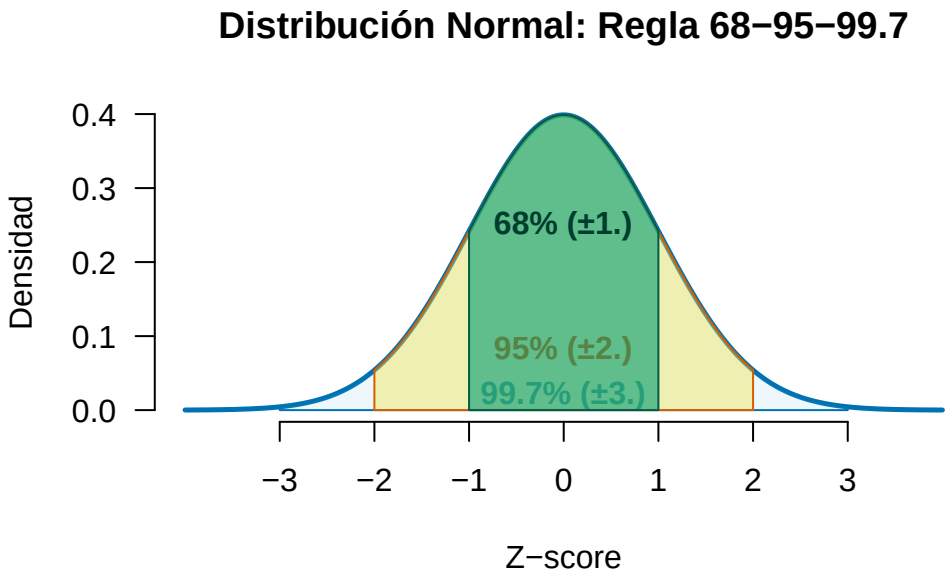
Cando una variable aleatoria X se transforma a puntuaciones estándar mediante: so z=-score, se dice que hay sido tipificada y se ha convertido en una variable aleatoria Z que sigue una distribución normal estándar N(0,1). La regla empírica (o regla 68-95-99.7) se interpreta directamente en términos de la distribución normal estándar N(0,1):

- Aproximadamente el **68 %** de los datos caen dentro de ± 1 desviación estándar ($\pm 1Z$)
- Aproximadamente el **95 %** de los datos caen dentro de ± 2 desviaciones estándar ($\pm 2Z$)
- Aproximadamente el **99.7 %** de los datos caen dentro de ± 3 desviaciones estándar ($\pm 3Z$)

En otras palabras, la tipificación "estandariza" cualquier normal a la misma escala, y esas proporciones se mantienen porque dependen únicamente de la forma de la curva normal estándar, no de la media o la desviación original.

Cuadro 2.8 Visualización de la Regla Empírica

```r
# Crear secuencia para curva normal
x <- seq(-4, 4, length.out = 200)
y <- dnorm(x)

# Graficar (paleta Okabe-Ito: azul oscuro #0072B2 + amarillo #F0E442 + vermellón #D55E00)
plot(x, y, type = "l", lwd = 2.5, col = "#0072B2",
     main = "Distribución Normal: Regla 68-95-99.7",
     xlab = "Z-score", ylab = "Densidad", axes = FALSE)
axis(1, at = -3:3)
axis(2, las = 1)

# Definir áreas y añadir etiquetas - colores diferenciables y de alto contraste
# 99.7%
x_99 <- seq(-3, 3, length.out = 150)
polygon(c(-3, x_99, 3), c(0, dnorm(x_99), 0),
        col = "#56B4E91A", border = "#0072B2")
text(0, 0.02, "99.7% (±3 )", col = "#0072B2", font = 2)

# 95%
x_95 <- seq(-2, 2, length.out = 100)
polygon(c(-2, x_95, 2), c(0, dnorm(x_95), 0),
        col = "#F0E44266", border = "#D55E00")
text(0, 0.08, "95% (±2 )", col = "#D55E00", font = 2)

# 68%
x_68 <- seq(-1, 1, length.out = 50)
polygon(c(-1, x_68, 1), c(0, dnorm(x_68), 0),
        col = "#009E7399", border = "#005C46")
text(0, 0.25, "68% (±1 )", col = "#003D2E", font = 2)
```



2.4. Diagramas de Caja (Boxplots)

Los diagramas de caja son una herramienta visual extremadamente útil en el AED. Permiten identificar rápidamente la mediana, los cuartiles, el rango y los valores atípicos.

2.4.1. Construcción de un Boxplot

Ingredientes del Boxplot

Para construir un diagrama de caja, necesitamos calcular:

1. **Mediana (Q2):** el valor central que divide los datos en dos mitades
2. **Primer cuartil (Q1):** el valor que deja el 25 % de los datos por debajo
3. **Tercer cuartil (Q3):** el valor que deja el 75 % de los datos por debajo
4. **Rango intercuartílico:** $\text{IQR} = Q3 - Q1$
5. **Límite inferior (lower whisker):** $Q1 - 1.5 \times \text{IQR}$
6. **Límite superior (upper whisker):** $Q3 + 1.5 \times \text{IQR}$
7. **Valores atípicos (outliers):** cualquier observación fuera de los límites

2.4.2. Interpretación Visual

Características Importantes

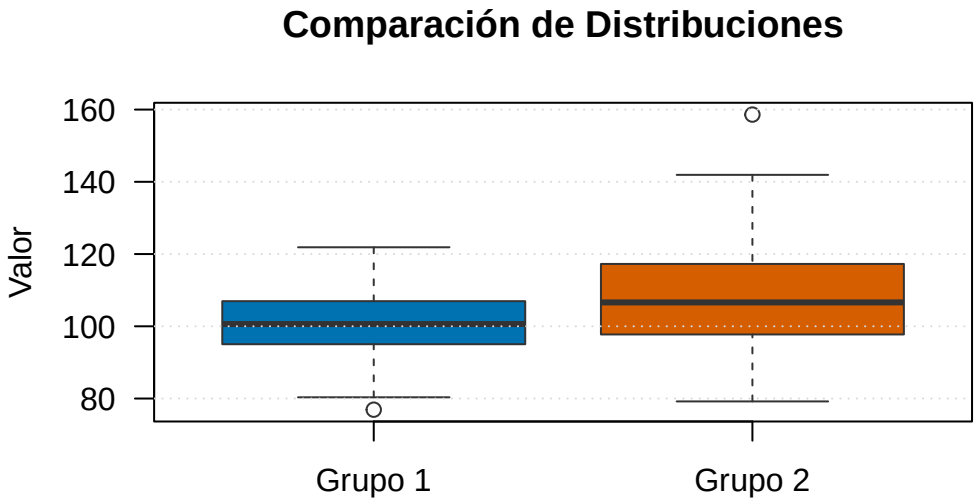
- La **caja** representa el 50 % central de los datos (entre Q1 y Q3)
- La **línea dentro de la caja** es la mediana (Q2)
- Los **bigotes** (whiskers) se extienden hasta los límites definidos
- Los **puntos individuales** fuera de los bigotes son valores atípicos
- Un boxplot **simétrico** sugiere una distribución aproximadamente normal

- Una mediana **no centrada** en la caja sugiere asimetría

Ejemplo 2.5: Boxplot Comparativo

Cuadro 2.9 Generar Boxplots

```r
# Generar datos simulados
set.seed(123)
grupo1 <- rnorm(100, mean = 100, sd = 10)
grupo2 <- rnorm(100, mean = 110, sd = 15)

# Boxplots comparativos con paleta Okabe-Ito (alto contraste, accesible)
boxplot(grupo1, grupo2,
        names = c("Grupo 1", "Grupo 2"),
        main  = "Comparación de Distribuciones",
        ylab  = "Valor",
        col   = c("#0072B2", "#D55E00"),
        border = "grey20",
        las   = 1,
        notch = FALSE)
grid(nx = NA, ny = NULL, col = "grey85", lty = "dotted")
```



2.5. Análisis Bivariante: Distribuciones Conjuntas

2.5.1. Distribución Conjunta

i Distribución Conjunta

La **distribución conjunta** (o distribución multivariante) describe simultáneamente dos o más variables. Para dos variables, hablamos de **distribución bivariante**.
Cuando observamos pares (x_i, y_i), podemos estudiar cómo ambas variables varían juntas, no solo separadamente.

2.5.2. Variables Discretas: Tablas de Frecuencia

i Tabla de Frecuencia (Contingencia)

Una **tabla de contingencia** o **tabla de frecuencia bidimensional** resume la relación entre dos variables discretas:

$X \backslash Y$	y_1	$\cdots$	y_j	$\cdots$	y_r	Total_X
x_1	h_{11}	$\cdots$	h_{1j}	$\cdots$	h_{1r}	$h_{1\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_i	h_{i1}	$\cdots$	h_{ij}	$\cdots$	h_{ir}	$h_{i\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_m	h_{m1}	$\cdots$	h_{mj}	$\cdots$	h_{mr}	$h_{m\bullet}$
Total_Y	$h_{\bullet 1}$	$\cdots$	$h_{\bullet j}$	$\cdots$	$h_{\bullet r}$	n

donde:

- h_{ij} es la **frecuencia absoluta** (número de casos con características x_i e y_j)
- $h_{i\bullet} = \sum_j h_{ij}$ es la distribución marginal de X
- $h_{\bullet j} = \sum_i h_{ij}$ es la distribución marginal de Y
- $f_{ij} = h_{ij}/n$ es la **frecuencia relativa**

2.5.3. Distribución Marginal

i Distribución Marginal

La **distribución marginal** de una variable es su distribución ignorando las otras variables.
En una tabla de contingencia:

- Distribución marginal de X: $h_{i\bullet} = \sum_j h_{ij}$
- Distribución marginal de Y: $h_{\bullet j} = \sum_i h_{ij}$

Las distribuciones marginales aparecen en los totales de filas y columnas.

2.5.4. Distribución Condicional

Distribución Condicional

La **distribución condicional** de una variable es su distribución cuando la otra variable toma un valor específico:

Distribución de X dado $Y = y_j$:

$$f(x_i|y_j) = \frac{h_{ij}}{h_{\bullet j}}$$

Distribución de Y dado $X = x_i$:

$$f(y_j|x_i) = \frac{h_{ij}}{h_{i\bullet}}$$

Ejemplo 2.6: Distribucines marginales y condicionales

Ocupación	Act. Rara	Act. Ocasional	Act. Regular
Obrero	240	120	70
No-manual	160	90	90
Oficina	30	30	30
Agricultor	37	7	6
Otros	40	32	18

Cuadro 2.10 Código R para porcentajes

```
options(scipen = 999)
# Crear tabla de contingencia
tabla <- rbind(
  c(240, 120, 70),
  c(160, 90, 90),
  c(30, 30, 30),
  c(37, 7, 6),
  c(40, 32, 18)
)

colnames(tabla) <- c("Rara", "Ocasional", "Regular")
rownames(tabla) <- c("Obrero", "No-manual", "Oficina", "Agricultor", "Otros")

# Mostrar tabla marginal y porcentajes
print(tabla)
```

```
           Rara Ocasional Regular
Obrero      240       120      70
No-manual   160        90      90
Oficina      30        30      30
Agricultor   37         7       6
Otros        40        32      18
```

Cuadro 2.11 Código R para porcentajes

```
print("Distribución marginal de ocupación:")
```

```
[1] "Distribución marginal de ocupación:"
```

Cuadro 2.12 Código R para porcentajes

```
round(rowSums(tabla), 3)
```

```
    Obrero  No-manual    Oficina Agricultor      Otros
       430        340         90         50         90
```

Cuadro 2.13 Código R para porcentajes

```
print("Distribución marginal de actividad:")
```

```
[1] "Distribución marginal de actividad:"
```

Cuadro 2.14 Código R para porcentajes

```
round(colSums(tabla), 3)
```

```
     Rara Ocasional   Regular
      507       279       214
```

Cuadro 2.15 Código R para porcentajes

```
# Mostrar tabla y calcular porcentajes condicionales
print(tabla)
```

```
           Rara Ocasional Regular
Obrero      240       120      70
No-manual   160        90      90
Oficina      30        30      30
Agricultor   37         7       6
Otros        40        32      18
```

Cuadro 2.16 Código R para porcentajes

```
print("Porcentajes por filas:")
```

```
[1] "Porcentajes por filas:"
```

Cuadro 2.17 Código R para porcentajes

```
round(prop.table(tabla, 1), 3)
```

```
            Rara Ocasional Regular
Obrero     0.558     0.279   0.163
No-manual  0.471     0.265   0.265
Oficina    0.333     0.333   0.333
Agricultor 0.740     0.140   0.120
Otros      0.444     0.356   0.200
```

Cuadro 2.18 Código R para porcentajes

```
print("Porcentajes por columnas:")
```

```
[1] "Porcentajes por columnas:"
```

Cuadro 2.19 Código R para porcentajes

```
round(prop.table(tabla, 2), 3)
```

```
            Rara Ocasional Regular
```

```
Obrero      0.473     0.430   0.327
No-manual   0.316     0.323   0.421
Oficina     0.059     0.108   0.140
Agricultor  0.073     0.025   0.028
Otros       0.079     0.115   0.084
```

Ejemplo 2.7: Distribuciones Condicionales y Probabilidad

Para ilustrar, consideremos la tabla de contingencia de un test de diagnóstico:

	Infectado	No infectado	Total
Test +	199	499	698
Test -	1	99301	99302
Total	200	99800	100000

Cuadro 2.20 Probabilidad condicional en R

```
options(scipen = 999)
# Crear tabla de contingencia 2x2: Test de VIH vs Infección
tabla_hiv <- rbind(
  c(199, 499),      # Test positivo
  c(1, 99301)       # Test negativo
)
colnames(tabla_hiv) <- c("Infectado", "No infectado")
rownames(tabla_hiv) <- c("Test +", "Test -")

# 1. Porcentajes por FILA (P(Infección | Test) - Valor Predictivo)
print("P(Infección | Test) - Porcentajes por FILA:")
```

```
[1] "P(Infección | Test) - Porcentajes por FILA:"
```

Cuadro 2.21 Probabilidad condicional en R

```
prop.table(tabla_hiv, 1)
```

```
           Infectado No infectado
Test + 0.28510028653    0.7148997
Test - 0.00001007029    0.9999899
```

Cuadro 2.22 Probabilidad condicional en R

```
# 2. Porcentajes por COLUMNA (Sensibilidad/Especificidad: P(Test | Infección))
print("P(Test | Infección) - Porcentajes por COLUMNA:")
```

```
[1] "P(Test | Infección) - Porcentajes por COLUMNA:"
```

Cuadro 2.23 Probabilidad condicional en R

```
prop.table(tabla_hiv, 2)
```

```
       Infectado No infectado
Test +     0.995        0.005
Test -     0.005        0.995
```

Interpretación Clínica: - **Por filas** (P(Infección | Test)): **Valor Predictivo** — responde a la pregunta clínica crucial: "Dado que el paciente tiene este resultado de test, ¿cuál es la probabilidad real de que esté infectado?". - **Por columnas** (P(Test | Infección)): **Sensibilidad y Especificidad** — qué tan bueno es el test para identificar la presencia o ausencia de enfermedad.

Métodos Avanzados

Para ampliar los contenidos sobre los test diagnósticos y otras técnicas estadísticas avanzadas, visita: **→ Bioestadística Avanzada — M.A. Luque Fernández**

Ejemplo 2.8: Covarianza y Correlación — Peso e IMC en pacientes diabéticos

Estudiamos la asociación lineal entre el **peso corporal** (kg) y el **índice de masa corporal** (IMC, kg/m²) en los 94 pacientes diabéticos del dataset `osteo` (Facultad de Medicina, UGR).

Cuadro 2.24 Covarianza y correlación entre peso e IMC en osteo

```
library(BioEstatR)
data(osteo)

# Eliminar valores perdidos en peso e IMC
dat <- na.omit(osteo[, c("peso", "imc", "sexo")])

# 1) Estadísticos descriptivos
cov_pi <- cov(dat$peso, dat$imc)
cor_pi <- cor(dat$peso, dat$imc)
test_cor <- cor.test(dat$peso, dat$imc)

cat(sprintf("Covarianza(peso, IMC)  =%.2f kg·(kg/m²)\n", cov_pi))
```

```
Covarianza(peso, IMC)  = 35.48 kg·(kg/m²)
```

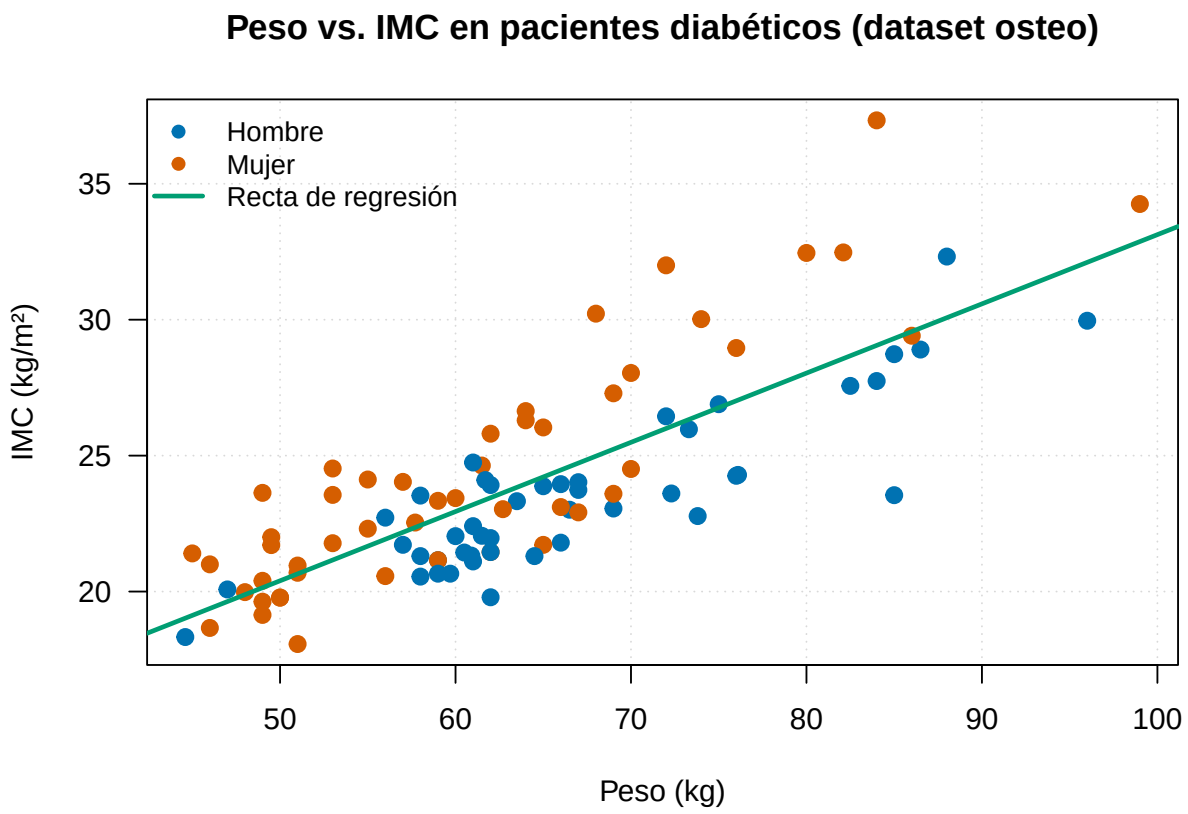
Cuadro 2.25 Covarianza y correlación entre peso e IMC en osteo

```
cat(sprintf("Correlación de Pearson r =%.3f  (IC 95%%: %.3f - %.3f, p =%.3g)\n",
            cor_pi, test_cor$conf.int[1], test_cor$conf.int[2], test_cor$p.value))
```

```
Correlación de Pearson r = 0.802  (IC 95%: 0.716 - 0.864, p = 2.68e-22)
```

Cuadro 2.26 Covarianza y correlación entre peso e IMC en osteo

```
# 2) Gráfico de dispersión coloreado por sexo, con línea de regresión
modelo <- lm(imc ~ peso, data = dat)

# Paleta Okabe-Ito (alta accesibilidad, segura para daltónicos)
col_pts <- ifelse(dat$sexo == "Hombre", "#0072B2", "#D55E00")

plot(dat$peso, dat$imc,
     main = "Peso vs. IMC en pacientes diabéticos (dataset osteo)",
     xlab = "Peso (kg)",
     ylab = "IMC (kg/m²)",
     pch  = 19,
     col  = col_pts,
     cex  = 1.2,
     las  = 1,
     panel.first = grid(col = "grey85", lty = "dotted"))

abline(modelo, col = "#009E73", lwd = 2.5)

legend("topleft",
       legend = c("Hombre", "Mujer", "Recta de regresión"),
       col    = c("#0072B2", "#D55E00", "#009E73"),
       pch    = c(19, 19, NA),
       lwd    = c(NA, NA, 2.5),
       bty    = "n",
       cex    = 0.9)
```



Interpretación:

- La **covarianza** muestral es $\widehat{\mathrm{Cov}}(\text{peso}, \text{IMC}) \approx 35.48$ kg $\cdot$ (kg/m^2) — positiva, lo que indica que peso e IMC tienden a crecer simultáneamente. Su magnitud depende de las unidades y no es directamente comparable entre estudios.
- La **correlación de Pearson** $r \approx 0.80$ (IC 95 %: 0.72 – 0.86; $p < 0.001$) cuantifica la asociación de forma adimensional: existe una asociación lineal positiva fuerte. Esto es **esperable por construcción** (IMC = peso/talla^2), pero la dispersión visible en el gráfico refleja la heterogeneidad debida a la talla.
- Los puntos en **azul oscuro** (`#0072B2`, hombres) y **vermellón** (`#D55E00`, mujeres) permiten apreciar visualmente que las mujeres se sitúan mayoritariamente en rangos de peso menores con IMC ligeramente más variable, mientras que los hombres se concentran en pesos superiores.
- La línea de regresión en **verde** (`#009E73`) muestra la mejor aproximación lineal de mínimos cuadrados.

  Paleta Okabe-Ito: accesibilidad cromática

  Los colores `#0072B2` (azul), `#D55E00` (vermellón) y `#009E73` (verde azulado) pertenecen a la paleta **Okabe-Ito**, diseñada específicamente para ser distinguible por personas con deuteranopía, protanopía y tritanopía (las formas más comunes de daltonismo). Se recomienda su uso sistemático en gráficos científicos en lugar de las combinaciones tradicionales rojo/verde o rojo/azul (Okabe & Ito, 2008).

2.5.5. Covarianza y Correlacion: Advertencia Visual

Limitaciones del coeficiente de correlación

Para comprender mejor por qué la correlación debe ir acompañada de un gráfico, observemos la siguiente figura:

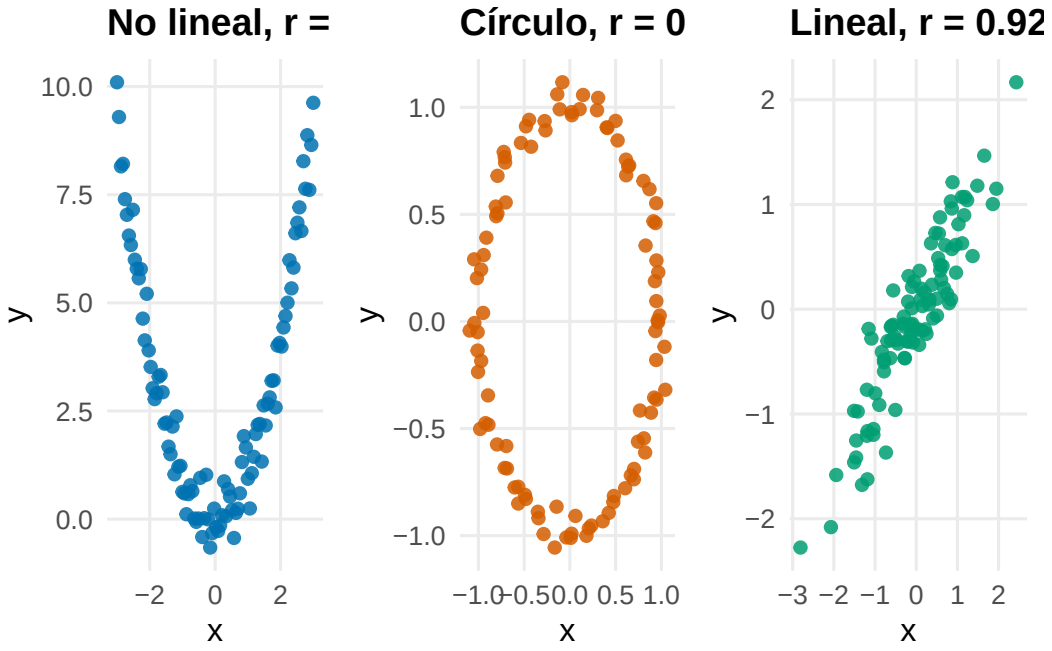


Figura 2.1: Limitaciones de la correlación de Pearson

En los dos primeros casos, existe una dependencia clara entre las variables, pero el coeficiente de correlación r es cercano a 0 porque no es una relación lineal.

La Paradoja de Simpson: Correlaciones Engañosas

La **Paradoja de Simpson** es un fenómeno donde una tendencia observada en varios grupos desaparece o se invierte al combinarlos.

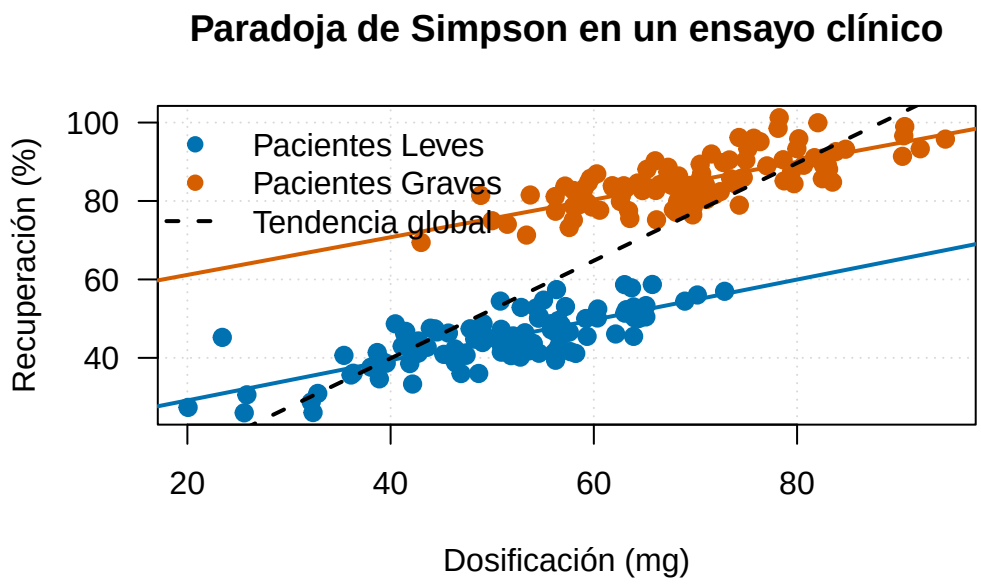
Relación con el ejemplo clínico:

En la figura a continuación (Ejemplo 2.11), vemos dos subgrupos (leves y graves). Si calculáramos la correlación total de todos los pacientes sin distinguir estos grupos, el resultado sería engañoso. La variable "gravedad" actúa aquí como un **factor de confusión**. Este ejemplo clínico ilustra perfectamente por qué la **"correlación no implica causalidad"** y subraya la importancia crítica de la estratificación y visualización de datos en medicina.

Ejemplo 2.9: Paradoja de Simpson en Ensayos Clínicos

Cuadro 2.27 Ilustración de la Paradoja de Simpson

```r
# Simulación: Efecto de un fármaco en dos subgrupos (ej. Pacientes leves vs graves)
set.seed(42)
n <- 100
# Subgrupo 1 (Leves)
x1 <- rnorm(n, mean = 50, sd = 10); y1 <- 20 + 0.5 * x1 + rnorm(n, sd = 5)
# Subgrupo 2 (Graves)
x2 <- rnorm(n, mean = 70, sd = 10); y2 <- 50 + 0.5 * x2 + rnorm(n, sd = 5)

# Gráfico (paleta Okabe-Ito: azul oscuro + vermellón, alto contraste)
plot(c(x1, x2), c(y1, y2),
     col  = rep(c("#0072B2", "#D55E00"), each = n),
     pch  = 19,
     cex  = 1.2,
     main = "Paradoja de Simpson en un ensayo clínico",
     xlab = "Dosificación (mg)", ylab = "Recuperación (%)",
     las  = 1,
     panel.first = grid(col = "grey85", lty = "dotted"))

# Líneas de regresión por subgrupo (refuerzan visualmente la paradoja)
abline(lm(y1 ~ x1), col = "#0072B2", lwd = 2)
abline(lm(y2 ~ x2), col = "#D55E00", lwd = 2)
# Línea de regresión global (efecto agregado, sentido opuesto)
abline(lm(c(y1, y2) ~ c(x1, x2)), col = "black", lwd = 2, lty = 2)

legend("topleft",
       legend = c("Pacientes Leves", "Pacientes Graves", "Tendencia global"),
       col    = c("#0072B2", "#D55E00", "black"),
       pch    = c(19, 19, NA),
       lwd    = c(NA, NA, 2),
       lty    = c(NA, NA, 2),
       bty    = "n")
```



2.5.6. Matriz de Correlaciones

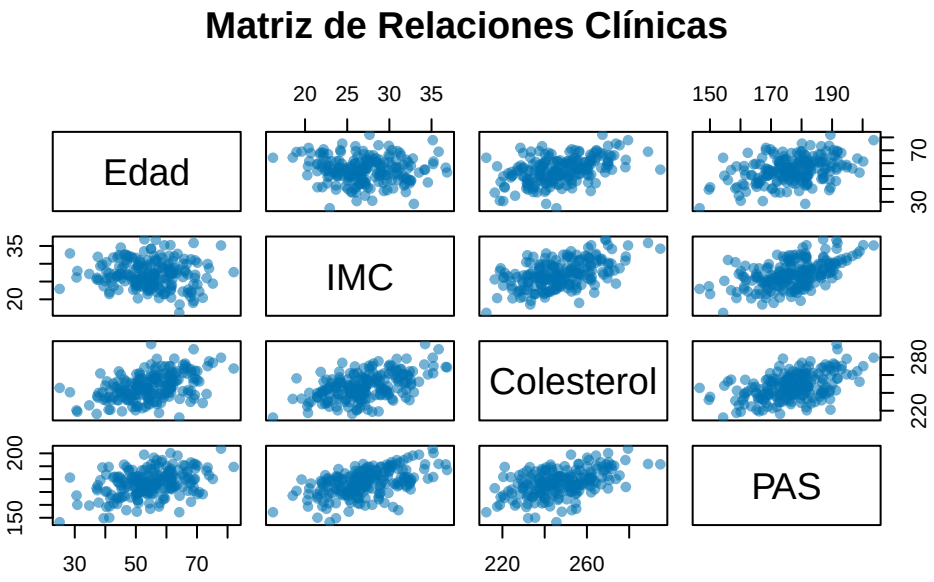
La matriz de correlaciones es una herramienta fundamental en investigación médica para evaluar la relación entre múltiples variables fisiológicas simultáneamente. Su utilidad principal radica en:

1. **Identificación de Predictores**: Detectar qué variables clínicas están más asociadas con un marcador de salud (ej. qué variables influyen más en la presión arterial).
2. **Detección de Colinealidad**: Identificar variables redundantes que aportan información similar, lo cual es crítico antes de construir modelos de regresión.
3. **Análisis de Clusters**: Agrupar variables que se comportan de manera similar en pacientes (ej. marcadores inflamatorios).

Ejemplo 2.10: Matriz de Correlaciones (Parámetros Clínicos)

Cuadro 2.28 Matriz de correlaciones clínicas

```
# Simulación de variables clínicas: Edad, Colesterol, Tensión Arterial (PAS), IMC
set.seed(42)
n <- 200
edad <- rnorm(n, 55, 10)
imc <- rnorm(n, 27, 4)
colesterol <- 150 + 0.8 * edad + 2 * imc + rnorm(n, 0, 10)
pas <- 110 + 0.5 * edad + 1.5 * imc + rnorm(n, 0, 8)

data_clinica <- data.frame(Edad = edad, IMC = imc, Colesterol = colesterol, PAS = pas)

# Visualización mediante matriz de dispersión (Pairs plot)
# Color Okabe-Ito (azul oscuro) con transparencia para que solapamientos sean visibles
pairs(data_clinica,
      main = "Matriz de Relaciones Clínicas",
      pch  = 19,
      cex  = 0.9,
      col  = adjustcolor("#0072B2", alpha.f = 0.55))
```



Cuadro 2.29 Matriz de correlaciones clínicas

```
# Correlaciones numéricas
cor(data_clinica)
```

```
             Edad        IMC Colesterol       PAS
```

```
Edad         1.00000000 -0.08036539  0.4484613 0.3906186
IMC         -0.08036539  1.00000000  0.5157946 0.5259363
Colesterol   0.44846130  0.51579465  1.0000000 0.4802219
PAS          0.39061863  0.52593630  0.4802219 1.0000000
```

2.6. Función de Distribución Empírica (ECDF)

En el Capítulo 1 estudiamos la Función de Distribución Empírica (ECDF) con datos agrupados en clases. Aquí presentamos cómo calcular y visualizar la ECDF con datos individuales usando la función `ecdf()` de R.

2.7. Ejemplo 2.11: ECDF de Alturas de Estudiantes

Se midieron las alturas (en cm) de 30 estudiantes de segundo año:

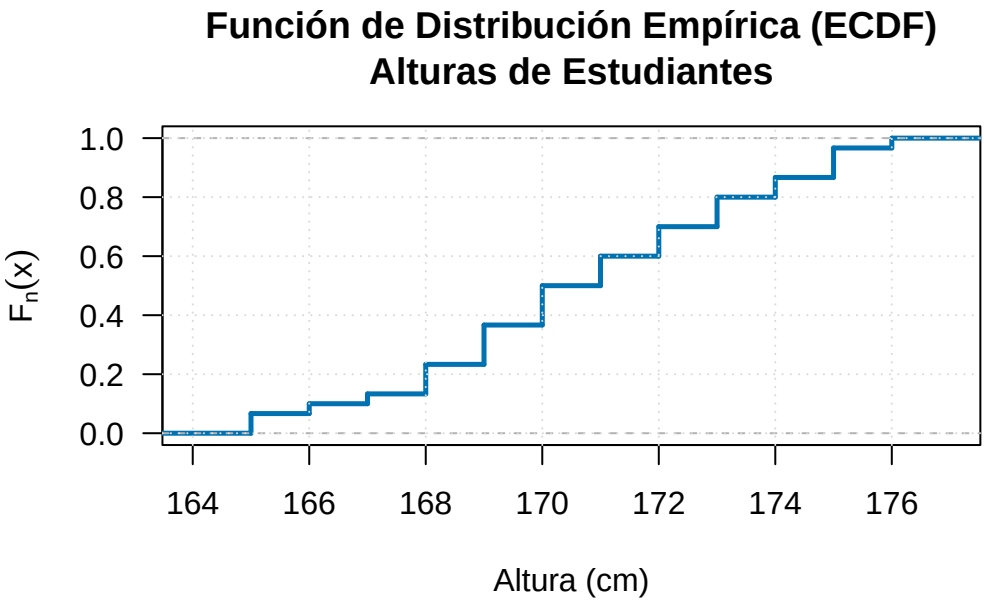
```
168, 172, 165, 175, 170, 169, 173, 166, 174, 171,
167, 176, 169, 172, 168, 170, 175, 169, 173, 171,
165, 174, 170, 172, 168, 175, 169, 171, 173, 170
```

La tabla siguiente muestra algunos valores y su función de distribución empírica acumulada:

Altura (cm)	$F_n(x)$
165	0.067
166	0.100
167	0.133
168	0.200
169	0.300
170	0.400
171	0.500
172	0.633
173	0.700
174	0.767
175	0.867
176	1.000

Cuadro 2.30 Código R para ECDF

```r
# Datos de alturas
alturas <- c(168, 172, 165, 175, 170, 169, 173, 166, 174, 171,
             167, 176, 169, 172, 168, 170, 175, 169, 173, 171,
             165, 174, 170, 172, 168, 175, 169, 171, 173, 170)

# Calcular ECDF
ecdf_alturas <- ecdf(alturas)

# Graficar la ECDF (paleta Okabe-Ito: azul oscuro de alto contraste)
plot(ecdf_alturas,
     main = "Función de Distribución Empírica (ECDF)\nAlturas de Estudiantes",
     xlab = "Altura (cm)",
     ylab = expression(F[n](x)),
     verticals = TRUE,
     do.points = FALSE,
     col  = "#0072B2",
     lwd  = 2.5,
     las  = 1)
grid(col = "grey85", lty = "dotted")
```



2.7.1. Interpretación

La ECDF $F_n(x)$ es una función escalonada que representa la proporción de datos menores o iguales a x:

- $F_n(165) \approx 0.067 \rightarrow$ El 6.7 % de estudiantes miden 165 cm
- $F_n(170) = 0.400 \rightarrow$ El 40 % de estudiantes miden 170 cm
- $F_n(175) \approx 0.867 \rightarrow$ El 86.7 % de estudiantes miden 175 cm

Propiedad fundamental: La ECDF es una **función no decreciente** que salta en cada valor observado.

2.8. Resumen

Este capítulo ha cubierto los conceptos esenciales para el análisis exploratorio bivariante de datos:

Concepto	Fórmula/Definición	Rango/Unidades
Varianza	$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$	$[0, \infty)$
Desviación Estándar	$s = \sqrt{s^2}$	Mismas unidades que x
Rango	$R = x_{\text{máx}} - x_{\text{mín}}$	Mismas unidades que x
Rango Intercuartílico	$\text{IQR} = Q3 - Q1$	Mismas unidades que x
Z-Score	$z = \frac{x-\bar{x}}{s}$	Adimensional
Covarianza	$s_{xy} = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$	Producto de unidades
Correlación Pearson	$r = \frac{s_{xy}}{s_x s_y}$	$[-1, 1]$

2.8.1. Puntos Clave

- Las **medidas de dispersión** (varianza, desviación estándar) cuantifican la variabilidad de los datos
- Las **transformaciones lineales** predicen cambios en las estadísticas de forma matemática
- Los **diagramas de caja** visualizan rápidamente la distribución y detectan outliers
- El **análisis bivariante** describe relaciones entre pares de variables
- La **correlación** mide la fuerza de relaciones lineales, pero **no implica causalidad**
- Siempre **visualiza los datos**: los resúmenes estadísticos pueden ocultar patrones subyacentes complejos
- La **ECDF** permite estimar cuantiles y visualizar la distribución acumulada sin asumir ninguna forma paramétrica

2.9. Ejercicios

Ejercicio 2.1: Estadísticos de Dispersión

Se registraron los tiempos de respuesta (en milisegundos) de 12 participantes en una tarea de tiempo de reacción: 250, 240, 300, 280, 270, 265, 290, 255, 310, 275, 260, 320

a) Calcula el rango y el rango intercuartílico (IQR).
b) Calcula la media y la desviación estándar.
c) ¿Qué observación es un valor atípico potencial según la regla de 1.5×IQR?
d) Dibuja un diagrama de caja.

Ejercicio 2.2: Transformaciones Lineales (Conversión de Unidades)

En un estudio de investigación, se registra la presión intraocular (PIO) de 50 pacientes en milímetros de mercurio (mmHg). La media registrada es $\bar{x} = 15$ mmHg con una desviación estándar de $s_x = 3$ mmHg. Para una publicación internacional, es necesario convertir estos valores a centímetros de agua (cmH O) utilizando la fórmula: $y = 1.36x$.

a) ¿Cuál será la media de la PIO en cmH O?
b) ¿Cuál será la desviación estándar en cmH O?
c) Si el IQR original es de 4 mmHg, ¿cuál será el IQR en cmH O?

Ejercicio 2.3: Z-Scores y Estandarización Clínica

En un estudio de cribado, se utilizan dos escalas distintas para medir el deterioro cognitivo. Un paciente obtiene las siguientes puntuaciones:

- Escala A: 85 puntos (media = 75, desviación estándar = 10)
- Escala B: 90 puntos (media = 80, desviación estándar = 15)

a) Calcula los z-scores para ambas escalas.
b) ¿En qué escala el deterioro del paciente es relativamente más severo?
c) Si se considera un z > 0.5 como "indicativo de riesgo", ¿en cuál(es) escala(s) el paciente supera este umbral?

Ejercicio 2.4: Tablas de Contingencia

Se encuestó a 300 personas sobre si fuman y si tienen problemas respiratorios. Los resultados fueron:

	Problemas	Sin problemas	Total
Fuman	70	40	110
No fuman	30	160	190
Total	100	200	300

a) Calcula la distribución marginal de fumadores.
b) Calcula la distribución condicional de problemas respiratorios **dado** que fuma.
c) Calcula la distribución condicional de problemas respiratorios **dado** que no fuma.
d) Interpreta las diferencias entre las distribuciones condicionales.

Ejercicio 2.5: Análisis de Datos Clínicos Simulados

Genera un dataset que represente la relación entre dos variables biomédicas:

Cuadro 2.31 Código R: Simulación Clínica

```r
set.seed(999)
edad <- rnorm(100, mean = 50, sd = 15)
presion_sistolica <- 100 + 0.6 * edad + rnorm(100, sd = 10)
```

Realiza las siguientes tareas: a) Calcula la media, desviación estándar y covarianza de la edad y la presión sistólica. b) Calcula el coeficiente de correlación de Pearson. c) Crea un gráfico de dispersión con línea de regresión. d) Crea boxplots para ambas variables. e) Estandariza ambas variables usando z-scores. f) Verifica que las variables estandarizadas tienen media 0 y desviación estándar 1.

Ejercicio 2.6: Distribuciones Conjuntas

Se miden presión sistólica (X) y diastólica (Y) en una muestra clínica. Se encuentra que $\text{Cov}(x, y) = 180$ (mmHg)^2, $s_x = 15$ mmHg, $s_y = 10$ mmHg.

a) Calcula la correlación entre X e Y.

b) Interpreta el valor obtenido.

c) ¿Son X e Y independientes?

2.10. Respuestas a los Ejercicios

Ejercicio 2.1: Estadísticos de Dispersión

a) Rango: 320 - 240 = 80 ms IQR: Q - Q = 292.5 - 258.75 = 33.75 ms (`quantile()` por defecto, type 7)

b) Media: 276.25 ms Desviación estándar: 24.69 ms Mediana: 272.5 ms

c) Outliers (> Q + 1.5×IQR = 292.5 + 50.625 = 343.125): Ninguno detectado Outliers (< Q - 1.5×IQR = 258.75 - 50.625 = 208.125): Ninguno

d) El boxplot muestra distribución simétrica alrededor de 275-280 ms sin outliers.

Ejercicio 2.2: Transformaciones Lineales (Conversión)

a) $\bar{y} = 1.36 \times 15 = 20.4$ cmH O

b) $s_y = 1.36 \times 3 = 4.08$ cmH O

c) $\text{IQR}_y = 1.36 \times 4 = 5.44$ cmH O

Ejercicio 2.3: Z-Scores y Estandarización Clínica

a) Escala A: $z = \frac{85-75}{10} = 1.0$ Escala B: $z = \frac{90-80}{15} = 0.67$

b) El deterioro es más severo en la Escala A, ya que el paciente se aleja más de la media poblacional (z=1.0 vs z=0.67).

c) El paciente supera el umbral en ambas escalas ($1.0 > 0.5$ y $0.67 > 0.5$).

Ejercicio 2.4: Tablas de Contingencia

a) Distribución marginal de fumadores: P(Fumar) = 110/300 = 0.367 P(No fumar) = 190/300 = 0.633

b) P(Problemas | Fuman) = 70/110 = 0.636

c) P(Problemas | No fuman) = 30/190 = 0.158

d) Los fumadores tienen 4 veces más probabilidad de problemas respiratorios (63.6 % vs 15.8 %)

Ejercicio 2.5: Análisis de Datos Clínicos Simulados

Para resolver este ejercicio en R: a) Usa `mean(edad)`, `sd(edad)`, `mean(presion_sistolica)`, `sd(presion_sistolica)` y `cov(edad, presion_sistolica)`. b) Usa `cor(edad, presion_sistolica)`. c) `plot(edad, presion_sistolica); abline(lm(presion_sistolica ~ edad))`. d) `boxplot(edad); boxplot(presion_sistolica)`. e) `z_edad <- scale(edad); z_presion <- scale(presion_sistolica)`. f) `mean(z_edad); sd(z_edad)` (el resultado será prácticamente 0 y 1).

Ejercicio 2.6: Distribuciones Conjuntas

a) Correlación $(r) = \frac{\text{Cov}(X,Y)}{s_x \cdot s_y} = \frac{180}{15 \times 10} = \frac{180}{150} = 1.2$. *Nota:* Este valor es **imposible** porque $|r| \leq 1$ siempre. La inconsistencia revela que los valores propuestos para Cov, s_x y s_y violan la **desigualdad de Cauchy–Schwarz** $|s_{xy}| \leq s_x \cdot s_y$ (aquí $|180| > 15 \times 10 = 150$). En la práctica, ante un resultado así habría que revisar los datos de partida.

b) Si los datos fuesen coherentes —por ejemplo, $\text{Cov}(X, Y) = 120$ con los mismos $s_x = 15$ y $s_y = 10$— se obtendría $r = 120/150 = 0.8$, lo cual indicaría una correlación lineal positiva fuerte entre la presión sistólica y diastólica en esta muestra.

c) No son independientes (Cov $\neq 0$), pero la relación lineal es despreciable. Independencia estadística requiere Cov = 0 y más condiciones.

Parte II

Parte II: Probabilidad y Variables Aleatorias

Capítulo 3

Semana 3 — Fundamentos de la Teoría de la Probabilidad

3.1. Introducción

La probabilidad es el fundamento del razonamiento estadístico. Esta semana cubriremos los conceptos básicos de la teoría de la probabilidad: experimentos aleatorios, espacios muestrales, eventos, y las herramientas matemáticas para cuantificar la incertidumbre. Aprenderemos los axiomas de Kolmogorov, el Teorema de Bayes, y cómo actualizar nuestras creencias con nueva información.

3.2. Experimentos Aleatorios y Eventos

3.2.1. Experimento Aleatorio

Definición: Experimento Aleatorio

Un **experimento aleatorio** es un proceso que:

- Puede repetirse bajo condiciones idénticas
- Tiene al menos dos resultados posibles diferentes
- Se realiza de una manera claramente especificada
- Tiene un resultado que no puede predecirse con certeza antes de realizarse

El **espacio muestral** Ω (o S) es el conjunto de todos los resultados posibles del experimento. Un **evento** A es un subconjunto del espacio muestral. Decimos que un evento "ocurre" si el resultado del experimento es un elemento de A.

Ejemplo 3.1: Aleatorización en Ensayo Clínico

Experimento: Aleatorización de un paciente a un grupo de tratamiento (T) o control (C) mediante el lanzamiento de una moneda.
Espacio muestral:

$$\Omega = \{T, C\}$$

Eventos posibles: - Evento A: "el paciente es asignado al grupo de tratamiento" $\rightarrow A = \{T\}$ - Evento B: "el paciente es asignado al grupo control" $\rightarrow B = \{C\}$

3.2.2. Evento Elemental y Partición

Definición: Evento Elemental

Un **evento elemental** es un evento que contiene exactamente un resultado del experimento. Por ejemplo, $\{2\}$ es un evento elemental en el experimento de lanzar un dado.

Definición: Partición Completa

Los eventos $A_1, A_2, \ldots, A_n$ forman una **partición completa** del espacio muestral Ω si:

1. $A_i \cap A_j = \emptyset$ para todo $i \neq j$ (son mutuamente disjuntos)
2. $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$ (cubren todo el espacio)
3. $\mathbb{P}(A_i) > 0$ para cada i (cada uno tiene probabilidad positiva)

3.3. Álgebra de Conjuntos

Los eventos se pueden combinar usando operaciones de conjuntos. A continuación presentamos las operaciones fundamentales:

Operaciones Básicas de Conjuntos

Conjunto vacío: $\emptyset$ es el conjunto sin elementos. En términos de eventos, $A = \emptyset$ significa que el evento es imposible.
Subconjunto: $A \subseteq B$ significa que cada elemento de A está también en B. En términos de eventos: "si A ocurre, entonces B ocurre".
Intersección: $A \cap B$ es el conjunto de elementos que están en ambos A y B. En términos de eventos: "tanto A como B ocurren".
Unión: $A \cup B$ es el conjunto de elementos que están en A o en B (o en ambos). En términos de eventos: "al menos uno de A o B ocurre".
Complemento: $\overline{A}$ (o A^c) es el conjunto de elementos en Ω que no están en A. En términos de eventos: "A no ocurre".

Diferencia: $A - B$ (o $A - B$) es el conjunto de elementos en A pero no en B. En términos de eventos: "A ocurre pero B no ocurre".

3.3.1. Relaciones y Operaciones de Eventos

La siguiente tabla resume las correspondencias entre descripciones de eventos y notación de conjuntos:

Descripción	Notación	Significado
A ocurre con certeza	$A = \Omega$	evento seguro
A es imposible	$A = \emptyset$	evento imposible
Si A ocurre, entonces B ocurre	$A \subseteq B$	A es subconjunto de B
A y B nunca ocurren juntos	$A \cap B = \emptyset$	eventos disjuntos
A y B son complementarios	$B = \overline{A}$	B ocurre si y solo si A no ocurre
Al menos uno de los A_i ocurre	$A = \bigcup_i A_i$	unión de eventos
Todos los A_i ocurren	$A = \bigcap_i A_i$	intersección de eventos

3.3.2. Diagramas de Venn

Los diagramas de Venn proporcionan representaciones visuales de operaciones entre conjuntos. A continuación se muestran las operaciones fundamentales:

3.3.3. Leyes de De Morgan

Las Leyes de De Morgan son reglas fundamentales que relacionan la unión, la intersección y el complemento de conjuntos. Son extremadamente útiles para simplificar expresiones probabilísticas complejas.

! Leyes de De Morgan

Para cualquier par de eventos A y B:

1. **Complemento de la unión:** El complemento de la unión de dos conjuntos es la intersección de sus complementos.

$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$

Significado: "Ni A ni B ocurren" es lo mismo que "A no ocurre Y B no ocurre".

2. **Complemento de la intersección:** El complemento de la intersección de dos conjuntos es la unión de sus complementos.

$$\overline{A \cap B} = \overline{A} \cup \overline{B}$$

Significado: "No ocurre que ambos A y B ocurran" es lo mismo que "A no ocurre O B

no ocurre".

Ejemplo 3.2: Visualización de De Morgan

Representación de $\overline{A \cup B} = \overline{A} \cap \overline{B}$:

Ejemplo 3.3: Aplicación Práctica con 3 Conjuntos

En estudios epidemiológicos es común analizar la comorbilidad de tres condiciones (A, B, C). Un diagrama de Venn de 3 conjuntos permite visualizar todas las posibles intersecciones:
Problema: En una muestra de 100 pacientes:

- 40 tienen Hipertensión (A)
- 30 tienen Diabetes (B)
- 20 tienen Obesidad (C)
- 10 tienen A y B
- 8 tienen A y C
- 5 tienen B y C
- 3 tienen las tres condiciones

¿Cuántos pacientes tienen **al menos una** condición? Usando el principio de Inclusión-Exclusión: $|A \cup B \cup C| = 40 + 30 + 20 - (10 + 8 + 5) + 3 = 90 - 23 + 3 = 70$ pacientes.

3.4. Definiciones de Probabilidad

Existen varias formas de formalizar el concepto de probabilidad, cada una reflejando una filosofia diferente sobre cómo interpretamos la "probabilidad".

3.4.1. Probabilidad Clásica (Laplace)

Definición: Probabilidad Clásica de Laplace

La probabilidad clásica se define como:

$$\mathbb{P}(A) = \frac{\text{Número de resultados favorables a } A}{\text{Número total de resultados posibles}} = \frac{|A|}{|\Omega|}$$

Supuestos: - Hay al menos dos resultados elementales posibles - Exactamente uno de los resultados posibles ocurre en cada experimento - El número de resultados elementales es finito - Cada resultado elemental ocurre con la misma probabilidad (equiprobabilidad)

Ejemplo 3.4: Moneda justa en Aleatorización

Experimento: Lanzar una moneda para decidir el grupo de tratamiento de un paciente.
Evento A: "el paciente es asignado al grupo de tratamiento (T)"
Resultados favorables a A: $\{T\} \to 1$ resultado
Resultados totales: $\{T, C\} \to 2$ resultados

$$\mathbb{P}(A) = \frac{1}{2} = 0.5$$

3.4.2. Probabilidad Frecuentista (von Mises)

Definición: Probabilidad Frecuentista

Sea n el número de repeticiones de un experimento y $f_n(A)$ el número de veces que ocurre el evento A. La **frecuencia relativa** es:

$$\text{Frecuencia relativa} = \frac{f_n(A)}{n}$$

La **probabilidad frecuentista** de A se define como el límite de la frecuencia relativa cuando $n \to \infty$:

$$\mathbb{P}(A) = \lim_{n\to\infty} \frac{f_n(A)}{n}$$

Esta definición requiere:

- Una secuencia de muestras independientes
- Bajo condiciones idénticas
- Repetibles arbitrariamente

Ejemplo 3.5: Simulación de aleatorización

Simulamos asignar n pacientes a un grupo de tratamiento:

Cuadro 3.1 Código R

```r
set.seed(123)

n <- 10
asignacion <- rbinom(n, size = 1, prob = 0.5)
prop_10 <- sum(asignacion) / n
cat("n =", n, ": proporción asignada a tratamiento =", prop_10, "\n")
```

```
n = 10 : proporción asignada a tratamiento = 0.6
```

Cuadro 3.2 Código R

```
n <- 100
asignacion <- rbinom(n, size = 1, prob = 0.5)
prop_100 <- sum(asignacion) / n
cat("n =", n, ": proporción asignada a tratamiento =", prop_100, "\n")
```

```
n = 100 : proporción asignada a tratamiento = 0.46
```

Cuadro 3.3 Código R

```
n <- 1000
asignacion <- rbinom(n, size = 1, prob = 0.5)
prop_1000 <- sum(asignacion) / n
cat("n =", n, ": proporción asignada a tratamiento =", prop_1000, "\n")
```

```
n = 1000 : proporción asignada a tratamiento = 0.497
```

Interpretación

La simulación ilustra la **Ley de los Grandes Números**: conforme incrementamos el tamaño muestral de 10 a 1,000 pacientes, la proporción observada de asignados al tratamiento converge hacia la probabilidad teórica de 0.5. Con apenas 10 pacientes observamos fluctuaciones notables (prop_10 = 0.60), con 100 ya se acerca al valor esperado (prop_100 = 0.46), y con 1,000 pacientes la proporción se estabiliza muy cerca de 0.5 (prop_1000 = 0.497). Esto demuestra empíricamente por qué los estudios clínicos requieren tamaños muestrales adecuados: garantizan que la frecuencia relativa refleje la probabilidad verdadera subyacente en la aleatorización.

A medida que n aumenta, la proporción de pacientes en el grupo de tratamiento se acerca a 0.5.

3.4.3. Probabilidad Axiomática (Kolmogorov)

Definición: Axiomas de Kolmogorov

Una función $\mathbb{P} : \mathcal{F} \to [0, 1]$ que asigna un número a cada evento en Ω es una **medida de probabilidad** si satisface los siguientes axiomas:

Axioma 1 (No negatividad):

$$\mathbb{P}(A) \geq 0 \text{ para todo evento } A$$

Axioma 2 (Normalización):

$$\mathbb{P}(\Omega) = 1$$

Axioma 3 (Aditividad contable): Si $A_1, A_2, A_3, ...$ son eventos mutuamente disjuntos (es decir, $A_i \cap A_j = \emptyset$ para $i \neq j$), entonces:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Componentes de un espacio de probabilidad: 1. Un espacio muestral Ω (el conjunto de todos los resultados posibles) 2. Una σ-álgebra $\mathcal{F}$ (una colección de eventos) 3. Una medida de probabilidad $\mathbb{P}$ (que satisface los axiomas anteriores)

La definición axiomática de Kolmogorov es la más general y es la base de toda la teoría moderna de la probabilidad.

3.5. Propiedades de la Probabilidad

De los axiomas de Kolmogorov se derivan varias propiedades útiles:

⚠ Teorema: Propiedades Fundamentales de la Probabilidad

Sea $A, B, A_1, A_2, ...$ eventos en un espacio de probabilidad. Entonces:
Propiedad 1: Probabilidad del complemento

$$\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$$

Prueba: Note que $A \cup \overline{A} = \Omega$ y $A \cap \overline{A} = \emptyset$. Por el Axioma 3:

$$\mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(\overline{A}) = 1$$

Propiedad 2: Probabilidad del conjunto vacío

$$\mathbb{P}(\emptyset) = 0$$

Prueba: $\emptyset = \overline{\Omega}$, así que $\mathbb{P}(\emptyset) = 1 - \mathbb{P}(\Omega) = 1 - 1 = 0$.
Propiedad 3: Monotonía
Si $A \subseteq B$, entonces $\mathbb{P}(A) \leq \mathbb{P}(B)$
Prueba: Si $A \subseteq B$, entonces $B = A \cup (B - A)$ donde A y $B - A$ son disjuntos. Por el Axioma 3:

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B - A) \geq \mathbb{P}(A)$$

Propiedad 4: Acotamiento

$$0 \leq \mathbb{P}(A) \leq 1$$

Se sigue de los Axiomas 1 y 2.

Propiedad 5: Regla de la suma (Inclusión-Exclusión)

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Prueba: Observe que $A \cup B = A \cup (B - A)$ donde estos son disjuntos, y $B = (A \cap B) \cup (B - A)$.

Propiedad 6: Aditividad para eventos disjuntos

Si $A \cap B = \emptyset$, entonces:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

Ejemplo 3.6: Aplicación de propiedades

Problema: En una consulta médica, el 30 % de los pacientes tienen hipertensión (H). Entre los pacientes hipertensos, el 80 % también presenta obesidad. ¿Cuál es la probabilidad de que un paciente NO tenga hipertensión?

Solución: Sea H = "tener hipertensión". Entonces $\mathbb{P}(H) = 0.30$.

Por la Propiedad 1:

$$\mathbb{P}(\text{sin hipertensión}) = \mathbb{P}(\overline{H}) = 1 - 0.30 = 0.70$$

3.6. Teorema de la Suma (Inclusión-Exclusión)

Teorema: Regla de la Suma (Inclusión-Exclusión)

Para dos eventos arbitrarios A y B:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Para tres eventos A, B, y C:

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$$

$$-\mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C)$$

$$+\mathbb{P}(A \cap B \cap C)$$

Ejemplo 3.7: Comorbilidad de pacientes

Problema: En una muestra de 52 pacientes, ¿cuál es la probabilidad de que un paciente tenga Hipertensión (A) O Diabetes (B)?
Definición de eventos: - A = "paciente con Hipertensión" $\rightarrow$ 4 pacientes - B = "paciente con Diabetes" $\rightarrow$ 13 pacientes - $A \cap B$ = "paciente con ambas enfermedades" $\rightarrow$ 1 paciente
Cálculo de probabilidades individuales:

$$\mathbb{P}(A) = \frac{4}{52}, \quad \mathbb{P}(B) = \frac{13}{52}, \quad \mathbb{P}(A \cap B) = \frac{1}{52}$$

Aplicación del Teorema de la Suma:

$$\mathbb{P}(A \cup B) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} \approx 0.308$$

3.7. Probabilidad Condicional

La probabilidad condicional permite actualizar nuestras creencias sobre un evento cuando tenemos información adicional.

Definición: Probabilidad Condicional

Dados dos eventos A y B con $\mathbb{P}(B) > 0$, la **probabilidad condicional** de A dado que B ha ocurrido es:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Interpretación: Es la probabilidad de que ocurra A cuando ya sabemos que B ocurrió. Formalmente, es la fracción de casos en que B ocurre que también incluyen la ocurrencia de A.

Ejemplo 3.8: Probabilidad condicional en aleatorización

Experimento: Asignación de dos pacientes consecutivos a grupos de tratamiento.
Eventos: - A = "el segundo paciente es asignado al grupo de tratamiento (T_2)" - B = "el primer paciente fue asignado al grupo de tratamiento (T_1)"
Cálculo: Si la asignación es aleatoria e independiente, $\mathbb{P}(T_1) = 0.5$, $\mathbb{P}(T_2) = 0.5$. La probabilidad condicional $\mathbb{P}(T_2|T_1) = \mathbb{P}(T_2) = 0.5$ (por independencia).

3.7.1. Teorema de la Multiplicación

⚠ Teorema: Regla de la Multiplicación

Para dos eventos A y B:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B|A) = \mathbb{P}(B) \cdot \mathbb{P}(A|B)$$

Para tres eventos A_1, A_2, A_3:

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_3|A_1 \cap A_2)$$

Para n eventos:

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_3|A_1 \cap A_2) \cdots \mathbb{P}(A_n|A_1 \cap \ldots \cap A_{n-1})$$

3.8. Independencia de Eventos

ℹ Definición: Independencia de Eventos

Dos eventos A y B son **independientes** si la ocurrencia de uno no afecta la probabilidad del otro:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

Equivalentemente:

$$\mathbb{P}(A|B) = \mathbb{P}(A) \quad (\text{si } \mathbb{P}(B) > 0)$$

$$\mathbb{P}(B|A) = \mathbb{P}(B) \quad (\text{si } \mathbb{P}(A) > 0)$$

Nota importante: Independencia es diferente de “disjuntos” (mutuamente excluyentes). Dos eventos disjuntos tienen $\mathbb{P}(A \cap B) = 0$, lo cual es diferente de independencia.

Ejemplo 3.9: Aleatorización independiente

Experimento: Asignar dos pacientes consecutivos al grupo de tratamiento (T) o control (C).
Eventos: - A = “primer paciente asignado al grupo T” - B = “segundo paciente asignado al grupo T”
Análisis:

$$\mathbb{P}(A) = 0.5, \quad \mathbb{P}(B) = 0.5$$

$$\mathbb{P}(A \cap B) = 0.25$$

Verificación de independencia:

$$\mathbb{P}(A) \cdot \mathbb{P}(B) = 0.5 \times 0.5 = 0.25 = \mathbb{P}(A \cap B)$$

Por lo tanto, A y B son independientes.

3.8.1. Independencia de Múltiples Eventos

Definición: Independencia Mutua

Los eventos $A_1, A_2, \ldots, A_n$ son **mutuamente independientes** si para cualquier subconjunto $\{i_1, i_2, \ldots, i_k\}$ de índices:

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k})$$

Esta condición debe cumplirse para todo subconjunto posible, no solo para pares de eventos.

3.9. Ley de la Probabilidad Total

Teorema: Ley de la Probabilidad Total

Sea $\{A_1, A_2, \ldots, A_n\}$ una partición completa del espacio muestral Ω. Entonces, para cualquier evento B:

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i)$$

Interpretación: La probabilidad de B se puede calcular como una suma ponderada de las probabilidades condicionales de B bajo cada posibilidad en la partición, ponderadas por las probabilidades de cada parte.

Ejemplo 3.10: Diagnóstico de una enfermedad (Ley de Probabilidad Total)

Escenario: Una enfermedad se distribuye en una población según tres grupos de riesgo:

- Grupo 1 (Bajo riesgo): 60 % de la población, tasa de enfermedad = 2 %
- Grupo 2 (Riesgo moderado): 30 % de la población, tasa de enfermedad = 10 %
- Grupo 3 (Alto riesgo): 10 % de la población, tasa de enfermedad = 25 %

Probabilidades a priori: - $\mathbb{P}(A_1) = 0.60$ - $\mathbb{P}(A_2) = 0.30$ - $\mathbb{P}(A_3) = 0.10$

Probabilidades condicionales de enfermedad (B): - $\mathbb{P}(B|A_1) = 0.02$ - $\mathbb{P}(B|A_2) = 0.10$ - $\mathbb{P}(B|A_3) = 0.25$

Aplicación de la Ley de la Probabilidad Total:

$$\mathbb{P}(B) = \mathbb{P}(B|A_1) \cdot \mathbb{P}(A_1) + \mathbb{P}(B|A_2) \cdot \mathbb{P}(A_2) + \mathbb{P}(B|A_3) \cdot \mathbb{P}(A_3)$$

$$= 0.02 \times 0.60 + 0.10 \times 0.30 + 0.25 \times 0.10$$

$$= 0.012 + 0.03 + 0.025 = 0.067$$

Interpretación: La probabilidad global de padecer la enfermedad en esta población es del 6.7 %.

3.10. Teorema de Bayes

El Teorema de Bayes es uno de los resultados más importantes de la teoría de la probabilidad y es fundamental en estadística moderna y aprendizaje automático.

3.10.1. Forma Simple del Teorema de Bayes

Teorema: Teorema de Bayes (Forma Simple)

Para dos eventos A y B con $\mathbb{P}(B) > 0$:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}$$

Terminología: - $\mathbb{P}(A|B)$ = probabilidad **a posteriori** (después de observar B)
- $\mathbb{P}(A)$ = probabilidad **a priori** (antes de observar B)
- $\mathbb{P}(B|A)$ = **verosimilitud** (probabilidad de la evidencia bajo el supuesto)
- $\mathbb{P}(B)$ = probabilidad **marginal** de la evidencia

3.10.2. Forma General con Partición

Teorema: Teorema de Bayes (Forma General)

Sea $\{A_1, A_2, \ldots, A_n\}$ una partición completa del espacio muestral. Para un evento B con $\mathbb{P}(B) > 0$:

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j) \cdot \mathbb{P}(A_j)}{\sum_{i=1}^{n} \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i)}$$

para cada $j = 1, 2, \ldots, n$.

3.10.3. Ejemplo: Prueba Diagnóstica

Ejemplo 3.11: Prueba diagnóstica médica

Escenario: Un test de diagnóstico para una enfermedad con:

- Prevalencia de la enfermedad: $\mathbb{P}(E) = 0.05$ (5 % de la población)
- Sensibilidad: $\mathbb{P}(+|E) = 0.95$ (95 % de enfermos dan positivo)
- Especificidad: $\mathbb{P}(-|\overline{E}) = 0.90$ (90 % de sanos dan negativo)

Pregunta: Si una persona da positivo, ¿cuál es la probabilidad de que realmente este enferma?

Solución:

Primero, calculamos $\mathbb{P}(+) = \mathbb{P}(+|E) \cdot \mathbb{P}(E) + \mathbb{P}(+|\overline{E}) \cdot \mathbb{P}(\overline{E})$
donde $\mathbb{P}(+|\overline{E}) = 1 - \mathbb{P}(-|\overline{E}) = 1 - 0.90 = 0.10$

$$\mathbb{P}(+) = 0.95 \times 0.05 + 0.10 \times 0.95 = 0.0475 + 0.095 = 0.1425$$

Ahora aplicamos el Teorema de Bayes:

$$\mathbb{P}(E|+) = \frac{\mathbb{P}(+|E) \cdot \mathbb{P}(E)}{\mathbb{P}(+)} = \frac{0.95 \times 0.05}{0.1425} = \frac{0.0475}{0.1425} \approx 0.333$$

Interpretación: Aunque el test tiene 95 % de sensibilidad, si una persona da positivo, la probabilidad de que realmente esté enferma es solo de aproximadamente 33 %. Esto sucede porque la enfermedad es rara (baja prevalencia) y hay muchos falsos positivos.

Cuadro 3.4 Código R

```r
prev <- 0.05
sens <- 0.95
spec <- 0.90

falso_pos_rate <- 1 - spec
p_positivo <- sens * prev + falso_pos_rate * (1 - prev)
vpp <- (sens * prev) / p_positivo

cat("Prevalencia:", prev, "\n")
```

```
Prevalencia: 0.05
```

Cuadro 3.5 Código R

```r
cat("Sensibilidad:", sens, "\n")
```

```
Sensibilidad: 0.95
```

Cuadro 3.6 Código R

```r
cat("Especificidad:", spec, "\n")
```

```
Especificidad: 0.9
```

Cuadro 3.7 Código R

```
cat("P(+):", p_positivo, "\n")
```

```
P(+): 0.1425
```

Cuadro 3.8 Código R

```
cat("Valor Predictivo Positivo (VPP):", round(vpp, 3), "\n")
```

```
Valor Predictivo Positivo (VPP): 0.333
```

Interpretación

Este cálculo implementa el **Teorema de Bayes** en contexto diagnóstico: aunque la prueba posee 95 % de sensibilidad (detecta enfermedad cuando existe), el Valor Predictivo Positivo (VPP 0.333) es sustancialmente menor debido a la baja prevalencia (5 %). Con una enfermedad rara, la mayoría de resultados positivos son falsos positivos. En la consulta clínica, este resultado orienta decisiones: ante test positivo con baja prevalencia, es prudente confirmar con prueba adicional antes de iniciar tratamiento, ilustrando cómo la probabilidad a priori (prevalencia) moldea la interpretación clínica de la evidencia.

:::

3.10.4. Ejemplo: Clasificación de pacientse

Ejemplo 3.12: Clasificación de pacientes (Bayes)

Escenario: Un médico quiere clasificar a un paciente como de alto riesgo o bajo riesgo de sufrir una patologîa cardîaca basándose en tres factores de riesgo (ej. hipertensión W_1, tabaquismo W_2, obesidad W_3).
Información disponible: - $\mathbb{P}(E)$ = probabilidad a priori de tener la patologîa - $\mathbb{P}(W_i|E)$ = probabilidad de que el factor W_i aparezca en pacientes enfermos - $\mathbb{P}(W_i|\overline{E})$ = probabilidad de que el factor W_i aparezca en pacientes sanos
Problema: Clasificar a un paciente que presenta los factores W_1, W_2, W_3.
Solución usando Bayes:
Se clasifica como de "Alto Riesgo" si la probabilidad a posteriori es alta:

$$\frac{\mathbb{P}(E|W_1 \cap W_2 \cap W_3)}{\mathbb{P}(\overline{E}|W_1 \cap W_2 \cap W_3)} > c$$

donde c es un umbral clînico.
Asumiendo independencia condicional de los factores:

$$\frac{\mathbb{P}(E|W_1 \cap W_2 \cap W_3)}{\mathbb{P}(\overline{E}|W_1 \cap W_2 \cap W_3)} \approx \frac{\mathbb{P}(E)}{\mathbb{P}(\overline{E})} \cdot \prod_{i=1}^{3} \frac{\mathbb{P}(W_i|E)}{\mathbb{P}(W_i|\overline{E})}$$

Aplicación clínica: - Permite integrar múltiples biomarcadores o factores de riesgo de manera eficiente. - Proporciona una medida de riesgo individualizada (probabilidad a posteriori). - Ayuda en la toma de decisiones clínicas para el cribado o tratamiento preventivo.

3.11. Resumen

3.11.1. Conceptos Clave

- **Experimento aleatorio:** Un proceso con resultado incierto pero repetible
- **Espacio muestral Ω:** El conjunto de todos los resultados posibles
- **Evento:** Un subconjunto del espacio muestral
- **Partición:** Una división del espacio muestral en eventos mutuamente disjuntos

3.11.2. Definiciones de Probabilidad

Definición	Ventajas	Limitaciones
Clásica (Laplace)	Simple, intuitiva	Requiere equiprobabilidad
Frecuentista (von Mises)	Empîrica, práctica	Requiere infinitas repeticiones
Axiomática (Kolmogorov)	General, rigurosa	Abstracta

3.11.3. Fórmulas Importantes

Propiedades básicas:

$$\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A), \quad \mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(\Omega) = 1$$

Regla de la suma:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Probabilidad condicional:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Regla de la multiplicación:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B|A)$$

Independencia:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) \text{ si y solo si } A \text{ y } B \text{ son independientes}$$

Ley de la probabilidad total:

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i)$$

Teorema de Bayes:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}$$

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j) \cdot \mathbb{P}(A_j)}{\sum_{i=1}^{n} \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i)}$$

3.12. Ejercicios

Ejercicio 3.1: En una clase hay 30 estudiantes de doctorado asistiendo a un seminario de Bioestadîstica avanzada en Medicina. De ellos, 18 se graduaron en Medicina, 15 tienen un grado en Estadistica y 10 estudiantes estudiaron ambos grados. Si seleccionamos un estudiante al azar:

a) ¿Cuál es la probabilidad de que se graduado en Medicina o estadîstica?
b) ¿Cuál es la probabilidad de que haya estudiado solo Medicina?
c) ¿Cuál es la probabilidad de que no hay estudiado ninguna de las dos?

Ejercicio 3.2: Un médico tiene dos opciones terapeúticas de antibióticos de primera lînea (A o B) y tres de segunda lînea de administración conjunta al tratamiento principal (1, 2, o 3) para el tratamiento de una enfermedad infecciosa y todas las opciones terapeúticas son igualmente eficaces. Si todas las opciones se eligen al azar:

a) ¿Cuál es el espacio muestral?
b) ¿Cuál es la probabilidad de elegir el antibiotico de primera lînea A?
c) ¿Cuál es la probabilidad de elegir el antibiotico de primera lînea B y el de segunda lînea 2?

Ejercicio 3.3: En un hospital, la probabilidad de que un paciente ingrese por fractura de cadera en un dîa es 0.1. Si evaluamos tres dîas independientemente:

a) ¿Cuál es la probabilidad de que no ingrese ningún paciente?
b) ¿Cuál es la probabilidad de que exactamente ingrese uno?

Ejercicio 3.4: Se sabe que:

- El 60 % de los pacientes de una consulta son mujeres
- El 40 % de las mujeres tinen menos de 30 años
- El 30 % de los hombres tienen menos de 30 años

Si seleccionamos un cliente al azar:

a) ¿Cuál es la probabilidad de que un paciente tenga menos de 30 años?

b) Si un paciente tiene 30 años, ¿cuál es la probabilidad de que sea mujer?

Ejercicio 3.5: Un test para detectar una enfermedad rara tiene:

- 98 % de sensibilidad (detecta la enfermedad cuando está presente)
- 95 % de especificidad (identifica correctamente a personas sanas)
- La enfermedad afecta al 0.5 % de la población

a) Si una persona da positivo en el test, ¿cuál es la probabilidad de que realmente esté enferma?
b) Interprete el resultado. ¿Por qué la probabilidad no es más alta?

Ejercicio 3.6: Una empresa privada vende ecografos a tres hospitales en tres ciudades distintas: Madrid (40 % de ventas), Barcelona (35 %) y Valencia (25 %). La probabilidad de cumplir la cuota de ventas es:

- Madrid: 0.8
- Barcelona: 0.85
- Valencia: 0.7

Si se selecciona un mes al azar:

a) ¿Cuál es la probabilidad de que se cumpla la cuota?
b) Si se cumplió la cuota, ¿cuál es la probabilidad de que haya sido en Barcelona?

3.13. Respuestas a los Ejercicios

Ejercicio 3.1: Diagramas de Venn
- a) P(M E) = P(M) + P(E) - P(M E) = (18+15-10)/30 = 23/30 0.767
- b) P(Solo M) = (18-10)/30 = 8/30 0.267
- c) P(Ninguna) = 1 - 23/30 = 7/30 0.233

Ejercicio 3.2: Espacio Muestral y Probabilidad
- a) Ω = {A1, A2, A3, B1, B2, B3}, $|\Omega| = 6$
- b) P(A) = 3/6 = 0.5
- c) P(B y 2) = 1/6 0.167

Ejercicio 3.3: Independencia
- a) P(Ningún ingreso) = $(0.9)^3 = 0.729$
- b) P(exactamente 1 ingreso) = C(3,1)$\times(0.1)^1\times(0.9)^2 = 3\times0.1\times0.81 = 0.243$
- Nota: Más adelante veremos como caalcular esta probabilidad usando la fórmula de la distribución Binomial (n, p).

Ejercicio 3.4: Probabilidad Total y Teorema de Bayes
- a) P(Menos de 30 años) = $0.6\times0.4 + 0.4\times0.3 = 0.24 + 0.12 = 0.36$
- b) P(Mujer|Menos de 30 años) = (0.6x0.4)/0.36 = 0.24/0.36 = 2/3 0.667

Ejercicio 3.5: Test Diagnóstico (Bayes)
- a) P(Enfermedad|Positivo) = $(0.98\times0.005)/(0.98\times0.005 + 0.05\times0.995) = 0.0049/(0.0049+0.04975)$

0.090

- b) Aunque el test es muy preciso (98 % sensibilidad), como la enfermedad es rara (0.5 %), la mayoría de positivos son falsos positivos.

Ejercicio 3.6: Probabilidad Total y Bayes (Hospitales)

- a) P(Cuota) = 0.4×0.8 + 0.35×0.85 + 0.25×0.7 = 0.32 + 0.2975 + 0.175 = 0.7925
- b) P(Barcelona|Cuota) = (0.35×0.85)/0.7925 = 0.2975/0.7925 0.375

Métodos Avanzados

Para ampliar los contenidos de este capítulo con técnicas estadísticas avanzadas, visita:
→ Bioestadística Avanzada — M.A. Luque Fernández

Capítulo 4

Semana 4 — Variables Aleatorias y Distribuciones de Probabilidad

4.1. Introducción

Las variables aleatorias son el concepto fundamental que conecta la teoría de la probabilidad con la estadística práctica. Esta semana exploramos los conceptos de variable aleatoria, sus funciones de distribución, y las distribuciones más importantes que usarás en bioestadística y análisis de datos.

4.2. Variables Aleatorias

Definición: Variable Aleatoria

Una **variable aleatoria** X es una función que asigna un número real a cada resultado en el espacio muestral. Formalmente:

$$X : \Omega \to \mathbb{R}, \quad X(\omega) = x$$

donde Ω es el espacio muestral.

4.2.1. Tipos de Variables Aleatorias

Variable Aleatoria Discreta: Toma un número finito o infinito numerable de valores.

$$X \in \{x_1, x_2, x_3, ...\}$$

Variable Aleatoria Continua: Puede tomar cualquier valor en un intervalo o en toda la recta real.

$$X \in [a, b] \text{ o } X \in \mathbb{R}$$

4.2.2. Función de Probabilidad y Función de Distribución

Para variables **discretas**, la **Función de Masa de Probabilidad (PMF)** es:

$$P(X = x_i) = f(x_i)$$

con propiedades:

- $f(x_i) \geq 0$ para todo i
- $\sum_i f(x_i) = 1$

Para variables **continuas**, la **Función de Densidad de Probabilidad (PDF)** es:

$$f(x) \geq 0 \text{ para todo } x$$

$$\int_{-\infty}^{+\infty} f(x)\, dx = 1$$

La **Función de Distribución Acumulada (CDF)** es:

$$F(x) = P(X \leq x)$$

con propiedades:

- $0 \leq F(x) \leq 1$
- F es no decreciente
- $\lim_{x \to -\infty} F(x) = 0$ y $\lim_{x \to +\infty} F(x) = 1$

Para variables continuas:

$$f(x) = \frac{dF(x)}{dx}$$

4.3. Funciones de Probabilidad en R

R ofrece un conjunto consistente de funciones para trabajar con distribuciones de probabilidad. Para cualquier distribución (`distribucion`), se dispone de cuatro funciones prefijadas:

- **`d` (density / mass):** `ddistribucion(x, ...)` — Calcula la función de densidad (PDF) o masa (PMF).
- **`p` (probability):** `pdistribucion(q, ...)` — Calcula la CDF: $P(X \leq q)$.
- **`q` (quantile):** `qdistribucion(p, ...)` — Calcula la función cuantil (inversa de CDF).
- **`r` (random):** `rdistribucion(n, ...)` — Genera números aleatorios.

4.4. Fundamentos de Cálculo de Probabilidades: Discreta vs. Continua

Antes de explorar distribuciones específicas, es vital comprender las diferencias fundamentales en cómo calculamos probabilidades dependiendo de la naturaleza de la variable.

4.4.1. La Singularidad del Punto en Variables Continuas

En distribuciones continuas (como la Normal), la probabilidad de que una variable tome un valor puntual exacto es siempre cero:

$$P(X = x) = 0$$

¿Por qué? Matemáticamente, la probabilidad de un punto se define como el área bajo la curva en un intervalo de ancho cero:

$$P(X = x) = \int_{x}^{x} f(t)\,dt = 0$$

En estadîstica práctica, esto significa que solo tiene sentido hablar de probabilidades en **intervalos**: $P(a \leq X \leq b) = \int_a^b f(t)\,dt$. Por lo tanto, para una continua, $P(X \leq x) = P(X < x)$.

4.4.2. El Cuidado de las Desigualdades en Variables Discretas

En variables discretas (como la Binomial), $P(X = k) > 0$. Esto implica que debemos ser extremadamente precisos al usar operadores de comparación ($<, \leq, >, \geq$), ya que el incluir o excluir un punto cambia el resultado significativamente.

La Regla de Oro: - $P(X \leq k) = P(X < k) + P(X = k)$ - $P(X > k) = 1 - P(X \leq k)$ - $P(X \geq k) = 1 - P(X \leq k - 1)$

4.4.2.1. Ejemplo comparativo en R

Cuadro 4.1 Cálculo comparativo discreto vs continuo

```
# Discreta: Binomial(n=10, p=0.5)
# P(X = 5) es distinta de cero
print(paste("P(X = 5) =", dbinom(5, 10, 0.5)))
```

```
[1] "P(X = 5) = 0.24609375"
```

Cuadro 4.2 Cálculo comparativo discreto vs continuo

```
print(paste("P(X <= 5) =", pbinom(5, 10, 0.5)))
```

```
[1] "P(X <= 5) = 0.623046875"
```

```
[1] "P(X < 5) = 0.376953125"
```

```
[1] "P(X = 0) = 0  (por definición de variable continua)"
```

Cuadro 4.3 Cálculo comparativo discreto vs continuo

```
print(paste("P(X < 5) =", pbinom(4, 10, 0.5)))
```

Cuadro 4.4 Cálculo comparativo discreto vs continuo

```
# Continua: Normal(mu=0, sd=1)
# Atención: en continuas P(X = 0) = 0 por definición.
# dnorm(0) NO es probabilidad: es la densidad f(0) en el punto 0.
print("P(X = 0) = 0  (por definición de variable continua)")
```

```
[1] "f(0) (densidad) = 0.398942280401433"
```

```
[1] "P(X <= 0) = 0.5"
```

```
[1] "P(X < 0) = 0.5"
```

Resumen de operadores

Operador	Continuo	Discreto
$P(X = k)$	0	$f(k)$
$P(X \leq k)$	`pdist(k)`	`pdist(k)`
$P(X < k)$	`pdist(k)`	`pdist(k-1)`
$P(X > k)$	`1 - pdist(k)`	`1 - pdist(k)`
$P(X \geq k)$	`1 - pdist(k)`	`1 - pdist(k-1)`

4.5. Varianza y Desviación Estándar

Definición: Varianza

La **varianza** mide la dispersión de una variable aleatoria respecto a su esperanza:

$$\text{Var}(X) = \sigma_X^2 = E[(X - E(X))^2]$$

Fórmula computacional equivalente:

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

Caso discreto:

$$\text{Var}(X) = \sum_{i=1}^{\infty} [x_i - E(X)]^2 \cdot f(x_i) = \sum_{i=1}^{\infty} x_i^2 \cdot f(x_i) - [E(X)]^2$$

Cuadro 4.5 Cálculo comparativo discreto vs continuo

```
print(paste("f(0) (densidad) =", dnorm(0)))
```

Cuadro 4.6 Cálculo comparativo discreto vs continuo

```
# P(X <= 0) es 0.5
print(paste("P(X <= 0) =", pnorm(0)))
```

Caso continuo:

$$\text{Var}(X) = \int_{-\infty}^{+\infty} [x - E(X)]^2 \cdot f(x)\, dx = \int_{-\infty}^{+\infty} x^2 \cdot f(x)\, dx - [E(X)]^2$$

4.5.1. Desviación Estándar poblacional

La **desviación estándar** es la raîz cuadrada positiva de la varianza:

$$\sigma_X = \sqrt{\text{Var}(X)}$$

Tiene las mismas unidades que la variable original.

4.5.2. Propiedades de la Varianza

⚠ Propiedades Importantes

1. **Transformación lineal:** Si $Y = a + bX$:

$$\text{Var}(Y) = b^2 \cdot \text{Var}(X)$$

(nota: la constante aditiva no afecta la varianza)

2. **Suma de variables independientes:** Si X e Y son independientes:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$$

Cuadro 4.7 Cálculo comparativo discreto vs continuo

```
# En continuas P(X < 0) = P(X <= 0) porque P(X = 0) = 0; por eso usamos pnorm(0). Lo anterior t
print(paste("P(X < 0) =", pnorm(0)))
```

4.6. Estandarización (Transformación Z)

i Definición: Transformación Z

La **estandarización** o **normalización** transforma una variable aleatoria para que tenga media 0 y varianza 1:

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}} = \frac{X - \mu_X}{\sigma_X}$$

Propiedades de Z:

- $E(Z) = 0$
- $\text{Var}(Z) = 1$
- Z es adimensional (sin unidades)

4.7. Variables Aleatorias Bivariantes

4.7.1. Distribuciones Conjunta, Marginal y Condicional (Caso Discreto)

i Distribuciones Discretas Bivariantes

La **función de probabilidad conjunta** es:

$$P(X = x_i, Y = y_j) = f(x_i, y_j)$$

con propiedades:

- $f(x_i, y_j) \geq 0$
- $\sum_i \sum_j f(x_i, y_j) = 1$

Las **distribuciones marginales** se obtienen sumando:

$$f(x_i) = P(X = x_i) = \sum_j f(x_i, y_j)$$

$$f(y_j) = P(Y = y_j) = \sum_i f(x_i, y_j)$$

La **distribución condicional** es:

$$P(X = x_i|Y = y_j) = \frac{f(x_i, y_j)}{f(y_j)}$$

⚠ Independencia

Dos variables aleatorias X e Y son **independientes** si:

$$f(x_i, y_j) = f(x_i) \cdot f(y_j) \quad \text{para todo } x_i, y_j$$

Importante: La independencia implica que la distribución condicional iguala la marginal:

$$P(X = x_i|Y = y_j) = P(X = x_i)$$

4.7.2. Distribuciones Conjunta, Marginal y Condicional (Caso Continuo)

ℹ Distribuciones Continuas Bivariantes

La **función de densidad conjunta** satisface:

$$f(x, y) \geq 0$$

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\, dx\, dy = 1$$

Las **funciones de densidad marginales** son:

$$f(x) = \int_{-\infty}^{+\infty} f(x, y)\, dy$$

$$f(y) = \int_{-\infty}^{+\infty} f(x, y)\, dx$$

La **densidad condicional** es:

$$f(x|y) = \frac{f(x, y)}{f(y)}$$

4.7.3. Covarianza y Correlación

ℹ Definición: Covarianza

La **covarianza** mide la variabilidad conjunta de dos variables aleatorias:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

Fórmula computacional:

$$\text{Cov}(X,Y) = E(XY) - E(X) \cdot E(Y)$$

Propiedades:

- $\text{Cov}(X,X) = \text{Var}(X)$
- Si X e Y son independientes: $\text{Cov}(X,Y) = 0$
- **Nota importante:** $\text{Cov}(X,Y) = 0$ NO implica independencia

i Definición: Correlacion

El **coeficiente de correlación** (o coeficiente de correlación de Pearson) en una población es:

$$\rho(X,Y) = \text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Propiedades:

- $-1 \leq \rho(X,Y) \leq 1$
- $\rho = 1$: correlación positiva perfecta
- $\rho = 0$: no hay correlación lineal
- $\rho = -1$: correlación negativa perfecta

4.8. Distribuciones Discretas Importantes

4.8.1. Distribución Uniforme Discreta $U(n)$

i Definición: Distribución Uniforme Discreta

Una variable aleatoria $X \sim U(n)$ toma n valores con igual probabilidad:

$$P(X = x_i) = \frac{1}{n} \quad \text{para } i = 1, 2, \ldots, n$$

Parámetros: n (número de valores posibles)
Esperanza y Varianza:

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))^2$$

En R: No existe función directa; usar `sample()` o implementar manualmente.

Ejemplo 4.1: Asignación aleatoria de pacientes a 6 brazos de un ensayo clínico

Un ensayo clínico **factorial 2×3** evalúa dos fármacos antihipertensivos (A, B) combinados con tres dosis (baja, media, alta), generando 6 brazos de tratamiento. Para evitar el sesgo de selección, cada paciente es asignado **aleatoriamente** a uno de los 6 brazos con la misma probabilidad. Si codificamos los brazos $k = 1, 2, \ldots, 6$, la variable X "brazo asignado" sigue una distribución uniforme discreta $X \sim U(6)$:

$$P(X = k) = \frac{1}{6} \quad \text{para } k = 1, 2, 3, 4, 5, 6$$

$$E(X) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

$$\text{Var}(X) = \frac{1}{6}\sum_{k=1}^{6}(k - 3.5)^2 \approx 2.92$$

La esperanza $E(X) = 3.5$ no tiene aquí interpretación clínica directa —los brazos son etiquetas, no magnitudes— sino que actúa como **verificación de balance**: si la aleatorización funciona correctamente, la media de los códigos asignados en una muestra grande tiende a 3.5. Una media muy alejada indicaría un problema en el procedimiento de aleatorización.

Cuadro 4.8 Código R

```r
# En R: calcular probabilidades para distribución uniforme discreta
k     <- 1:6
probs <- rep(1/6, 6)
mean_x <- sum(k * probs)
var_x  <- sum((k - mean_x)^2 * probs)

cat(sprintf("P(X = k) =%.4f para cada brazo k = 1..6\n", probs[1]))
```

```
P(X = k) = 0.1667 para cada brazo k = 1..6
```

Cuadro 4.9 Código R

```r
cat(sprintf("E(X) =%.2f   Var(X) =%.3f\n", mean_x, var_x))
```

```
E(X) = 3.50   Var(X) = 2.917
```

Cuadro 4.10 Código R

```r
# Simulacion de la asignacion aleatoria de 600 pacientes
set.seed(2026)
n_pac <- 600
asig  <- sample(1:6, size = n_pac, replace = TRUE)

# Verificacion del balance entre brazos
tabla_asig <- table(factor(asig, levels = 1:6,
                           labels = c("A-baja","A-media","A-alta",
                                      "B-baja","B-media","B-alta")))
print(tabla_asig)
```

```
 A-baja A-media  A-alta  B-baja B-media  B-alta
    100     112      76     102     108     102
```

Cuadro 4.11 Código R

```r
cat(sprintf("\nMedia muestral de los codigos asignados =%.3f (esperada: 3.5)\n",
            mean(asig)))
```

```
Media muestral de los codigos asignados = 3.520 (esperada: 3.5)
```

Interpretación clínica: Con 600 pacientes asignados aleatoriamente, cada brazo recibe aproximadamente $100 \approx n/6$ pacientes y la media de los códigos asignados se aproxima al valor teórico 3.5, confirmando que la aleatorización está bien implementada. Este principio se utiliza en todos los ensayos clínicos aleatorizados (RCT) para garantizar la **comparabilidad de los grupos** y eliminar el sesgo de confusión por características basales.

4.8.2. Distribución de Bernoulli $B(p)$

Definición: Distribución de Bernoulli

Una variable aleatoria $X \sim B(p)$ modela un experimento con dos resultados: éxito ($X = 1$) o fracaso ($X = 0$):

$$P(X = x) = \begin{cases} p & \text{si } x = 1 \\ 1 - p & \text{si } x = 0 \\ 0 & \text{en otro caso} \end{cases}$$

O compactamente: $P(X = x) = p^x (1 - p)^{1-x}$ para $x \in \{0, 1\}$

Parámetro: $p \in [0, 1]$ (probabilidad de éxito)

Esperanza y Varianza:

$$E(X) = p$$

$$\text{Var}(X) = p(1-p)$$

En R: `dbinom(x, size=1, prob=p)` o `rbinom(n, size=1, prob=p)`

4.8.3. Distribución Binomial $B(n, p)$

Definición: Distribución Binomial

Una variable aleatoria $X \sim B(n, p)$ cuenta el número de éxitos en n ensayos Bernoulli independientes:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{para } k = 0, 1, \ldots, n$$

donde $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
Parámetros: n (número de ensayos), p (probabilidad de éxito)
Esperanza y Varianza:

$$E(X) = n \cdot p$$

$$\text{Var}(X) = n \cdot p \cdot (1-p)$$

En R: - `dbinom(k, size=n, prob=p)` — función de probabilidad - `pbinom(k, size=n, prob=p)` — función de distribución - `rbinom(N, size=n, prob=p)` — generar muestras aleatorias

Ejemplo 4.2: Distribución Binomial

Un medicamento cura el 70 % de los pacientes. Si se trata a 10 pacientes, ¿cuál es la probabilidad de que exactamente 7 sanen?

$$P(X = 7) = \binom{10}{7} (0.7)^7 (0.3)^3 = 120 \cdot 0.0823 \cdot 0.027 \approx 0.267$$

En R:

Cuadro 4.12 Código R

```
# Probabilidad de exactamente 7 curaciones
dbinom(7, size=10, prob=0.7)
```

```
[1] 0.2668279
```

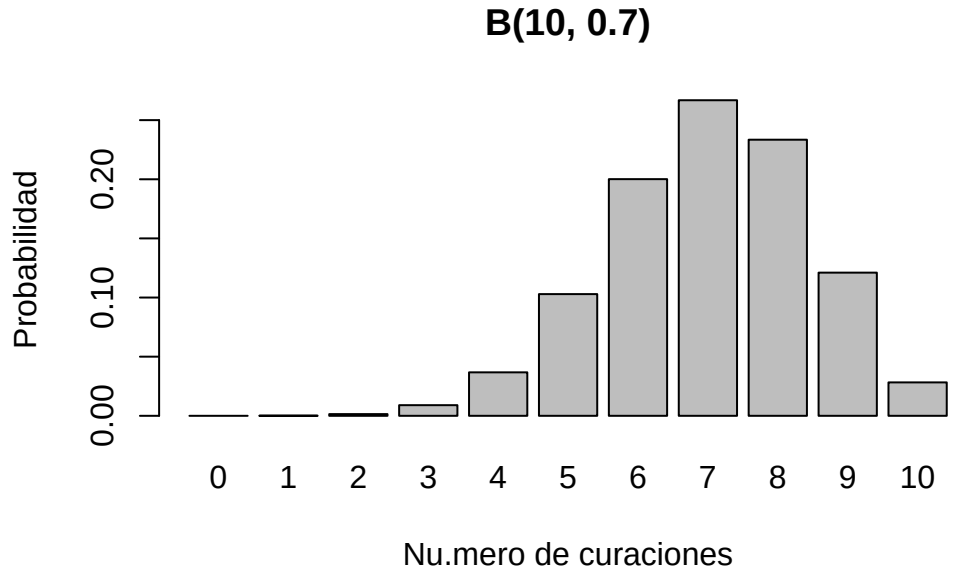
Cuadro 4.13 Código R

```
# Probabilidad de 7 o menos curaciones
pbinom(7, size=10, prob=0.7)
```

```
[1] 0.6172172
```

Cuadro 4.14 Código R

```
# Visualización
x <- 0:10
probs <- dbinom(x, size=10, prob=0.7)
barplot(probs, names.arg=x, main="B(10, 0.7)",
        xlab="Número de curaciones", ylab="Probabilidad")
```



4.8.4. Distribución de Poisson Po(λ)

Definición: Distribución de Poisson

Una variable aleatoria $X \sim \text{Po}(\lambda)$ modela el número de eventos raros que ocurren en un intervalo de tiempo o espacio fijo:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{para } k = 0, 1, 2, \ldots$$

Parámetro: $\lambda > 0$ (tasa de eventos)
Esperanza y Varianza:

$$E(X) = \lambda$$

$$\text{Var}(X) = \lambda$$

(Nota: en Poisson, media y varianza son iguales)
Propiedad Reproductiva: Si $X \sim \text{Po}(\lambda_1)$ e $Y \sim \text{Po}(\lambda_2)$ son independientes:

$$X + Y \sim \text{Po}(\lambda_1 + \lambda_2)$$

En R: - `dpois(k, lambda= )` — función de probabilidad - `ppois(k, lambda= )` — función de distribución - `rpois(N, lambda= )` — generar muestras aleatorias

Ejemplo 4.3: Distribución de Poisson

El número de reacciones adversas en una clínica sigue una distribución Poisson con $\lambda = 3$ por día. ¿Cuál es la probabilidad de que ocurran exactamente 5 reacciones mañana?

$$P(X = 5) = \frac{3^5 e^{-3}}{5!} = \frac{243 \cdot 0.0498}{120} \approx 0.1008$$

En R:

Cuadro 4.15 Código R

```
# Probabilidad de exactamente 5 eventos
dpois(5, lambda=3)
```

```
[1] 0.1008188
```

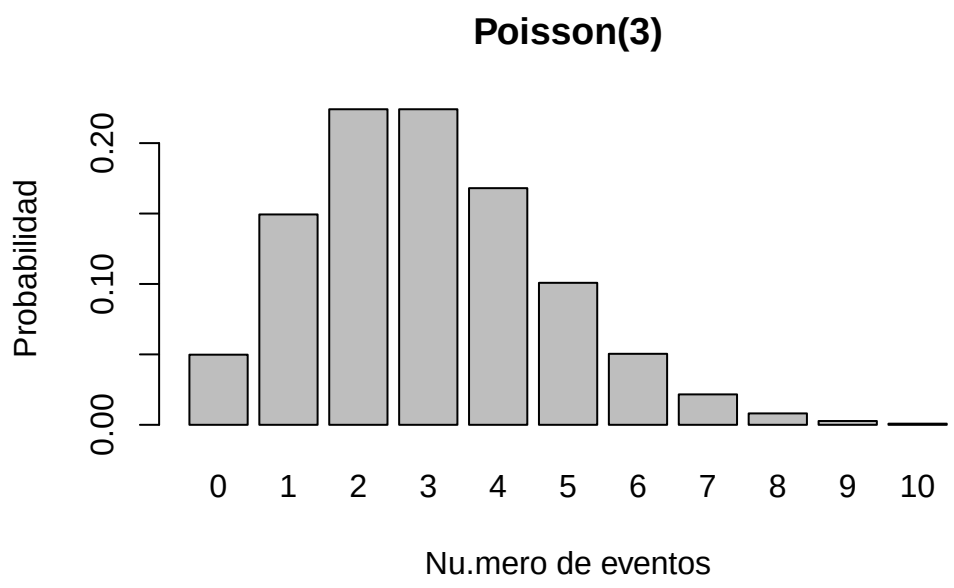
Cuadro 4.16 Código R

```
# Probabilidad de 5 o menos eventos
ppois(5, lambda=3)
```

```
[1] 0.9160821
```

Cuadro 4.17 Código R

```
# Visualización
x <- 0:10
probs <- dpois(x, lambda=3)
barplot(probs, names.arg=x, main="Poisson(3)",
        xlab="Número de eventos", ylab="Probabilidad")
```



4.9. Distribuciones Continuas Importantes

4.9.1. Distribución Uniforme Continua $U(a, b)$

Definición: Distribución Uniforme Continua

Una variable aleatoria $X \sim U(a, b)$ tiene densidad constante en el intervalo $[a, b]$:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{en otro caso} \end{cases}$$

Función de Distribución:

$$F(x) = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{si } x > b \end{cases}$$

Parámetros: a, b (límites del intervalo)
Esperanza y Varianza:

$$E(X) = \frac{a+b}{2}$$

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

En R: `dunif(x, min=a, max=b)`, `punif(x, min=a, max=b)`, `runif(n, min=a, max=b)`

Ejemplo: Tiempo de espera

Un paciente debe esperar entre 0 y 30 minutos para ser atendido ($X \sim U(0, 30)$).

$$E(X) = \frac{0 + 30}{2} = 15 \text{ minutos}$$

$$\text{Var}(X) = \frac{(30 - 0)^2}{12} = 75$$

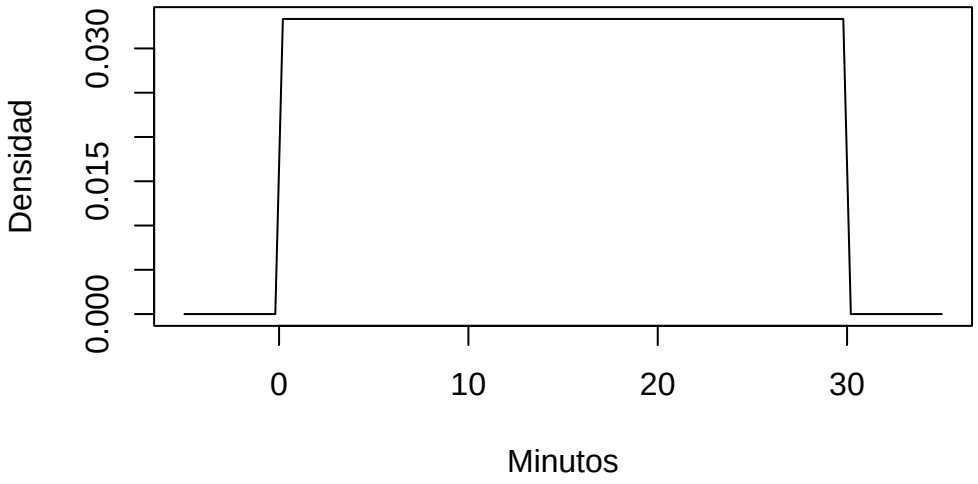
Probabilidad de esperar menos de 10 minutos:

$$P(X \leq 10) = \frac{10 - 0}{30 - 0} = \frac{1}{3} \approx 0.333$$

En R:

Cuadro 4.18 Código R

```
# Densidad
curve(dunif(x, min=0, max=30), from=-5, to=35,
      xlab="Minutos", ylab="Densidad")
```



Cuadro 4.19 Código R

```
# P(X   10)
punif(10, min=0, max=30)
```

```
[1] 0.3333333
```

4.9.2. Distribución Exponencial Exp(λ)

Definición: Distribución Exponencial

Una variable aleatoria $X \sim \text{Exp}(\lambda)$ modela el tiempo hasta el siguiente evento en un proceso de Poisson:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Función de Distribución:

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Parámetro: $\lambda > 0$ (tasa de evento)
Esperanza y Varianza:

$$E(X) = \frac{1}{\lambda}$$

$$\text{Var}(X) = \frac{1}{\lambda^2}$$

Propiedad de Falta de Memoria: $P(X > s + t | X > s) = P(X > t)$
En R: `dexp(x, rate= )`, `pexp(x, rate= )`, `rexp(n, rate= )`

Ejemplo: Tiempo de supervivencia

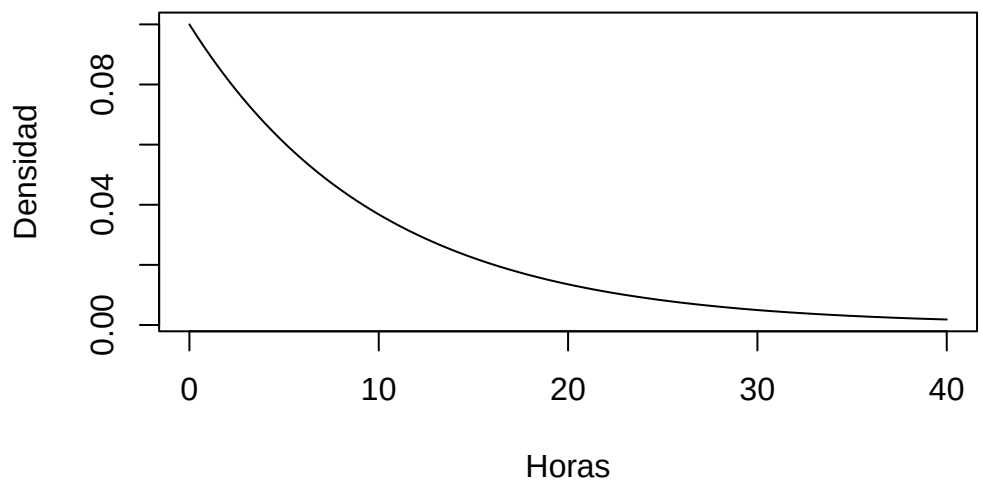
El tiempo de supervivencia de células en cultivo sigue Exp($\lambda = 0.1$) (en horas).
Tiempo medio de supervivencia: $E(X) = 1/0.1 = 10$ horas
Probabilidad de que una célula sobreviva más de 5 horas:

$$P(X > 5) = e^{-0.1 \cdot 5} = e^{-0.5} \approx 0.606$$

En R:

Cuadro 4.20 Código R

```
# Densidad
curve(dexp(x, rate=0.1), from=0, to=40,
      xlab="Horas", ylab="Densidad")
```



Cuadro 4.21 Código R

```
# P(X > 5)
1 - pexp(5, rate=0.1)
```

```
[1] 0.6065307
```

Cuadro 4.22 Código R

```
# O equivalentemente:
pexp(5, rate=0.1, lower.tail=FALSE)
```

```
[1] 0.6065307
```

4.9.3. Distribución Normal $N(\mu, \sigma^2)$

Definición: Distribución Normal

Una variable aleatoria $X \sim N(\mu, \sigma^2)$ tiene la función de densidad:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad \text{para} \ -\infty < x < +\infty$$

Parámetros: μ (media), σ^2 (varianza); $\sigma > 0$
Propiedades: - Simétrica alrededor de μ - Media = Mediana = Moda = μ - $E(X) = \mu$ - $\text{Var}(X) = \sigma^2$
Función de Distribución: No tiene forma cerrada; se usa notación $\Phi(z)$ para la CDF estándar.
En R: `dnorm(x, mean= , sd= )`, `pnorm(x, mean= , sd= )`, `rnorm(n, mean= , sd= )`

Distribución Normal Estándar $N(0,1)$

La **distribución normal estándar** tiene $\mu = 0$ y $\sigma = 1$:

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Transformación Z (Estandarización):

$$Z = \frac{X - \mu}{\sigma}$$

Si $X \sim N(\mu, \sigma^2)$, entonces $Z \sim N(0,1)$
Esta transformación es fundamental porque:

- Permite usar tablas únicas de probabilidades
- Valores de Z son directamente comparables entre variables diferentes

Ejemplo: Medidas antropométricas

El peso de adultos sigue $N(\mu = 75, \sigma = 12)$ kg.
¿Cuál es la probabilidad de que una persona pese menos de 90 kg?

$$P(X \leq 90) = P\left(Z \leq \frac{90 - 75}{12}\right) = P(Z \leq 1.25)$$

En R:

Cuadro 4.23 Código R

```
# Directamente con parámetros
pnorm(90, mean=75, sd=12)
```

```
[1] 0.8943502
```

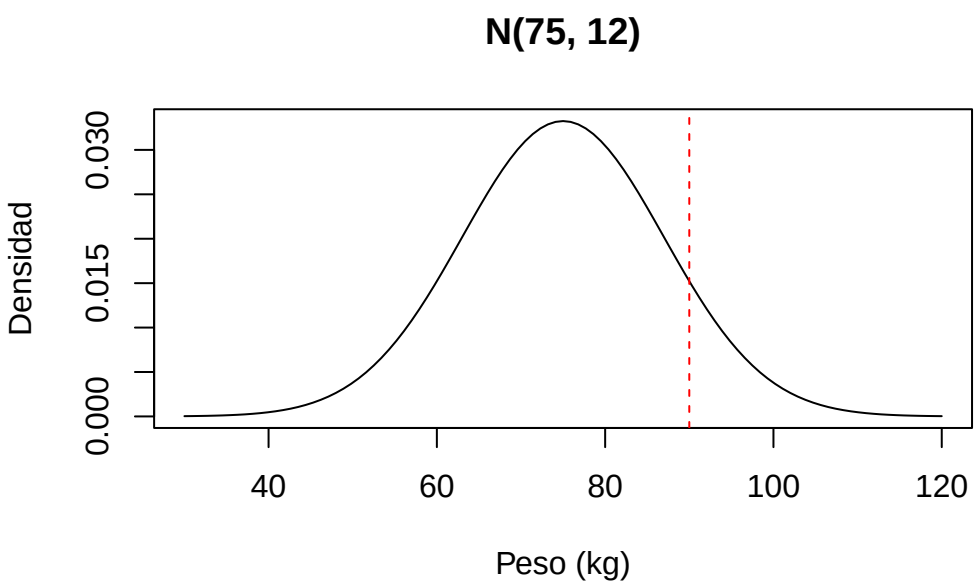
Cuadro 4.24 Código R

```
# O usando estandarización
z <- (90 - 75) / 12
pnorm(z)  # Usa N(0,1) por defecto
```

```
[1] 0.8943502
```

Cuadro 4.25 Código R

```
# Visualización
curve(dnorm(x, mean=75, sd=12), from=30, to=120,
      xlab="Peso (kg)", ylab="Densidad",
      main="N(75, 12)")
abline(v=90, col="red", lty=2)
```



La probabilidad es aproximadamente **0.894** (89.4 %).

Ejemplo: Interpretación de z-scores

Para pacientes con presión sistólica $N(\mu = 120, \sigma = 15)$ mmHg:

- Paciente A con presión = 120 mmHg: $z = 0$ (media)
- Paciente B con presión = 135 mmHg: $z = 1$ (1 DE arriba de la media)
- Paciente C con presión = 105 mmHg: $z = -1$ (1 DE abajo de la media)
- Paciente D con presión = 150 mmHg: $z = 2$ (2 DE arriba de la media)

Los z-scores permiten comparar directamente aunque las variables originales tengan unidades diferentes.

4.9.4. Aproximación Normal a la Binomial y Corrección de Continuidad

Históricamente, cuando el cálculo manual o las tablas estadísticas eran las principales herramientas, la aproximación de distribuciones discretas (como la binomial) mediante distribuciones continuas (como la normal) era una técnica invaluable para estimar probabilidades. Esta aproximación, una aplicación directa del Teorema Central del Límite, permitía simplificar cálculos complejos. Sin embargo, con el advenimiento del software estadístico moderno, que utiliza algoritmos exactos (como la Función Beta Incompleta Regularizada para la binomial) con alta precisión computacional, el uso de estas aproximaciones para el cálculo de probabilidades exactas se ha vuelto menos recomendable. Aunque la aproximación sigue siendo fundamental para la comprensión teórica y para la derivación de métodos como los intervalos de confianza o tests Z para proporciones, para la obtención de valores de probabilidad específicos, se prefiere el cálculo directo para evitar posibles inexactitudes (véase Saad et al. [2020]).

La **corrección de continuidad** es un ajuste que mejora la aproximación al compensar el paso de una variable discreta a una continua. Cada barra del histograma binomial en k ocupa el intervalo

$[k-0.5,\ k+0.5]$.

Definición: Aproximación Normal a la Binomial

Si $X \sim B(n,p)$ con $np \geq 5$ y $n(1-p) \geq 5$, entonces:

$$X \overset{\text{aprox.}}{\sim} N(\mu = np,\ \sigma^2 = np(1-p))$$

La **corrección de continuidad** mejora la aproximación al ajustar ± 0.5 para compensar el paso de una variable discreta a una continua. Cada barra del histograma binomial en k ocupa el intervalo $[k-0.5,\ k+0.5]$.

Probabilidad exacta	Sin corrección	**Con corrección**
$P(X \leq k)$	$\Phi\left(\frac{k-\mu}{\sigma}\right)$	$\Phi\left(\frac{k+0.5-\mu}{\sigma}\right)$
$P(X \geq k)$	$1-\Phi\left(\frac{k-\mu}{\sigma}\right)$	$1-\Phi\left(\frac{k-0.5-\mu}{\sigma}\right)$
$P(X = k)$	≈ 0	$\Phi\left(\frac{k+0.5-\mu}{\sigma}\right) - \Phi\left(\frac{k-0.5-\mu}{\sigma}\right)$

Condición de Aplicación

La aproximación normal a la binomial es válida cuando:

$$np \geq 5 \quad \text{y} \quad n(1-p) \geq 5$$

Para p cercano a 0 o 1 con n pequeño, usar la binomial exacta (`pbinom()`).

Ejemplo: Aproximación Normal a $B(20, 0.4)$ con Corrección de Continuidad

Un fármaco es efectivo en el 40 % de los pacientes. En un ensayo con $n = 20$ pacientes, ¿cuál es la probabilidad de que a lo sumo 10 respondan?
Parámetros: $\mu = 20 \times 0.4 = 8$, $\sigma = \sqrt{20 \times 0.4 \times 0.6} \approx 2.19$

	Cálculo	Resultado
Exacto (binomial)	`pbinom(10, 20, 0.4)`	**0.8725**
Normal sin corrección	$\Phi\left(\frac{10-8}{2.19}\right) = \Phi(0.913)$	0.8193

Normal con corrección	$\Phi\left(\frac{10.5-8}{2.19}\right) = \Phi(1.142)$	**0.8731**

La corrección reduce el error de aproximadamente 5.3 pp a 0.06 pp.

Cuadro 4.26 Código R

```
n <- 20; p <- 0.4
mu <- n*p; sigma <- sqrt(n*p*(1-p))
k  <- 10

cat("Exacto:               ", pbinom(k, n, p), "\n")
```

```
Exacto:               0.8724788
```

Cuadro 4.27 Código R

```
cat("Normal sin corrección:", pnorm(k, mu, sigma), "\n")
```

```
Normal sin corrección: 0.8193448
```

Cuadro 4.28 Código R

```
cat("Normal con corrección:", pnorm(k + 0.5, mu, sigma), "\n")
```

```
Normal con corrección: 0.8730835
```

La figura siguiente muestra las barras de la binomial con la curva normal superpuesta. Las líneas punteadas delimitan la barra $k = 10$ en $[9.5, 10.5]$, que es el intervalo integrado con la corrección de continuidad.

Cuadro 4.29 Código R

```r
n <- 20; p <- 0.4
mu <- n * p; sigma <- sqrt(n * p * (1 - p))
x_vals  <- 0:n
probs_b <- dbinom(x_vals, n, p)

# Histograma tipo barras binomiales
bp <- barplot(probs_b,
              names.arg = x_vals,
              col    = ifelse(x_vals == 10, "#f39200", "#cce0f5"),
              border = "white",
              xlab   = "k (número de respuestas)",
              ylab   = "Probabilidad / Densidad",
              main   = expression("Aproximación normal a " * B(20, 0.4)),
              ylim   = c(0, max(probs_b) * 1.25),
              space  = 0)

# Curva normal superpuesta (coordenadas del barplot: posición = k + 0.5)
x_c <- seq(-0.5, n + 0.5, length.out = 400)
lines(x_c + 0.5, dnorm(x_c, mu, sigma), col = "#003087", lwd = 2.5)

# Límites corrección de continuidad para k = 10
abline(v = c(10, 11), col = "#f39200", lty = 2, lwd = 1.8)

legend("topright",
       legend = c(sprintf("B(%d,%.1f)  exacto", n, p),
                  sprintf("N(%.0f,%.2f²) aprox.", mu, sigma),
                  "Límites ±0.5  (k = 10)"),
       fill   = c("#cce0f5", NA, NA),
       lty    = c(NA, 1, 2),
       lwd    = c(NA, 2.5, 1.8),
       col    = c(NA, "#003087", "#f39200"),
       bty    = "n", cex = 0.85)
```

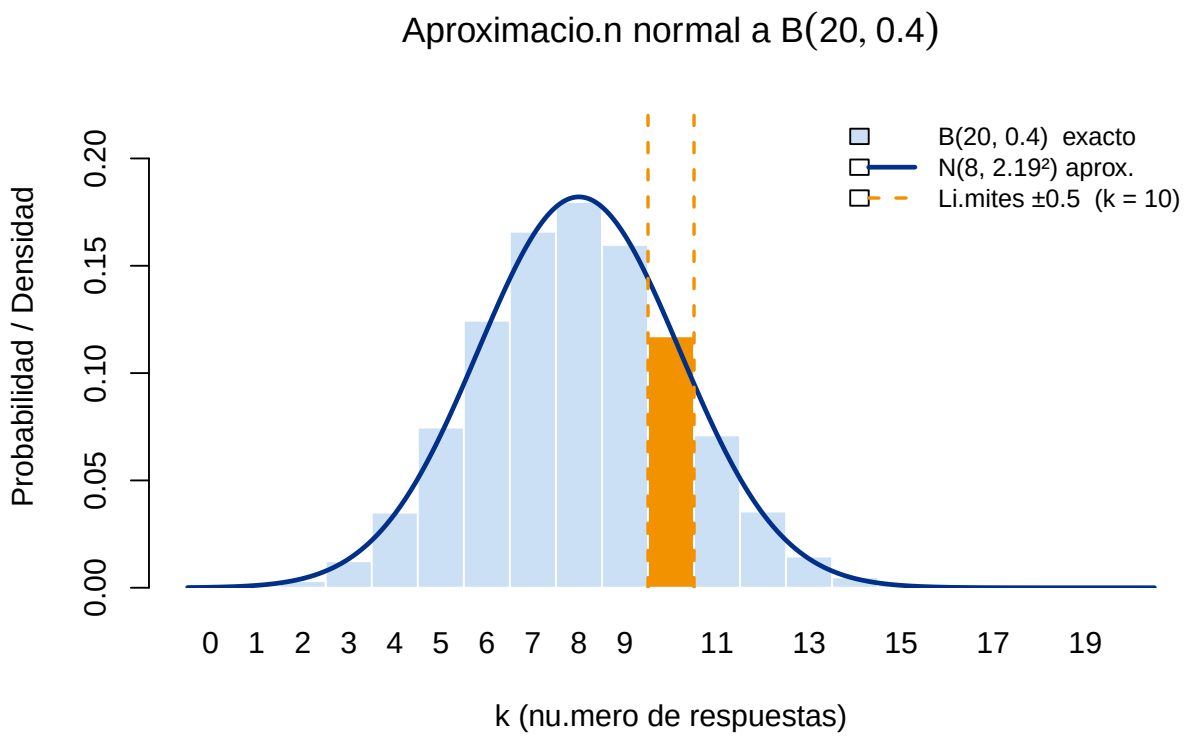


Figura 4.1: Aproximación normal a B(20, 0.4). Las barras son probabilidades binomiales exactas; la curva azul es la densidad N(8, 2.19²). La barra naranja (k = 10) y las líneas punteadas ilustran la corrección de continuidad: el área bajo la curva normal en [9.5, 10.5] corresponde a P(X = 10).

4.10. Distribuciones para Muestreo

4.10.1. Distribución Chi-Cuadrado χ^2_k

Definición: Distribución Chi-Cuadrado

Si $Z_1, Z_2, \ldots, Z_k$ son variables normales estándar independientes, entonces:

$$X = Z_1^2 + Z_2^2 + \cdots + Z_k^2 \sim \chi^2_k$$

Parámetro: k (grados de libertad)
Propiedades: - Solo toma valores positivos - $E(X) = k$ - $\text{Var}(X) = 2k$ - A medida que k aumenta, se aproxima a una distribución normal (TCL)
En R: `dchisq(x, df=k)`, `pchisq(x, df=k)`, `rchisq(n, df=k)`
Usos: Pruebas de bondad de ajuste, tablas de contingencia, intervalos de confianza para varianza.
La morfologia de la distribucion chi-cuadrado depende del numero de grados de libertad: para $k = 1$ es altamente asimetrica, pero a medida que k aumenta se vuelve mas simetrica y se

aproxima a una normal.

Vease la figura siguiente que muestra la densidad de χ^2_k para $k = 1, 3, 5, 10$. Para $k = 1$ la distribucion es muy asimetrica, con una cola larga hacia la derecha. Para $k = 3$ y $k = 5$ la distribucion se vuelve mas simetrica, aunque sigue teniendo una cola hacia la derecha. Para $k = 10$ la distribucion ya se aproxima bastante a una normal, aunque sigue siendo ligeramente asimetrica.

Cuadro 4.30 Código R

```r
library(ggplot2)
k_values <- c(1, 3, 5, 10)
x <- seq(0, 30, length.out = 400)
# Configuración de paneles
par(mfrow = c(2, 2),
    mar = c(4, 4, 2, 1))
for (k in k_values) {
  y <- dchisq(x, df = k)
  plot(
    x, y,
    type = "l",
    lwd = 2,
    col = "#003087",
    main = paste("k =", k),
    xlab = "x",
    ylab = "Densidad"
  )
  polygon(
    c(x, rev(x)),
    c(y, rep(0, length(y))),
    col = adjustcolor("#4F81BD", alpha.f = 0.35),
    border = NA
  )
  lines(x, y, lwd = 2, col = "#003087")
}
```

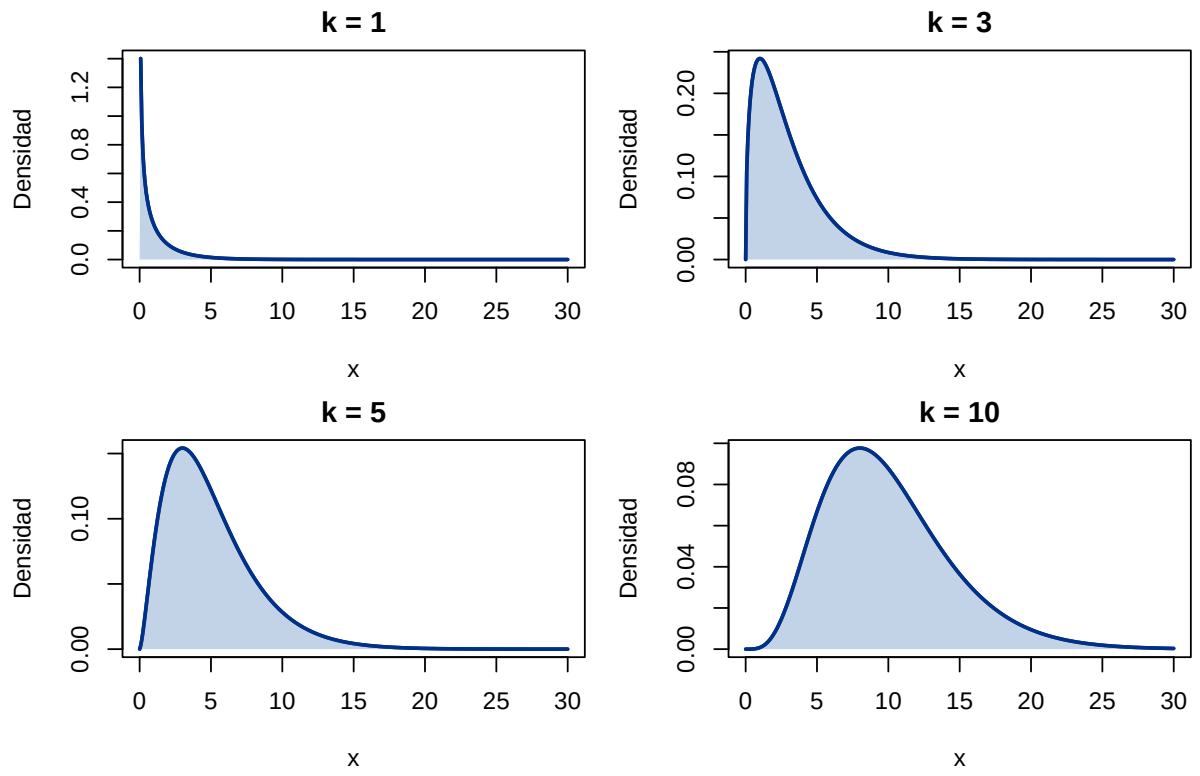


Figura 4.2: c("k = 1", "k = 3", "k = 5", "k = 10")

Densidad de la distribución chi-cuadrado para diferentes grados de libertad. A medida que k aumenta, la distribución se vuelve más simétrica y se aproxima a una normal.

4.10.2. Distribución t de Student t_k

Definición: Distribución t de Student

Si $Z \sim N(0,1)$ e $Y \sim \chi^2_k$ son independientes, entonces:

$$T = \frac{Z}{\sqrt{Y/k}} \sim t_k$$

Parámetro: k (grados de libertad)
Propiedades: - Simétrica alrededor de 0 - $E(T) = 0$ (para $k > 1$) - $\text{Var}(T) = \frac{k}{k-2}$ (para $k > 2$) - Colas más pesadas que la normal - Cuando $k \to \infty$, converge a $N(0,1)$
En R: `dt(x, df=k)`, `pt(x, df=k)`, `rt(n, df=k)`
Usos: Intervalos de confianza para la media (varianza desconocida), pruebas t de Student.

Cuadro 4.31 Código R

```r
k_values <- c(1, 3, 5, 10)
x <- seq(-5, 5, length.out = 500)

par(mfrow = c(2, 2))

for (k in k_values) {
  y <- dt(x, df = k)

  plot(
    x, y,
    type = "l",
    lwd = 2,
    col = "steelblue",
    main = paste("k =", k),
    xlab = "x",
    ylab = "Densidad"
  )

  curve(
    dnorm(x),
    add = TRUE,
    col = "red",
    lty = 2,
    lwd = 2
  )

  legend(
    "topright",
    legend = c("t de Student", "Normal estándar"),
    col = c("steelblue", "red"),
    lty = c(1, 2),
    bty = "n"
  )
}
```

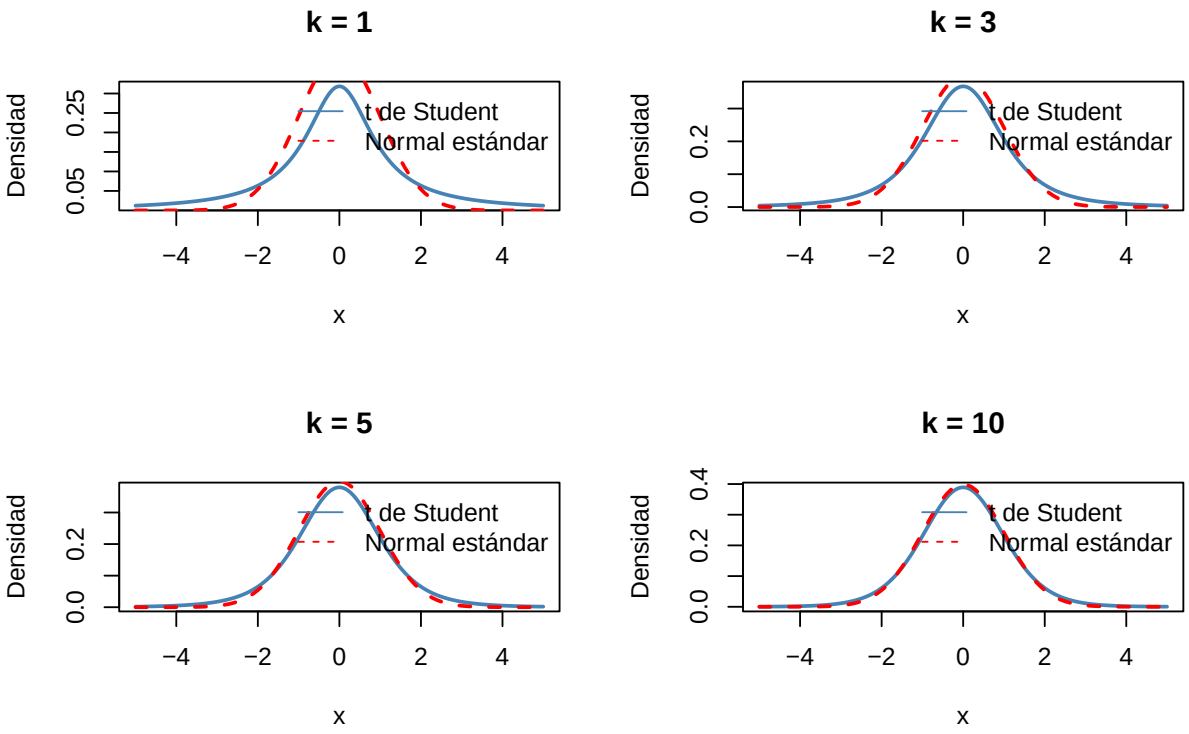


Figura 4.3: k = 1

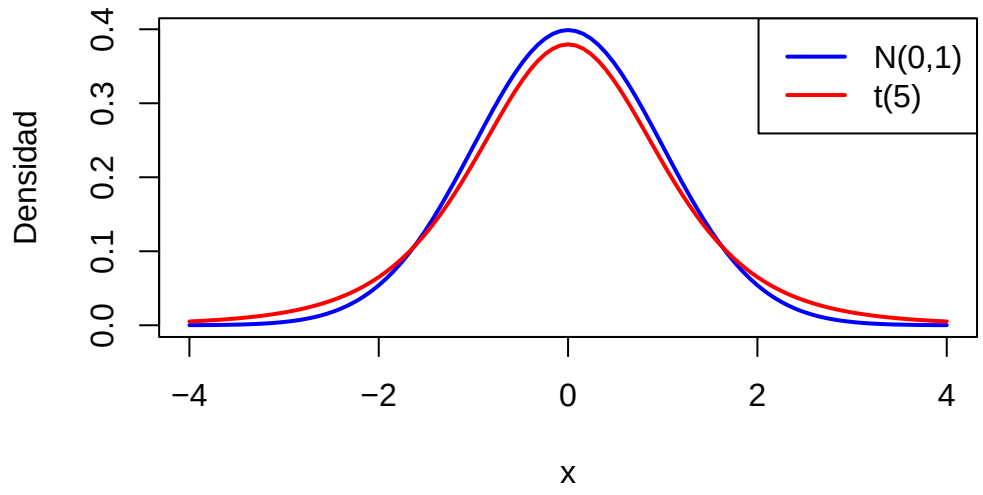
Densidad de la distribución t de Student para diferentes grados de libertad. A medida que k aumenta, la distribución se vuelve más similar a una normal estándar, aunque siempre tiene colas más pesadas.

Tip

La distribución t tiene colas más pesadas, reflejando mayor incertidumbre con muestras pequeñas. ## Comparación: Normal vs t de Student
Para $k = 5$ grados de libertad:

Cuadro 4.32 Código R

```
# Graficar ambas distribuciones
curve(dnorm(x, mean=0, sd=1), from=-4, to=4, col="blue", lwd=2,
      main="N(0,1) vs t con 5 gl", xlab="x", ylab="Densidad")
curve(dt(x, df=5), from=-4, to=4, col="red", lwd=2, add=TRUE)
legend("topright", legend=c("N(0,1)", "t(5)"),
       col=c("blue", "red"), lwd=2)
```



4.10.3. Distribución F de Snedecor F_{k_1,k_2}

Definición: Distribución F

Si $Y_1 \sim \chi^2_{k_1}$ e $Y_2 \sim \chi^2_{k_2}$ son independientes, entonces:

$$F = \frac{Y_1/k_1}{Y_2/k_2} \sim F_{k_1,k_2}$$

Parámetros: k_1, k_2 (grados de libertad del numerador y denominador)
Propiedades: - Solo toma valores positivos - $E(F) = \frac{k_2}{k_2-2}$ (para $k_2 > 2$) - Asimétrica hacia la derecha
En R: `df(x, df1=k1, df2=k2)`, `pf(x, df1=k1, df2=k2)`, `rf(n, df1=k1, df2=k2)`
Usos: Análisis de varianza (ANOVA), pruebas de igualdad de varianzas, análisis de regresión.

Cuadro 4.33 Código R

```r
params <- list(
  c(1, 1),
  c(5, 2),
  c(10, 10),
  c(20, 30)
)

x <- seq(0.001, 5, length.out = 500)

par(mfrow = c(2, 2))

for (p in params) {

  d1 <- p[1]
  d2 <- p[2]

  y <- df(x, df1 = d1, df2 = d2)

  plot(
    x, y,
    type = "l",
    lwd = 2,
    col = "blue",
    main = paste("d1 =", d1, ", d2 =", d2),
    xlab = "x",
    ylab = "Densidad"
  )
}
```

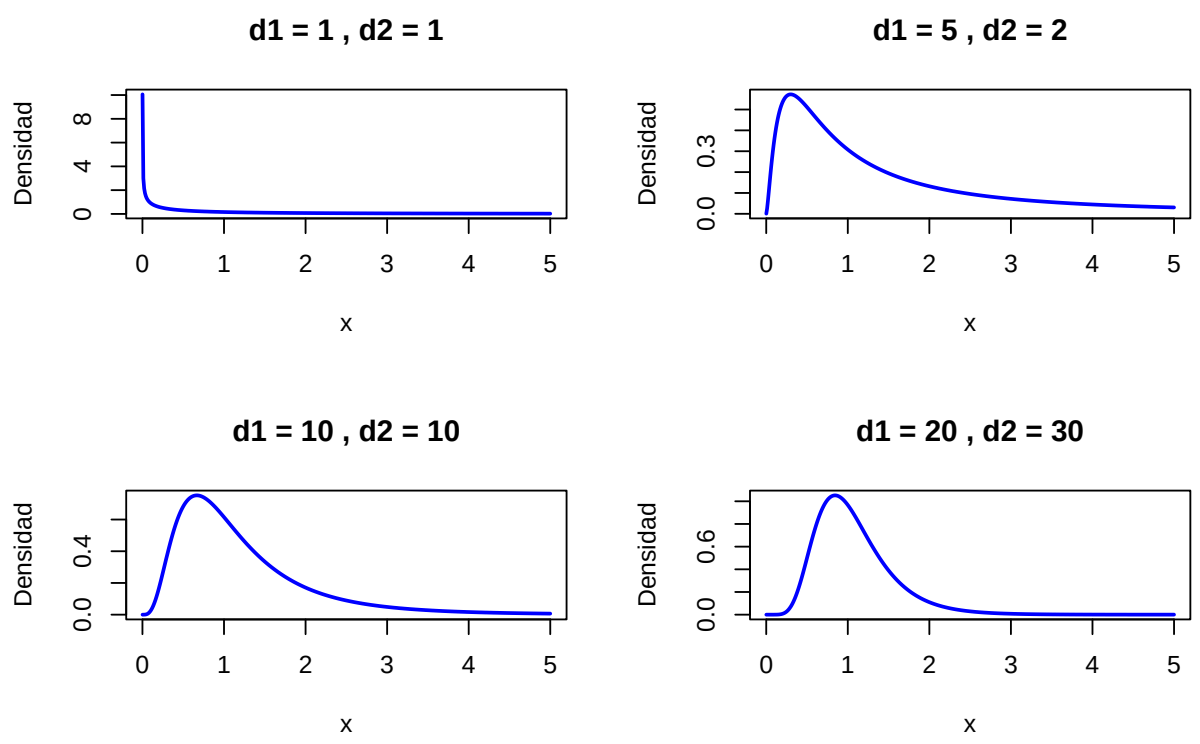


Figura 4.4: d1 = 1, d2 = 1

Densidad de la distribución F de Snedecor para diferentes combinaciones de grados de libertad. La distribución F es asimétrica hacia la derecha y solo toma valores positivos.

4.11. Teorema Central del Límite

Teorema Central del Límite

Enunciado: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuidas (iid) con $E(X_i) = \mu$ y $\text{Var}(X_i) = \sigma^2 < \infty$, entonces:

$$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z)$$

donde Φ es la CDF de la distribución normal estándar.

Consecuencias prácticas:

1. La media muestral es aproximadamente normal para n grande:

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

2. La suma de variables es aproximadamente normal:

$$S_n = \sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2)$$

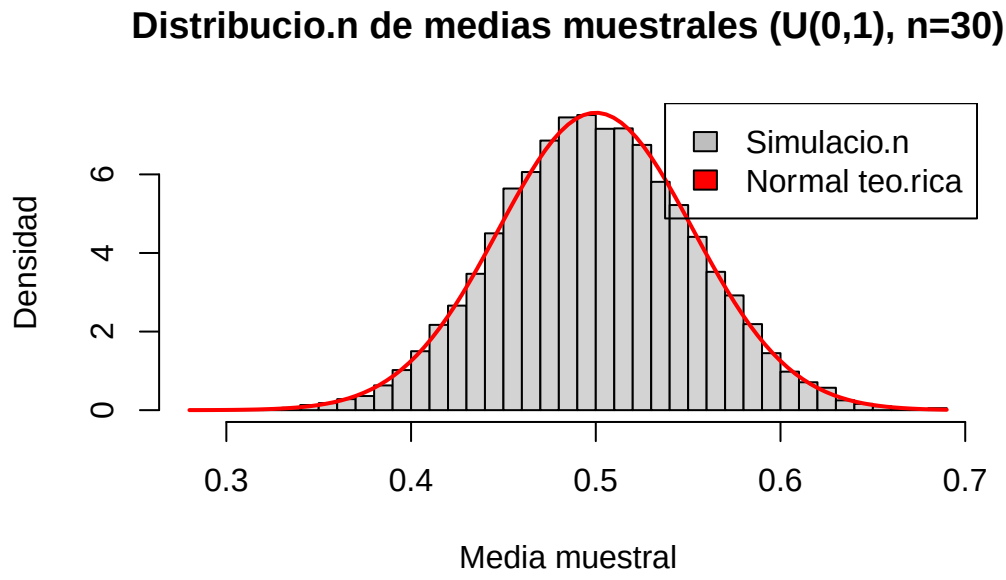
3. Vale **sin importar la distribución original** de las X_i (si es discreta, continua, etc.)

Ejemplo: Simulación del TCL

Demostración del TCL con datos uniformes:

Cuadro 4.34 Código R

```
set.seed(123)

# Generar medias muestrales a partir de muestras de una U(0,1)
n_simulaciones <- 10000
tamano_muestra <- 30

medias <- replicate(n_simulaciones, mean(runif(tamano_muestra, 0, 1)))

# Graficar histograma con curva normal teórica
hist(medias, breaks=40, probability=TRUE,
     main="Distribución de medias muestrales (U(0,1), n=30)",
     xlab="Media muestral", ylab="Densidad")

# Añadir curva normal teórica: N(0.5, 1/(12*30))
mu_teorica <- 0.5
sigma_teorica <- sqrt(1/(12*tamano_muestra))
curve(dnorm(x, mean=mu_teorica, sd=sigma_teorica),
      add=TRUE, col="red", lwd=2)

legend("topright", legend=c("Simulación", "Normal teórica"),
       fill=c("gray", "red"))
```



Aunque los datos originales son uniformes, las medias muestrales siguen una distribución aproximadamente normal.

4.12. Aproximaciones entre distribuciones vistas como aplicaciones del Teorema Central del Lîmite

4.12.1. Binomial → Normal

Ya la vimos anteriormente. Sin embargo, ahora la podemos ver como una de las aplicaciones más conocidas del Teorema Central del Lîmite. Cuando el número de ensayos (n) es suficientemente grande, y la probabilidad de éxito (p) no es extremadamente cercana a 0 o 1, la forma de la distribución Binomial se asemeja a la de una campana, caracterîstica de la distribución Normal.

Condiciones para la aproximación:
La aproximación es generalmente considerada adecuada si se cumplen las siguientes condiciones:

- $np \geq 5$ (o $np \geq 10$ para una mejor aproximación)
- $n(1-p) \geq 5$ (o $n(1-p) \geq 10$ para una mejor aproximación)

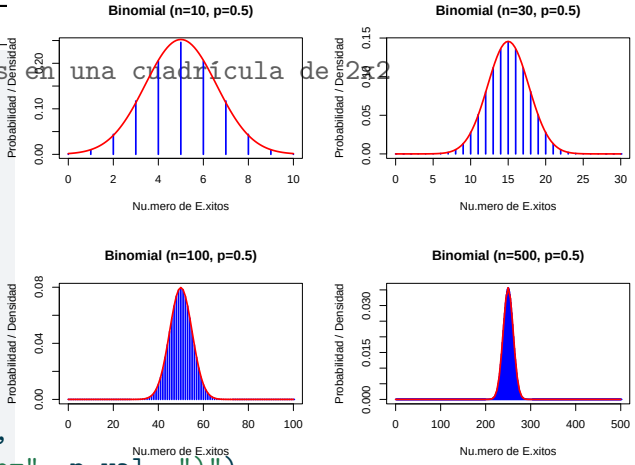
Bajo estas condiciones, si $X \sim B(n,p)$, entonces X puede ser aproximada por $Y \sim N(\mu = np, \sigma^2 = np(1-p))$.

Visualización:

Vamos a visualizar cómo la distribución Binomial se aproxima a la Normal a medida que n aumenta. Consideremos una probabilidad de éxito $p = 0.5$ (simétrica) y variemos n.

Cuadro 4.35 Código R

```
par(mfrow=c(2, 2)) # Organizar 4 gráficos en una cuadrícula de 2x2

p_val <- 0.5 # Probabilidad de éxito

# Caso 1: n pequeña (n=10)
n1 <- 10
x1 <- 0:n1
y1 <- dbinom(x1, size=n1, prob=p_val)
mu1 <- n1 * p_val
sigma1 <- sqrt(n1 * p_val * (1 - p_val))
plot(x1, y1, type="h", lwd=2, col="blue",
     main=paste0("Binomial (n=", n1, ", p=", p_val, ")"),
     xlab="Número de Exitos", ylab="Probabilidad / Densidad")
curve(dnorm(x, mean=mu1, sd=sigma1), add=TRUE, col="red", lwd=2)

# Caso 2: n mediana (n=30)
n2 <- 30
x2 <- 0:n2
y2 <- dbinom(x2, size=n2, prob=p_val)
mu2 <- n2 * p_val
sigma2 <- sqrt(n2 * p_val * (1 - p_val))
plot(x2, y2, type="h", lwd=2, col="blue",
     main=paste0("Binomial (n=", n2, ", p=", p_val, ")"),
     xlab="Número de Exitos", ylab="Probabilidad / Densidad")
curve(dnorm(x, mean=mu2, sd=sigma2), add=TRUE, col="red", lwd=2)

# Caso 3: n grande (n=100)
n3 <- 100
x3 <- 0:n3
y3 <- dbinom(x3, size=n3, prob=p_val)
mu3 <- n3 * p_val
sigma3 <- sqrt(n3 * p_val * (1 - p_val))
plot(x3, y3, type="h", lwd=2, col="blue",
     main=paste0("Binomial (n=", n3, ", p=", p_val, ")"),
     xlab="Número de Exitos", ylab="Probabilidad / Densidad")
curve(dnorm(x, mean=mu3, sd=sigma3), add=TRUE, col="red", lwd=2)

# Caso 4: n muy grande (n=500)
n4 <- 500
x4 <- 0:n4
y4 <- dbinom(x4, size=n4, prob=p_val)
mu4 <- n4 * p_val
sigma4 <- sqrt(n4 * p_val * (1 - p_val))
plot(x4, y4, type="h", lwd=2, col="blue",
     main=paste0("Binomial (n=", n4, ", p=", p_val, ")"),
     xlab="Número de Exitos", ylab="Probabilidad / Densidad")
curve(dnorm(x, mean=mu4, sd=sigma4), add=TRUE, col="red", lwd=2)
```



(a) Aproximación de la distribución Binomial a la Normal para diferentes valores de n. A medida que n aumenta, la forma de la binomial (barras) se asemeja más a la curva de densidad Normal (línea roja).

Cuadro 4.36 Código R

```
par(mfrow=c(1,1)) # Reset layout
```

4.12.2. Poisson → Normal

Similar a la Binomial, la distribución de Poisson también puede ser aproximada por la distribución Normal cuando su parámetro de tasa (λ) es suficientemente grande. Esto se debe a que la distribución de Poisson puede verse como el límite de una Binomial cuando $n \to \infty$ y $p \to 0$ tal que $np = \lambda$. Por lo tanto, el Teorema Central del Límite también juega un papel aquí.

Condiciones para la aproximación: La aproximación es generalmente aceptable si:

- $\lambda \geq 10$ (o $\lambda \geq 5$ en algunos contextos, pero $\lambda \geq 10$ es más seguro para una buena aproximación).

Bajo esta condición, si $X \sim \text{Po}(\lambda)$, entonces X puede ser aproximada por $Y \sim N(\mu = \lambda, \sigma^2 = \lambda)$.

Ejemplo y Visualización:

Vamos a observar cómo la distribución de Poisson se asemeja a la Normal a medida que λ aumenta.

4.12.3. Exponencial Poisson (dualidad tiempo-conteo)

La relación entre las distribuciones Exponencial y de Poisson es una de las dualidades más fundamentales en la teoría de procesos estocásticos, particularmente en los Procesos de Poisson. Un Proceso de Poisson describe el número de eventos que ocurren en un intervalo de tiempo fijo (con distribución de Poisson), mientras que la distribución Exponencial describe el tiempo entre eventos consecutivos (conocido como "tiempo de inter-llegada") en un proceso de Poisson.

Conceptualización de la dualidad: - Si el número de eventos en un intervalo de tiempo t sigue una distribución de Poisson con tasa λ, es decir, $N(t) \sim \text{Po}(\lambda t)$, entonces:

- El tiempo hasta que ocurre el primer evento (o el tiempo entre dos eventos consecutivos) sigue una distribución Exponencial con parámetro de tasa λ.

Esta dualidad es crucial para modelar fenómenos como el tiempo de espera hasta el próximo cliente en una cola, el tiempo de vida de componentes electrónicos, o el tiempo entre mutaciones genéticas.

Ejemplo y Visualización:

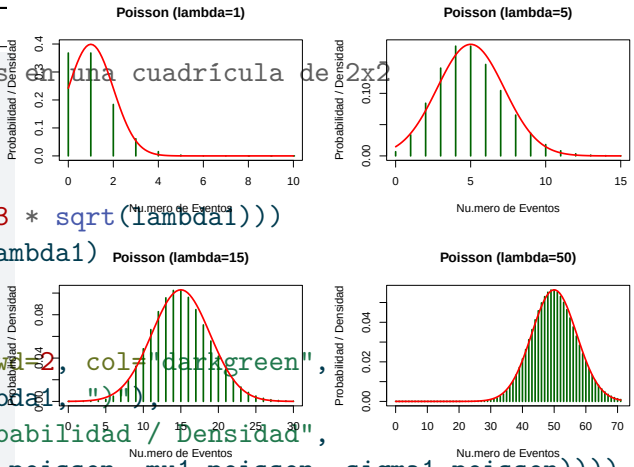
Vamos a simular eventos de un Proceso de Poisson y luego examinar la distribución de los tiempos entre estos eventos para ver cómo se ajusta a una distribución Exponencial.

Cuadro 4.39 Código R

```r
set.seed(42) # Para reproducibilidad

lambda_rate <- 2 # Tasa de eventos por unidad de tiempo
max_time <- 10  # Tiempo total de observación

# Simular un proceso de Poisson: número de eventos en max_time
num_events_poisson <- rpois(1, lambda = lambda_rate * max_time)
cat("Número simulado de eventos en", max_time, "unidades de tiempo:", num_events_poisson, "\n")
```

Cuadro 4.37 Código R

```
par(mfrow=c(2, 2)) # Organizar 4 gráficos en una cuadrícula de 2x2

# Caso 1: lambda pequeña (lambda=1)
lambda1 <- 1
x1_poisson <- 0:max(10, round(lambda1 + 3 * sqrt(lambda1)))
y1_poisson <- dpois(x1_poisson, lambda=lambda1)
mu1_poisson <- lambda1
sigma1_poisson <- sqrt(lambda1)
plot(x1_poisson, y1_poisson, type="h", lwd=2, col="darkgreen",
     main=paste0("Poisson (lambda=", lambda1, ")"),
     xlab="Número de Eventos", ylab="Probabilidad / Densidad",
     ylim=c(0, max(y1_poisson, dnorm(mu1_poisson, mu1_poisson, sigma1_poisson))))
curve(dnorm(x, mean=mu1_poisson, sd=sigma1_poisson),
      from=min(x1_poisson), to=max(x1_poisson), add=TRUE, col="red", lwd=2)


# Caso 2: lambda mediana (lambda=5)
lambda2 <- 5
x2_poisson <- 0:max(15, round(lambda2 + 3 * sqrt(lambda2)))
y2_poisson <- dpois(x2_poisson, lambda=lambda2)
mu2_poisson <- lambda2
sigma2_poisson <- sqrt(lambda2)
plot(x2_poisson, y2_poisson, type="h", lwd=2, col="darkgreen",
     main=paste0("Poisson (lambda=", lambda2, ")"),
     xlab="Número de Eventos", ylab="Probabilidad / Densidad",
     ylim=c(0, max(y2_poisson, dnorm(mu2_poisson, mu2_poisson, sigma2_poisson))))
curve(dnorm(x, mean=mu2_poisson, sd=sigma2_poisson),
      from=min(x2_poisson), to=max(x2_poisson), add=TRUE, col="red", lwd=2)

# Caso 3: lambda grande (lambda=15)
lambda3 <- 15
x3_poisson <- 0:max(30, round(lambda3 + 3 * sqrt(lambda3)))
y3_poisson <- dpois(x3_poisson, lambda=lambda3)
mu3_poisson <- lambda3
sigma3_poisson <- sqrt(lambda3)
plot(x3_poisson, y3_poisson, type="h", lwd=2, col="darkgreen",
     main=paste0("Poisson (lambda=", lambda3, ")"),
     xlab="Número de Eventos", ylab="Probabilidad / Densidad",
     ylim=c(0, max(y3_poisson, dnorm(mu3_poisson, mu3_poisson, sigma3_poisson))))
curve(dnorm(x, mean=mu3_poisson, sd=sigma3_poisson),
      from=min(x3_poisson), to=max(x3_poisson), add=TRUE, col="red", lwd=2)

# Caso 4: lambda muy grande (lambda=50)
lambda4 <- 50
x4_poisson <- 0:max(70, round(lambda4 + 3 * sqrt(lambda4)))
y4_poisson <- dpois(x4_poisson, lambda=lambda4)
mu4_poisson <- lambda4
sigma4_poisson <- sqrt(lambda4)
plot(x4_poisson, y4_poisson, type="h", lwd=2, col="darkgreen",
     main=paste0("Poisson (lambda=", lambda4, ")"),
     xlab="Número de Eventos", ylab="Probabilidad / Densidad",
```



(a) Aproximación de la distribución de Poisson a la Normal para diferentes valores de lambda. A medida que lambda aumenta, la forma de la Poisson (barras) se asemeja más a la curva de densidad Normal (línea roja).

```
Número simulado de eventos en 10 unidades de tiempo: 26
```

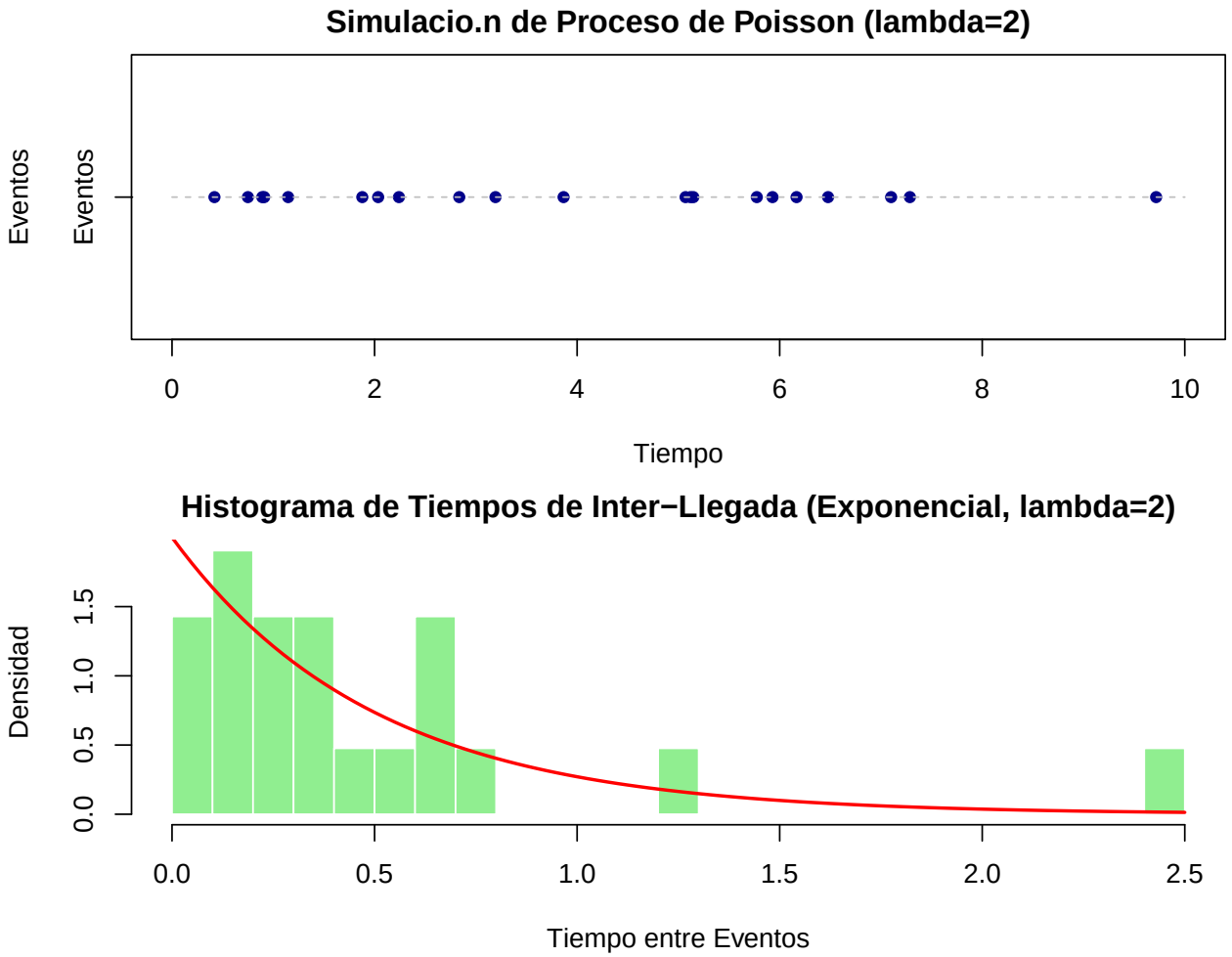


Figura 4.7: Dualidad entre la distribución de Poisson y la Exponencial. La parte superior muestra la simulación de un proceso de Poisson con eventos marcados. La parte inferior es el histograma de los tiempos entre eventos, mostrando un ajuste a la distribución Exponencial.

4.13. Tabla de Referencia de Distribuciones

Distribución	Parámetros	$E(X)$	$\text{Var}(X)$	Función R
Discretas				
Uniforme	n	$\frac{1}{n}\sum x_i$	$\frac{1}{n}\sum(x_i-\mu)^2$	`sample()`
Bernoulli	p	p	$p(1-p)$	`dbinom(,size=1)`
Binomial	n, p	np	$np(1-p)$	`dbinom()`
Poisson	λ	λ	λ	`dpois()`
Continuas				

Distribución	Parámetros	$E(X)$	$\text{Var}(X)$	Función R
Uniforme	a, b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	`dunif()`
Exponencial	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	`dexp()`
Normal	μ, σ	μ	σ^2	`dnorm()`
Muestreo				
Chi-cuadrado	k	k	$2k$	`dchisq()`
t	k	0	$\frac{k}{k-2}$	`dt()`
F	k_1, k_2	$\frac{k_2}{k_2-2}$	—	`df()`

4.14. Resumen de Conceptos Clave

Conceptos Fundamentales

1. **Variables Aleatorias:** Funciones que asignan números a resultados de experimentos aleatorios.
2. **Esperanza:** Media ponderada; resume la localización.
3. **Varianza:** Dispersión alrededor de la media; mide variabilidad.
4. **Estandarización:** Transforma variables para comparación directa (z-score).
5. **Distribuciones Discretas:** Binomial (conteos en n ensayos), Poisson (eventos raros).
6. **Distribuciones Continuas:** Normal (datos naturales), Exponencial (tiempos de espera), Uniforme (selección aleatoria).
7. **Distribuciones para Muestreo:** Chi-cuadrado, t, F (inferenciales, no descriptivas).
8. **Teorema Central del Límite:** Las medias muestrales son normales, sin importar la distribución original.

4.15. Ejercicios

4.15.1. Ejercicio 1: Distribución Binomial

Un ensayo clínico prueba un tratamiento con probabilidad de éxito $p = 0.65$ en 50 pacientes.

a) ¿Cuál es la probabilidad de que exactamente 35 pacientes se curen?
b) ¿Cuál es la probabilidad de que al menos 40 se curen?
c) Calcula la esperanza y varianza del número de curaciones.

4.15.2. Ejercicio 2: Distribución Normal

El colesterol LDL en una población sigue $N(130, 40)$ mg/dL.

a) ¿Qué proporción de personas tienen LDL > 160 mg/dL (riesgo alto)?
b) ¿Cuál es el percentil 90 del LDL?

c) Un paciente tiene z-score = 1.5. ¿Cuál es su valor de LDL?

4.15.3. Ejercicio 3: Distribución de Poisson

El número de infecciones nosocomiales en una clínica sigue una Poisson con $\lambda = 2.5$ por mes.

a) ¿Cuál es la probabilidad de que no haya infecciones en un mes?
b) ¿Cuál es la probabilidad de que haya más de 5 infecciones?
c) ¿Cuál es el número esperado de infecciones en 6 meses?

4.15.4. Ejercicio 4: Teorema Central del Límite

Se muestran 100 recién nacidos de bajo peso al nacer y prematuridad con una altura media 45 cm y una desviación estándar 8 cm.

a) ¿Cuál es la distribución aproximada de la media muestral $\bar{X}$?
b) ¿Cual es $P(\bar{X} \geq 46)$?
c) ¿Qué tamaño de muestra se necesita para que el error estándar sea menor de 0.5 cm?

4.15.5. Ejercicio 5: Transformación Z

Las medidas de glucosa en sangre siguen $N(100, 100)$ mg/dL.

a) Estandariza los valores 90, 100 y 120 mg/dL.
b) ¿Qué valor original corresponde a z = -1.96?
c) Interpreta: "un paciente tiene z-score = 2.5".

4.16. Respuestas a los Ejercicios

Ejercicio 1: Distribución Binomial
- a) P(X=35) = C(50,35)(0.65)³ (0.35)¹ 0.0622
- b) P(X 40) = Σ P(X=k) para k=40..50 0.0339
- c) E[X] = np = 50×0.65 = 32.5; Var[X] = np(1-p) = 50×0.65×0.35 = 11.375

Ejercicio 2: Distribución Normal N(130,40)
- a) Z = (160-130)/√40 = 4.74; P(LDL>160) 0.0001 (muy raro)
- b) P : z=1.28, LDL = 130 + 1.28√40 130 + 8.1 = 138.1 mg/dL
- c) LDL = 130 + 1.5√40 130 + 9.49 = 139.49 mg/dL

Ejercicio 3: Distribución Poisson (=2.5)
- a) P(X=0) = e^(-2.5) 0.0821
- b) P(X>5) = 1 - P(X 5) 0.0420
- c) En 6 meses: _total = 2.5×6 = 15; E[X] = 15

Ejercicio 4: Teorema Central del Límite
- a) $\bar{X} \sim N(45, \sigma^2_{\bar{X}})$ donde $\sigma_{\bar{X}} = 8/\sqrt{100} = 0.8$ cm

- b) Z = (46-45)/0.8 = 1.25; P(46) 0.1056
- c) SE < 0.5 8/√n < 0.5 n > 256

Ejercicio 5: Estandarización
- a) z(90) = (90-100)/10 = -1; z(100) = 0; z(120) = 2
- b) x = 100 + (-1.96)(10) = 80.4 mg/dL
- c) El paciente está 2.5 desviaciones estándar por encima de la media (valor muy alto, atîpico)

4.17. Lecturas Complementarias

- **Forbes et al. (2011):** Distribuciones de probabilidad continuadas y sus propiedades (referencia técnica).
- **Tablas Normales:** Los z-scores están tabulados en libros de estadîstica y disponibles en R mediante `qnorm()`.
- **Relaciones Distribucionales:** Saad, F. A., Freer, C. E., Rinard, M. C., & Mansinghka, V. K. (2020).** Optimal Approximate Sampling from Discrete Probability Distributions. *Proceedings of the ACM on Programming Languages*, 4(POPL), Article 36. DOI: 10.1145/3371104. Este trabajo discute la exactitud y los trade-offs entre métodos de muestreo exactos y aproximados para distribuciones discretas, un concepto extensible a la precisión de cálculos de probabilidad.

Métodos Avanzados

Para ampliar los contenidos de este capîtulo con técnicas estadîsticas avanzadas, visita:
→ **Bioestadîstica Avanzada — M.A. Luque Fernández**

Cuadro 4.40 Código R

```
# Simular los tiempos de inter-llegada (distribución Exponencial)
inter_arrival_times <- rexp(num_events_poisson + 1, rate = lambda_rate)
# Asegurarse de que la suma de tiempos no exceda max_time
arrival_times <- cumsum(inter_arrival_times)
arrival_times <- arrival_times[arrival_times <= max_time]

# Eliminar el último tiempo si excede el max_time para el histograma
if (length(arrival_times) > 0 && sum(inter_arrival_times[1:length(arrival_times)]) > max_time) {
  inter_arrival_times_for_hist <- inter_arrival_times[1:length(arrival_times) - 1]
} else {
  inter_arrival_times_for_hist <- inter_arrival_times[1:length(arrival_times)]
}
inter_arrival_times_for_hist <- inter_arrival_times_for_hist[inter_arrival_times_for_hist > 0]

# Preparar el gráfico
par(mfrow=c(2, 1), mar=c(4, 4, 2, 2) + 0.1)

# Gráfico superior: Proceso de Poisson (eventos en el tiempo)
plot(arrival_times, rep(1, length(arrival_times)), pch=16, col="darkblue",
     xlab="Tiempo", ylab="Eventos",
     main=paste0("Simulación de Proceso de Poisson (lambda=", lambda_rate, ")"),
     xlim=c(0, max_time), ylim=c(0.5, 1.5), yaxt='n')
axis(side=2, at=1, labels="Eventos")
segments(0, 1, max_time, 1, col="gray", lty=2)

# Gráfico inferior: Histograma de tiempos de inter-llegada (Exponencial)
if (length(inter_arrival_times_for_hist) > 1) {
  hist(inter_arrival_times_for_hist, breaks=20, probability=TRUE,
       main=paste0("Histograma de Tiempos de Inter-Llegada (Exponencial, lambda=", lambda_rate
       xlab="Tiempo entre Eventos", ylab="Densidad",
       col="lightgreen", border="white")
  curve(dexp(x, rate = lambda_rate), add=TRUE, col="red", lwd=2)
} else {
  plot(1, type="n", main="No hay suficientes eventos para histograma de inter-llegada",
       xlab="", ylab="", xaxt='n', yaxt='n')
  text(1,1, "Aumenta max_time o lambda_rate para más eventos.")
}
```

Cuadro 4.41 Código R

```
par(mfrow=c(1,1)) # Reset layout
```

Capítulo 5

Semana 5 — Muestreo y Distribuciones Muestrales

En esta semana estudiamos los fundamentos del muestreo y las propiedades estadísticas de los estimadores. Los conceptos que aprenderemos son la base de toda inferencia estadística: cómo usar datos de una muestra para hacer afirmaciones sobre la población completa.

5.1. Conceptos Fundamentales de Muestreo

5.1.1. ¿Qué es el Muestreo?

> **Definición: Muestreo**
>
> El **muestreo** es el proceso de seleccionar una parte de la población para observar y estudiar, de manera que podamos hacer inferencias sobre características de interés en toda la población.

Imagine que queremos estimar el gasto en alimentos de las familias en España. Tenemos dos opciones:

- **Censo**: Encuestar a TODAS las familias del país (costoso, largo, casi imposible)
- **Muestra**: Encuestar al 5 % de las familias e inferir sobre toda la población (barato, rápido, práctico)

5.1.1.1. Notación Fundamental

Símbolo	Significado
N	Tamaño de la población
n	Tamaño de la muestra
$f = n/N$	Fracción muestral
X_i	i-ésima observación (variable aleatoria)
x_i	i-ésimo valor observado (número)

5.1.2. Parámetros Poblacionales

Los parámetros son características desconocidas de la población que queremos estimar:

⚠ Parámetros de Interés

- **Media poblacional**: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- **Varianza poblacional**: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- **Proporción poblacional** (variables binarias): $\pi = \frac{1}{N}\sum_{i=1}^{N} x_i$, donde $x_i \in \{0, 1\}$

5.1.3. Parámetros vs. Estadísticos

ℹ Definiciones Clave

Un **parámetro** es una característica desconocida de la población (p.ej., μ, σ^2, π).
Un **estadístico** es una función de la muestra calculada a partir de datos observados. Los estadísticos se usan para estimar parametros desconocidos.

5.1.4. El Problema del Sesgo de Muestreo: El Estudio Nurses' Health (NHS) y la Terapia Hormonal Sustitutiva

La importancia de un muestreo y un diseño adecuados se ilustra con uno de los casos más comentados en la epidemiología moderna: la aparente discrepancia entre los estudios observacionales y los ensayos clínicos sobre la **terapia hormonal sustitutiva** (THS) en mujeres postmenopáusicas y el riesgo cardiovascular.

⚠ Advertencia: El sesgo de selección en el Nurses' Health Study (NHS)

Durante los años 80 y 90, el **Nurses' Health Study** (Universidad de Harvard) — un estudio observacional prospectivo de gran tamaño (n 120 000 enfermeras) — concluyó que las mujeres que tomaban **terapia hormonal sustitutiva** presentaban una reducción **40–50 %** del riesgo de enfermedad coronaria. Sobre la base de estos resultados, durante años se prescribió la THS de forma sistemática como **cardioprotección primaria**.
Sin embargo, el ensayo aleatorizado **Women's Health Initiative (WHI)** (publicado en *JAMA*, 2002, con n 16 000 mujeres asignadas al azar a THS o placebo) mostró exactamente lo contrario: un **incremento** del riesgo cardiovascular, de cáncer de mama y de eventos tromboembólicos. El ensayo se interrumpió prematuramente por razones de seguridad.
¿Por qué dos estudios sobre la misma intervención llegan a conclusiones opuestas?

- En el NHS, las mujeres que **se autoseleccionaban** para tomar THS eran sistemáticamente **más sanas, más delgadas, con mejor nivel socioeconómico, no fumadoras y con mejor acceso a controles médicos** que las que no la tomaban.
- Este **sesgo de selección** (también llamado *healthy user bias*) confundía el efecto del

tratamiento con el efecto del perfil basal de las pacientes.

- En el WHI, en cambio, la **aleatorización** equilibró todas las características basales (medidas y no medidas) entre los grupos. Aquí sí podía atribuirse causalmente la diferencia observada a la intervención.

Tamaños muestrales:

Estudio	Diseño	n	Conclusión sobre THS y riesgo cardiovascular
NHS (1985–2000)	Observacional, autoseleccionado	120 000	**−40 %** (protector)
WHI (1993–2002)	Aleatorizado, doble ciego	16 000	**+29 %** (perjudicial)

Lección epidemiológica: Un tamaño muestral 8 veces mayor **no compensa** un mal diseño de muestreo. El sesgo de selección puede invertir completamente la conclusión clínica, con consecuencias para millones de pacientes. La aleatorización sigue siendo el estándar de oro precisamente porque elimina la confusión por características basales medidas y no medidas— entre los grupos. Este episodio motivó el desarrollo de los modernos métodos de **inferencia causal** (Hernán & Robins, *Causal Inference: What If*, 2020) para tratar de "emular" un ensayo a partir de datos observacionales.

5.2. Tipos de Muestreo

Para obtener muestras representativas y hacer inferencias válidas, existen varios diseños de muestreo:

5.2.1. Muestreo Aleatorio Simple

i Definición

En el **muestreo aleatorio simple**, cada elemento de la población tiene la misma probabilidad de ser seleccionado.

Distinguimos dos variantes:

1. Muestreo sin reemplazo: - Cada unidad puede ser seleccionada como máximo una vez - Los elementos muestreados NO se devuelven a la población - Los valores muestrales NO son independientes - Aplicable a poblaciones finitas

2. Muestreo con reemplazo: - Cada unidad puede ser seleccionada más de una vez - Los elementos se devuelven a la población después de cada extracción - Los valores muestrales SON independientes - Matemáticamente equivalente a una población infinita

5.2.2. Muestreo Estratificado

> **Definición**
>
> En el **muestreo estratificado**, la población se divide en subgrupos (estratos) no solapados, y se extrae una muestra aleatoria simple de cada estrato.

Ejemplo: Para estudiar opiniones políticas, estratificar por región, edad, o nivel educativo asegura que cada grupo esté representado adecuadamente.

Ventaja: Proporciona estimaciones más precisas que el muestreo simple cuando existe variación dentro de los estratos.

5.2.3. Muestreo por Conglomerados

> **Definición**
>
> En el **muestreo por conglomerados**, la población se divide en conglomerados (clusters), se seleccionan aleatoriamente algunos conglomerados, y se observan TODOS los elementos dentro de esos conglomerados.

Ejemplo: Para encuestar hogares en un país, seleccionar 30 ciudades al azar y encuestar todos los hogares en esas ciudades.

Ventaja: Más barato que muestreo simple cuando los elementos están geográficamente dispersos. **Desventaja**: Menos preciso si hay mucha variación entre conglomerados.

5.2.4. Muestreo por Criterio (Judgment Sampling)

> **Definición**
>
> En el **muestreo por criterio**, los elementos se seleccionan según el juicio de expertos, no aleatoriamente.

Advertencia: Este tipo de muestreo puede introducir sesgo y no permite hacer inferencia estadística formal.

5.3. Estadísticos Muestrales como Variables Aleatorias

5.3.1. El Concepto Clave

Antes de observar una muestra, los valores que obtendremos son aleatorios. Por lo tanto:

- Los valores $X_1, X_2, \ldots, X_n$ son **variables aleatorias**
- Cualquier función de estos valores es también una **variable aleatoria**
- Esto incluye la media muestral $\bar{X}$, la proporción muestral $\hat{\pi}$, y la varianza muestral S^2

5.3.2. Estadísticos Importantes

⚠ Estadísticos Muestrales

- **Media muestral**: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- **Proporción muestral** (variables binarias): $\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- **Varianza muestral (conocida)**: $S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$
- **Varianza muestral (desconocida)**: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- **Varianza muestral (desconocida, alternativa)**: $S'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Cada uno de estos estadísticos tiene su propia **distribución muestral**: la distribución de probabilidad del estadístico a través de todas las posibles muestras de tamaño n.

5.4. Distribución Muestral de la Media

5.4.1. Caso: Muestreo con Reemplazo

Cuando extraemos la muestra con reemplazo (o de una población infinita), los valores muestrales son independientes.

⚠ Propiedades de $\bar{X}$ (con reemplazo)

Para variables aleatorias independientes X_i con $E(X_i) = \mu$ y $\text{Var}(X_i) = \sigma^2$:

$$E(\bar{X}) = \mu$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}} \quad \text{(error estándar)}$$

Interpretación: Mientras mayor sea n, menor será la varianza de la media muestral. Esto explica por qué muestras grandes dan estimaciones más precisas.

5.4.2. Caso: Muestreo sin Reemplazo

Cuando la muestra se extrae sin reemplazo de una población finita, existe un factor de corrección:

⚠ Propiedades de $\bar{X}$ (sin reemplazo)

$$E(\bar{X}) = \mu$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

donde $\frac{N-n}{N-1}$ es el **factor de corrección por población finita** (CPF).

Nota: Cuando $N \gg n$, el CPF ≈ 1 y recuperamos la fórmula del caso con reemplazo.

5.4.3. Distribución Exacta Cuando $X \sim N(\mu, \sigma^2)$

⚠ Resultado Fundamental

Si $X_1, X_2, \ldots, X_n$ son i.i.d. $N(\mu, \sigma^2)$, entonces:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Estandarizando:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Este estadístico Z sigue una **distribución normal estándar**.

5.5. Distribución Muestral de la Proporción

Cuando estudiamos variables binarias (éxito/fracaso, sí/no), la proporción muestral es el estadístico de interés.

5.5.1. Con Reemplazo

Distribución de $\hat{\pi}$ (con reemplazo)

Para una muestra de variables Bernoulli independientes con parámetro π:

$$E(\hat{\pi}) = \pi$$

$$\text{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$$

$$\sigma(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Ejemplo práctico: Si queremos estimar la proporción de votantes que apoyan una política, con $\pi = 0.5$ y $n = 1000$:

$$\sigma(\hat{\pi}) = \sqrt{\frac{0.5 \times 0.5}{1000}} = \sqrt{0.00025} \approx 0.0158$$

Esto significa que la proporción muestral varía tipicamente en $\pm 1.58\,\%$ entre diferentes muestras.

5.5.2. Normalidad Asintótica

Para muestras grandes ($n > 30$ o $n\pi > 5$ y $n(1-\pi) > 5$):

$$Z = \frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} \approx N(0,1)$$

5.6. Distribución Muestral de la Varianza

La varianza muestral también es una variable aleatoria y su distribución depende de si conocemos la media poblacional.

5.6.1. Caso: μ Conocida

Cuando la media poblacional μ es conocida:

Distribucion de S^{*2} (conocida)

Si $X_i \sim N(\mu, \sigma^2)$ y μ es conocida:

$$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$\frac{nS^{*2}}{\sigma^2} \sim \chi^2_n$$

donde χ^2_n es la distribución chi-cuadrado con n grados de libertad.
Propiedades: - $E(S^{*2}) = \sigma^2$ (insesgada) - $\text{Var}(S^{*2}) = \frac{2\sigma^4}{n}$

5.6.2. Caso: μ Desconocida

Cuando estimamos la media con $\bar{X}$, pierden un grado de libertad:

⚠ Distribución de S^2 (desconocida)

Si $X_i \sim N(\mu, \sigma^2)$ y μ es desconocida:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Propiedades: - $E(S^2) = \sigma^2$ (insesgada) - $\text{Var}(S^2) = \frac{2\sigma^4}{n-1}$

Alternativa sesgada:

$$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$$

- $E(S'^2) = \frac{n-1}{n}\sigma^2$ (sesgada, subestima σ^2)

Conclusión: Use S^2 (dividir por $n-1$) para estimaciones insesgadas de σ^2.

5.7. Teorema Central del Límite (Versión Formal)

El Teorema Central del Límite (TCL) es el resultado más importante en estadística. Explica por qué la distribución normal es ubicua.

⚠ Teorema Central del Límite

Sea $X_1, X_2, \ldots$ una secuencia de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) con media μ y varianza σ^2 finitas.
Define $S_n = X_1 + X_2 + \cdots + X_n$. Entonces:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1) \quad \text{cuando} \quad n \to \infty$$

Equivalentemente, para la media muestral:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

5.7.1. Interpretación

Mensaje principal: La suma (y la media) de muchas variables aleatorias ES APROXIMADAMENTE normal, sin importar la distribución original de cada variable.

Requisito único: Que las variables tengan varianza finita. La distribución original puede ser cualquiera: exponencial, uniforme, Bernoulli, etc.

Implicación práctica: Para n suficientemente grande (tîpicamente $n \geq 30$), la media muestral es aproximadamente normal, incluso si los datos originales no son normales. Esto justifica el uso de técnicas basadas en la normalidad para inferencia estadîstica, como intervalos de confianza y pruebas de hipótesis, en una amplia variedad de situaciones.

Advertencia

Notese que si la población es extremadamente asimétrica, se podrîa necesitar $n \geq 60$ o más para que la aproximación normal sea adecuada.

5.8. La Distribución t de Student

En la práctica, casi nunca conocemos σ. Cuando estimamos la desviación tîpica con la muestra, la distribución cambia.

La Distribución t de Student

Si $X_i \sim N(\mu, \sigma^2)$ (normales) y σ^2 es DESCONOCIDA, entonces:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

donde t_{n-1} es la **distribucion t de Student con $n-1$ grados de libertad**.
$S = \sqrt{S^2}$ es la desviación tîpica muestral (raîz cuadrada del estimador insesgado de varianza).

5.8.1. Propiedades de la Distribución t

- **Forma**: Simétrica, con colas más pesadas que la normal estándar $N(0,1)$

- **Media**: $E(T) = 0$ (para $n > 1$)
- **Varianza**: $\text{Var}(T) = \frac{n-1}{n-3}$ (para $n > 3$)
- **Convergencia**: Conforme $n \to \infty$, $t_n \to N(0,1)$
- **Regla práctica**: Para $n > 30$, $t_n \approx N(0,1)$

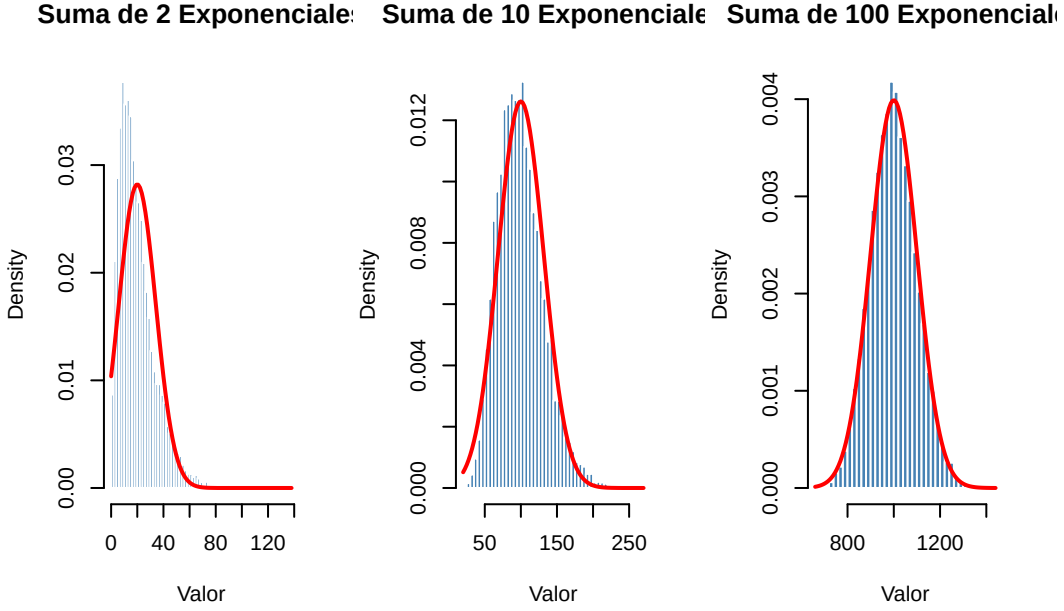
5.8.2. Cuándo Usar Cada Distribución

Situación	Distribución
$X \sim N(\mu, \sigma^2)$, σ conocida	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
$X \sim N(\mu, \sigma^2)$, σ desconocida	$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$
Cualquier distribución, $n > 30$, σ conocida	$Z \approx N(0,1)$ (TCL)
Cualquier distribución, $n > 30$, σ desconocida	$T \approx N(0,1)$ (TCL aproximado)

5.9. Simulaciones del Teorema Central del Límite

Ejemplo 5.1: TCL con Variables Exponenciales

Simularemos el TCL mostrando que la suma de variables exponenciales (muy asimétricas) se aproxima a una distribución normal conforme aumentamos el número de variables.



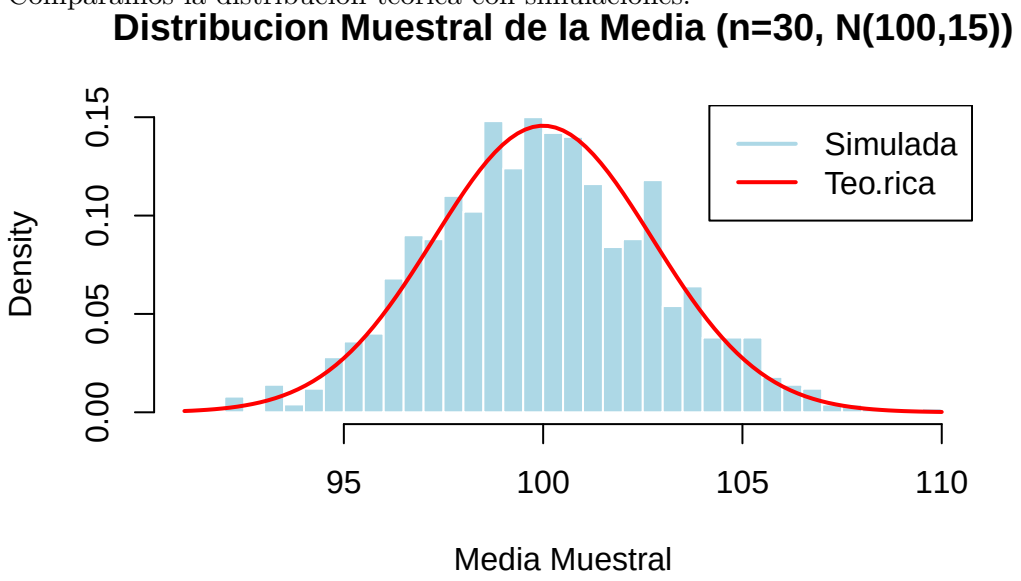
Interpretación

La simulación demuestra empíricamente el Teorema Central del Límite. La suma de dos variables exponenciales (distribución altamente asimétrica hacia la derecha) aún muestra desviación respecto a la normalidad, con solapamiento limitado entre

histograma y curva teórica. Con diez exponenciales sumadas, la concordancia mejora notablemente. Con cien exponenciales, la distribución muestral es prácticamente indistinguible de una curva normal (línea roja), a pesar de que cada variable individual es exponencial pura. Este comportamiento es independiente de la distribución original y requiere únicamente varianza finita, ilustrando por qué la normalidad emerge como distribución universal en estadística aplicada.

Ejemplo 5.2: Distribuciones Muestrales de la Media

Comparamos la distribución teórica con simulaciones.



Interpretación

La simulación de 1000 muestras (n=30 cada una) de una población normal N(100,15) genera medias muestrales distribuidas teóricamente como N(100, 15/√30) = N(100, 2.74). El histograma (barras azules) se superpone casi perfectamente con la curva normal teórica (línea roja), confirmando que el error estándar de la media es /√n = 15/√30 2.74. Esta concordancia exacta entre simulación y teoría valida la predicción analítica de distribuciones muestrales y justifica el uso de la distribución normal para intervalos de confianza e inferencia en muestras grandes.

5.10. Resumen de Distribuciones Muestrales

Tabla Resumen: Estadísticos y sus Distribuciones

Estadístico	Condiciones	$E(\cdot)$	Var$(\cdot)$	Distribución
$\bar{X}$	RS c/ reemplazo, σ conocida	μ	σ^2/n	$N(\mu, \sigma^2/n)$
$\bar{X}$	RS s/ reemplazo, σ conocida	μ	$\sigma^2/n \cdot \frac{N-n}{N-1}$	$N(\mu, \cdots)$
$\bar{X}$	$X \sim N(\mu, \sigma^2)$, σ conocida	μ	σ^2/n	$N(\mu, \sigma^2/n)$
$\bar{X}$	$X \sim N(\mu, \sigma^2)$, σ desconocida	μ	S^2/n	t_{n-1}
$\hat{\pi}$	Variables binarias, n grande	π	$\pi(1-\pi)/n$	$N(\pi, \frac{\pi(1-\pi)}{n})$
S^2	$X \sim N(\mu, \sigma^2)$, μ desconocida	σ^2	$\frac{2\sigma^4}{n-1}$	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

5.11. Corrección por Población Finita

Cuando la población es finita y la fracción muestral $f = n/N$ es apreciable (tipicamente $f > 0.05$), se aplica la **corrección por población finita**:

Factor de Corrección por Población Finita

$$\text{CPF} = \sqrt{\frac{N-n}{N-1}}$$

La varianza del estimador se multiplica por este factor:

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

Casos extremos: - Si $N \to \infty$, entonces CPF $\to 1$ (población infinita, no hay corrección) - Si $n = N$ (censo), entonces CPF $= 0$ (varianza $= 0$, estimación perfecta) - Si $n = N/2$, entonces CPF ≈ 0.71

5.12. Ejercicios

Ejercicio 5.1: Varianza de la Media Muestral

Un hospital desea estimar el peso medio al nacer de los neonatos en una región. De una población de 5000 nacimientos anuales, toma una muestra de 100. Se sabe que la desviación típica poblacional es $\sigma = 200$ gramos.

(a) Calcule la varianza de la media muestral $\bar{X}$.
(b) Calcule el error estándar $\sigma(\bar{X})$.
(c) ¿Cuál sería el error estándar si la muestra fuera de 400 neonatos?
(d) ¿Qué efecto tiene aumentar el tamaño muestral?

Ejercicio 5.2: Corrección por Población Finita

Un centro de salud tiene 8,000 pensionistas a su cargo. Se desea estimar el gasto mensual medio por paciente en tratamientos anticoagulantes y protectores cardiovasculares avanzados. Se planea tomar una muestra de 320 pacientes. Sabiendo que el gasto medio estimado ronda los 65 euros y la desviación típica poblacional es de $\sigma = 45$ euros:

(a) Calcule la fracción muestral f.
(b) ¿Debería aplicarse la corrección por población finita?
(c) Calcule $\text{Var}(\bar{X})$ con y sin corrección.
(d) ¿Cuál es la diferencia porcentual?

Ejercicio 5.3: Distribución de la Proporción

Un candidato jefe de servicio quiere conocer el porcentaje de facultativos que lo apoyan para ser director del hospital. Un sondeo con $n = 600$ facultativos de área encuentra que 324 lo apoyan.

(a) Estime la proporción muestral $\hat{\pi}$.
(b) Calcule $\sigma(\hat{\pi})$ asumiendo que la proporción poblacional verdadera es $\pi = 0.5$ (máxima incertidumbre).
(c) Construya un intervalo del 95 % alrededor de $\hat{\pi}$ usando la normalidad asintótica.
(d) ¿Es compatible el verdadero apoyo de 50 % con el sondeo observado?

Ejercicio 5.4: Aplicación del Teorema Central del Límite

Se sabe que el ingreso mensual de los trabajadores sanitarios de un hospital sigue una distribución exponencial con media $\mu = 2000$ euros y varianza $\sigma^2 = 4.000.000$.

(a) ¿Por qué NO podemos usar directamente una distribución normal para un trabajador individual?
(b) Si tomamos una muestra aleatoria de 100 trabajadores, ¿cual es la distribución aproximada de $\bar{X}$?

(c) ¿Cuál es la probabilidad aproximada de que la media muestral esté entre 1900 y 2100 euros?
(d) ¿Qué tamaño muestral sería necesario para que $P(|\bar{X} - 2000| < 100) \approx 0.95$?

Ejercicio 5.5: Distribución t de Student

Una clínica desea estimar el peso medio de recién nacidos. Una muestra de 16 bebés tiene peso promedio $\bar{x} = 3500$ gramos con desviación típica muestral $s = 400$ gramos. Se asume que los pesos siguen una distribución normal.

(a) Calcule el error estándar estimado $s(\bar{X})$.
(b) ¿Cuál es el valor crítico $t_{0.975,15}$ para un intervalo del 95 %? (Pista: use tablas o R)
(c) Construya un intervalo de confianza del 95 % para el peso medio poblacional.
(d) ¿Por qué usamos distribución t en lugar de normal estándar?

Ejercicio 5.6: Distribución Chi-Cuadrado y Varianza

Se mide el contenido de grasa en leche de 25 muestras. La varianza muestral es $s^2 = 1.44$ (porcentaje al cuadrado). Asuma que el contenido sigue una distribución normal con varianza poblacional $\sigma^2 = 1.0$.

(a) Calcule el estadístico $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$.
(b) ¿Cuál es la distribución de este estadístico?
(c) ¿Cuál es la probabilidad de observar una varianza muestral mayor o igual a 1.44?
(d) ¿Es la varianza observada compatible con $\sigma^2 = 1.0$?

5.13. Respuestas a los Ejercicios

Ejercicio 5.1: - (a) $\text{Var}(\bar{X}) = \frac{200^2}{100} \cdot \frac{5000-100}{5000-1} \approx 392.08$ g² (con CPF) - (b) $\sigma(\bar{X}) = \sqrt{392.08} \approx 19.80$ gramos - (c) Con $n = 400$: $\text{Var}(\bar{X}) \approx 92.02$ g² y $\sigma(\bar{X}) \approx 9.59$ gramos - (d) Aumentar n reduce el error estándar aproximadamente proporcional a $1/\sqrt{n}$

Ejercicio 5.2:
- (a) $f = 320/8000 = 0.04$ (justo al borde)
- (b) Como $f = 0.04 < 0.05$, la corrección es pequeña pero aplicable
- (c) Sin CPF: Var $= 45^2/320 \approx 6.33$; Con CPF: Var $= 6.33 \times \dfrac{8000-320}{8000-1} \approx 6.08$ (diferencia 4 %)

Ejercicio 5.3:
- (a) $\hat{\pi} = 324/600 = 0.54$
- (b) $\sigma(\hat{\pi}) = \sqrt{0.5 \times 0.5/600} \approx 0.0204$
- (c) IC: $0.54 \pm 1.96 \times 0.0204 \approx [0.500, 0.580]$ - (d) Sí, 0.50 está dentro del intervalo

Ejercicio 5.4:

- (a) La distribución exponencial es sesgada a la derecha
- (b) $\bar{X} \approx N(2000, 40000)$ con $\sigma(\bar{X}) = 200$
- (c) $P(1900 < \bar{X} < 2100) = P(-0.5 < Z < 0.5) \approx 0.383$
- (d) $n \approx 1537$ para precisión de ± 100 euros al 95 %

Ejercicio 5.5:
- (a) $s(\bar{X}) = 400/\sqrt{16} = 100$ gramos
- (b) $t_{0.975,15} \approx 2.131$
- (c) IC: $3500 \pm 2.131 \times 100 \approx [3287, 3713]$ gramos
- (d) Porque σ es desconocida; la distribución t tiene colas más pesadas, reflejando incertidumbre adicional

Ejercicio 5.6:
- (a) $\chi^2 = (24 \times 1.44)/1.0 = 34.56$
- (b) χ^2_{24}
- (c) $P(\chi^2_{24} \geq 34.56) \approx 0.075$ (`1 - pchisq(34.56, 24)`) - (d) El valor es algo alto pero no rechazable al nivel 5 % (el valor crîtico es $\chi^2_{24,\,0.95} \approx 36.42$, mayor que 34.56)

Parte III

Parte III: Inferencia Estadîstica

Capítulo 6

Semana 6 — Estimación de Parámetros

En este capítulo estudiamos los métodos fundamentales para estimar parámetros poblacionales desconocidos basándonos en datos muestrales. Cubriremos desde los conceptos básicos de estimadores puntuales hasta la construcción de intervalos de confianza, pasando por dos de los métodos más importantes: el método de momentos y la estimación por máxima verosimilitud.

6.1. El Modelo Estadístico Paramétrico

Un modelo estadístico paramétrico especifica una familia de distribuciones, cada una de las cuales es indexada por parámetros desconocidos que queremos estimar.

Definición: Modelo Estadístico Paramétrico

Un **modelo estadístico paramétrico** es una colección de distribuciones de probabilidad:

$$\{f(x;\theta) : \theta \in \Theta\}$$

donde:

- $f(x;\theta)$ es la función de densidad o de probabilidad
- $\theta = (\theta_1, \ldots, \theta_p)$ es el vector de parámetros desconocidos
- Θ es el **espacio de parámetros**, el conjunto de todos los valores posibles que θ puede tomar

Ejemplo: Si modelamos una variable aleatoria como Normal, el modelo paramétrico es $\{N(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$, donde los parámetros desconocidos son la media μ y la desviación estándar σ.

6.2. Estimadores Puntuales

Un estimador puntual es una función de la muestra que proporciona un único valor estimado para el parámetro de interés.

Definición: Estimador Puntual

Un **estimador puntual** de un parámetro θ es una función de la muestra:

$$\hat{\theta} = g(X_1, X_2, \dots, X_n)$$

Antes de observar la muestra, $\hat{\theta}$ es una **variable aleatoria** que varía de muestra a muestra. Después de observar los datos muestrales $(x_1, \dots, x_n)$, el estimador toma un valor concreto llamado **estimación** o **valor estimado**:

$$\hat{\theta} = g(x_1, x_2, \dots, x_n)$$

Observación clave: Es importante distinguir entre:

- El **estimador** (antes de ver los datos): variable aleatoria
- El **valor estimado** (después de ver los datos): número concreto

6.3. Propiedades de los Estimadores

Un buen estimador debe satisfacer ciertas propiedades deseables. Las más importantes son la insesgadez, la eficiencia, la consistencia y el error cuadrático medio bajo.

6.3.1. Insesgadez

Definición: Insesgadez

Un estimador $\hat{\theta}$ es **insesgado** para θ si:

$$E(\hat{\theta}) = \theta$$

El **sesgo** (o bias) del estimador se define como:

$$\text{Sesgo}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Un estimador es insesgado si y solo si su sesgo es cero.

⚠ Nota Importante: Insesgadez Asintótica

Un estimador es **asintóticamente insesgado** si:

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta$$

Es decir, el sesgo desaparece cuando el tamaño de muestra crece indefinidamente.

6.3.2. Eficiencia

La eficiencia compara la varianza de estimadores insesgados. Cuanto menor sea la varianza, más eficiente es el estimador.

ℹ Definición: Eficiencia Relativa y Absoluta

Sean $\hat{\theta}^{(1)}$ y $\hat{\theta}^{(2)}$ dos estimadores insesgados de θ (es decir, $E(\hat{\theta}^{(1)}) = E(\hat{\theta}^{(2)}) = \theta$).

- **Eficiencia Relativa:** $\hat{\theta}^{(1)}$ es más eficiente que $\hat{\theta}^{(2)}$ si:

$$\text{Var}(\hat{\theta}^{(1)}) \leq \text{Var}(\hat{\theta}^{(2)})$$

- **Eficiencia Absoluta (o Eficiencia de Mínima Varianza):** $\hat{\theta}^{(1)}$ es eficiente para θ si tiene la menor varianza entre todos los estimadores insesgados de θ.

6.3.3. Consistencia

ℹ Definición: Consistencia

Un estimador $\hat{\theta}$ es **consistente** para θ si se cumplen ambas condiciones:

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta \quad \text{y} \quad \lim_{n\to\infty} \text{Var}(\hat{\theta}) = 0$$

Es decir, a medida que el tamaño de muestra crece, tanto el sesgo como la varianza del estimador se acercan a cero.

La consistencia garantiza que con suficientes datos, el estimador converge en probabilidad al parámetro verdadero.

6.3.4. Error Cuadrático Medio (ECM)

⚠ Resultado Importante: Descomposición del ECM

El **Error Cuadrático Medio** (o Mean Squared Error) se define como:

$$\text{ECM}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

Se puede descomponer como:

$$\text{ECM}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{Sesgo}(\hat{\theta})]^2$$

Esta descomposición muestra que el error total proviene de dos fuentes: la variabilidad del estimador y su sesgo.

Un estimador con menor ECM es preferible, incluso si es ligeramente sesgado. A veces, un estimador sesgado con baja varianza tiene ECM menor que un estimador insesgado con alta varianza.

6.4. Tabla Comparativa de Estimadores Comunes

La siguiente tabla resume propiedades de estimadores frecuentemente utilizados:

Estimador	Parámetro	Sesgo	Varianza	ECM	Condiciones
$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	μ	0	$\frac{\sigma^2}{n}$	$\frac{\sigma^2}{n}$	Insesgado
$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} X_i$	π (proporción)	0	$\frac{\pi(1-\pi)}{n}$	$\frac{\pi(1-\pi)}{n}$	Insesgado, $X_i \sim \text{Bernoulli}(\pi)$
$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$	σ^2 (media conocida)	0	$\frac{2\sigma^4}{n}$	$\frac{2\sigma^4}{n}$	Insesgado si μ es conocida
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$	σ^2 (media desconocida)	0	$\frac{2\sigma^4}{n-1}$	$\frac{2\sigma^4}{n-1}$	Insesgado, varianza muestral
$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$	σ^2 (media desconocida)	$-\frac{\sigma^2}{n}$	$\frac{2\sigma^4(n-1)}{n^2}$	$\frac{\sigma^4(2n-1)}{n^2}$	Sesgado pero asintóticamente insesgado

Nota: Para tamaños de muestra grandes, todos estos estimadores son aproximadamente insesgados y consistentes. La diferencia clave es que S^2 (con divisor $n-1$) es insesgado incluso para muestras pequeñas.

6.5. Método de los Momentos (MoM)

El método de los momentos es uno de los procedimientos más antiguos y simples para estimar parámetros. Se basa en igualar los momentos poblacionales con los momentos muestrales.

Definición: Método de los Momentos

El **método de los momentos** estima los parámetros igualando los primeros k momentos poblacionales con los k momentos muestrales, donde k es el número de parámetros a estimar:
Momentos poblacionales: $E(X^j) = \mu_j$ para $j = 1, 2, \ldots, k$
Momentos muestrales: $m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$ para $j = 1, 2, \ldots, k$
Igualación: $m_j = \mu_j$ para $j = 1, 2, \ldots, k$

6.5.1. Procedimiento del Método de Momentos

El procedimiento consiste en tres pasos:

1. **Calcular momentos poblacionales:** Expresar $E(X^j)$ en términos de los parámetros desconocidos $\theta_1, \ldots, \theta_k$.
2. **Invertir las expresiones:** Despejar los parámetros en términos de los momentos poblacionales.
3. **Sustituir momentos muestrales:** Reemplazar los momentos poblacionales por sus versiones muestrales para obtener los estimadores.

6.5.2. Ejemplo 6.5.1: Distribución Normal

Supongamos que tenemos una muestra de una distribución Normal $N(\mu, \sigma^2)$ con ambos parámetros desconocidos.

Paso 1: Momentos poblacionales:

- Primer momento: $E(X) = \mu$
- Segundo momento: $E(X^2) = \mu^2 + \sigma^2$

Paso 2: Inversión:

- $\mu = E(X)$
- $\sigma^2 = E(X^2) - [E(X)]^2$

Paso 3: Estimadores MoM:

$$\hat{\mu}_{\text{MoM}} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2_{\text{MoM}} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = S'^2$$

Ejemplo: Estimación por Momentos en R

Cuadro 6.1 Código R

```
# Simular niveles de glucosa basal en ayunas (mg/dL) de 100 pacientes
# Modelo: Normal con   = 95 mg/dL (rango normoglucemia) y   = 12 mg/dL
set.seed(123)
glucosa <- rnorm(100, mean = 95, sd = 12)

# Momentos muestrales
m1 <- mean(glucosa)           # Primer momento muestral
m2 <- mean(glucosa^2)         # Segundo momento muestral

# Estimadores por método de momentos
mu_mom <- m1
sigma2_mom <- m2 - m1^2

cat("Estimadores por Método de Momentos (Glucosa basal, mg/dL):\n")
```

```
Estimadores por Método de Momentos (Glucosa basal, mg/dL):
```

Cuadro 6.2 Código R

```
cat("  (MoM)  =", round(mu_mom, 2), "mg/dL\n")
```

```
  (MoM)  = 96.1 mg/dL
```

Cuadro 6.3 Código R

```
cat(" ² (MoM) =", round(sigma2_mom, 2), "\n")
```

```
 ² (MoM) = 119
```

Cuadro 6.4 Código R

```
# Comparar con los parámetros verdaderos
cat("\nParámetros verdaderos:\n")
```

```

Parámetros verdaderos:
```

Cuadro 6.5 Código R

```
cat("         =", 95, "mg/dL\n")
```

```
         = 95 mg/dL
```

Cuadro 6.6 Código R

```
cat(" ²       =", 144, "\n")
```

```
²       = 144
```

Interpretación

El método de momentos recupera una estimación de la media próxima al verdadero = 95 mg/dL (96.1 mg/dL) y una estimación de la varianza algo inferior al verdadero ² = 144 (² 119) — esta discrepancia refleja principalmente la **variabilidad muestral** del estimador (con $n = 100$, el error estándar de S'^2 es 20); el sesgo teórico ($-\sigma^2/n \approx -1.44$) contribuye solo marginalmente. Asintóticamente, ambas estimaciones convergen a los parámetros poblacionales, evidenciando que el procedimiento de igualar momentos muestrales con poblacionales proporciona estimadores consistentes. Esta propiedad es fundamental en estudios epidemiológicos sobre control glucémico, donde se busca estimar parámetros poblacionales de biomarcadores (glucemia, HbA1c, insulina) a partir de cohortes muestrales. El MoM es un método práctico y computacionalmente eficiente para caracterizar distribuciones de variables clínicas continuas.

6.5.3. Ejemplo 6.5.2: Distribución de Poisson

Para una distribución de Poisson Poisson(λ):

- Momento poblacional: $E(X) = \lambda$
- Estimador MoM: $\hat{\lambda}_{\text{MoM}} = \bar{X}$

6.5.4. Ejemplo 6.5.3: Distribución Gamma

Para una distribución Gamma con parámetros α (forma) y β (tasa):

- $E(X) = \alpha/\beta$
- $\text{Var}(X) = \alpha/\beta^2$

Resolviendo:

$$\hat{\alpha}_{\text{MoM}} = \frac{\bar{X}^2}{S^2}, \quad \hat{\beta}_{\text{MoM}} = \frac{\bar{X}}{S^2}$$

6.6. Estimación por Máxima Verosimilitud (MLE)

El método de máxima verosimilitud es el procedimiento más importante y ampliamente utilizado en estadística. Se basa en el principio de que la muestra observada es la "más probable" bajo el verdadero modelo.

6.6.1. La Función de Verosimilitud

Definición: Función de Verosimilitud

Para una muestra aleatoria simple $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ con valores observados $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, la **función de verosimilitud** es:

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta)$$

Si las observaciones son independientes (asunción estándar en muestreo aleatorio simple), la verosimilitud conjunta es el producto de las verosimilitudes individuales.
Interpretación: $L(\theta; \mathbf{x})$ mide la probabilidad (o densidad de probabilidad) de observar la muestra $\mathbf{x}$ para cada posible valor del parámetro θ.

Después de observar los datos, $L(\theta)$ se considera como una función de θ solamente (los datos son fijos). Diferentes valores de θ producen diferentes valores de $L(\theta)$.

6.6.2. El Estimador de Máxima Verosimilitud

Definición: Estimador de Máxima Verosimilitud (EMV)

El **estimador de máxima verosimilitud** (o MLE, del inglés) de θ es el valor que maximiza la función de verosimilitud:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(\theta; \mathbf{x})$$

Si la verosimilitud es diferenciable, el EMV satisface:

$$\left.\frac{\partial L(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

y verificar que es un máximo: $\left.\frac{\partial^2 L(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0$

6.6.3. La Función Log-Verosimilitud

En la práctica, es más fácil maximizar el **logaritmo** de la verosimilitud, llamado **log-verosimilitud**:

$$\ell(\theta; \mathbf{x}) = \ln L(\theta; \mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i; \theta)$$

Propiedad Fundamental

Las funciones $L(\theta)$ y $\ell(\theta)$ tienen el mismo máximo porque el logaritmo es una función monótona creciente:

$$\max_{\theta} L(\theta) \Leftrightarrow \max_{\theta} \ell(\theta)$$

Por lo tanto:

$$\left.\frac{\partial \ell(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0 \quad \Leftrightarrow \quad \left.\frac{\partial L(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

Usar log-verosimilitud es computacionalmente más conveniente porque:

- Convierte productos en sumas
- Es numéricamente más estable
- Es más fácil de derivar

6.6.4. Ejemplo 6.6.1: MLE para Distribución Binomial

Supongamos que observamos x éxitos en n ensayos Bernoulli independientes, donde p es la probabilidad desconocida de éxito.

Función de verosimilitud:

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Log-verosimilitud:

$$\ell(p) = \ln \binom{n}{x} + x \ln p + (n-x) \ln(1-p)$$

Derivada:

$$\frac{\partial \ell}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p}$$

Igualando a cero:

$$\frac{x}{p} = \frac{n-x}{1-p}$$

$$x(1-p) = (n-x)p$$

$$x = np$$

Estimador MLE:

$$\hat{p}_{\text{MLE}} = \frac{x}{n}$$

6.6.5. Ejemplo 6.6.2: MLE para Distribución Normal

Para una muestra de una distribución Normal $N(\mu, \sigma^2)$:

Log-verosimilitud:

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Derivadas parciales:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

Igualando a cero y resolviendo:

$$\hat{\mu}_{\text{MLE}} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = S'^2$$

Nota: Para la media, el EMV coincide con el estimador MoM. Para la varianza, el EMV es S'^2 (con divisor n), que es sesgado, mientras que el estimador por el **método de momentos** es también S'^2 (ambos métodos coinciden en este caso, aunque son procedimientos distintos: MoM no es un EMV en general).

6.6.6. Propiedades del EMV

⚠ Propiedades Importantes del EMV

Bajo condiciones de regularidad (que suelen cumplirse en la práctica):

1. **Consistencia:** El EMV es consistente, es decir, $\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta_0$ cuando $n \to \infty$.
2. **Eficiencia Asintótica:** El EMV es asintóticamente eficiente; es decir, entre todos los estimadores consistentes, el EMV tiene la menor varianza asintótica.
3. **Distribución Asintótica:** Para tamaños de muestra grandes,

$$\hat{\theta}_{\text{MLE}} \approx N\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

donde $I(\theta)$ es la información de Fisher.

4. **Invariancia:** Si $\hat{\theta}$ es el EMV de θ, entonces $g(\hat{\theta})$ es el EMV de $g(\theta)$ para cualquier función g.

6.7. Optimización Numérica para MLE

En muchos casos, no es posible encontrar una solución analítica para el EMV. En estos casos, usamos algoritmos numéricos de optimización.

Ejemplo: Optimización Numérica con optim() en R

Modelamos el **tiempo (en días) hasta la primera hipoglucemia** en 100 pacientes con diabetes tipo 1 mediante una distribución Gamma. Los parámetros `shape` y `rate` no tienen forma analítica cerrada en su EMV conjunto, por lo que recurrimos a optimización numérica.

Cuadro 6.7 Código R

```
# Tiempo (días) hasta primera hipoglucemia - 100 pacientes DM1
# Modelo poblacional: Gamma(shape = 2, rate = 0.5)
# E[T] =  /  = 4 días, Var[T] =  / ² = 8
set.seed(42)
tiempo_hipo <- rgamma(100, shape = 2, rate = 0.5)

# Función de log-verosimilitud negativa
# (minimizamos lo negativo para usar funciones de minimización)
neg_log_lik <- function(params, datos) {
  shape <- params[1]
  rate <- params[2]

  # Evitar parámetros no válidos
  if (shape <= 0 || rate <= 0) return(Inf)

  # Log-verosimilitud para Gamma
  # f(t;  ,  ) = ( ^  / Γ( )) * t^( -1) * exp(- t)
  -sum(dgamma(datos, shape = shape, rate = rate, log = TRUE))
}

# Valores iniciales
inicial <- c(shape = 1, rate = 1)

# Optimización numérica de Nelder-Mead
resultado <- optim(inicial, neg_log_lik, datos = tiempo_hipo,
                   method = "Nelder-Mead")

cat("Estimadores MLE numéricos - Tiempo hasta hipoglucemia (días):\n")
```

```
Estimadores MLE numéricos - Tiempo hasta hipoglucemia (días):
```

Cuadro 6.8 Código R

```
cat("Shape ( ) =", round(resultado$par[1], 4), "\n")
```

```
Shape ( ) = 2.1
```

Cuadro 6.9 Código R

```
cat("Rate ( )  =", round(resultado$par[2], 4), "\n")
```

```
Rate ( )  = 0.55
```

Cuadro 6.10 Código R

```
cat("Tiempo medio estimado E[T] =  /  =",
    round(resultado$par[1]/resultado$par[2], 2), "días\n")
```

```
Tiempo medio estimado E[T] =  /  = 3.81 días
```

Cuadro 6.11 Código R

```
# Valores verdaderos del modelo
cat("\nParámetros poblacionales verdaderos:\n")
```

```
Parámetros poblacionales verdaderos:
```

Cuadro 6.12 Código R

```
cat("Shape ( ) = 2 ; Rate ( ) = 0.5 ; E[T] = 4 días\n")
```

```
Shape ( ) = 2 ; Rate ( ) = 0.5 ; E[T] = 4 días
```

Interpretación

La optimización numérica recupera estimadores MLE de los parámetros de forma (ˆ 2.10) y tasa (ˆ 0.55) próximos a los parámetros poblacionales verdaderos (= 2, = 0.5), con tiempo medio estimado hasta hipoglucemia E[T] = ˆ/ˆ 3.81 días (frente al valor poblacional 4 días). La pequeña discrepancia es consistente con la variabilidad muestral esperable para $n = 100$. El algoritmo de Nelder-Mead converge adecuadamente sin necesidad de derivadas analíticas, lo cual es fundamental en aplicaciones biomédicas de análisis de supervivencia (tiempo hasta evento clínico) donde la verosimilitud carece de forma cerrada. La penalización con `Inf` para parámetros no admisibles y el uso de log-densidades garantizan estabilidad numérica. En diabetes tipo 1, conocer la distribución del tiempo hasta hipoglucemia permite diseñar pautas de monitorización glucémica y educación al paciente basadas en cuantiles poblacionales.

6.8. Intervalos de Confianza

Un intervalo de confianza proporciona un rango de valores plausibles para el parámetro desconocido, cuantificando la incertidumbre en nuestra estimación.

6.8.1. Definición de Intervalo de Confianza

Definición: Intervalo de Confianza

Un **intervalo de confianza** de nivel $1-\alpha$ (o nivel de confianza $100(1-\alpha)\,\%$) para el parámetro θ es un intervalo aleatorio $[L(\mathbf{X}), U(\mathbf{X})]$ tales que:

$$P(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})) = 1 - \alpha$$

donde $L(\mathbf{X})$ y $U(\mathbf{X})$ son funciones de la muestra aleatoria, y $1-\alpha$ se llama **nivel de confianza** o **coeficiente de confianza**.
Típicamente, $1 - \alpha = 0.90$, 0.95, o 0.99.

6.8.2. Interpretación Frecuentista

Interpretación Correcta de un IC

Antes de observar los datos: El intervalo $[L(\mathbf{X}), U(\mathbf{X})]$ es **aleatorio**. La afirmación $P(L \leq \theta \leq U) = 1 - \alpha$ significa que si repitiéramos el procedimiento de muestreo muchas veces, aproximadamente $100(1 - \alpha)\,\%$ de los intervalos resultantes contendrían el verdadero parámetro.
Después de observar los datos: El intervalo observado $[l, u]$ es **fijo** (con números reales). No es correcto decir "la probabilidad de que θ esté en $[l, u]$ es $1-\alpha$" porque θ es una constante desconocida, no una variable aleatoria. Simplemente decimos que el intervalo observado es una realización de un procedimiento que acierta $100(1 - \alpha)\,\%$ de las veces.

6.8.3. Intervalo de Confianza para la Media (Varianza Conocida)

Supuestos: - Población Normal: $X \sim N(\mu, \sigma)$ - Varianza poblacional σ^2 **conocida** - Muestra aleatoria simple de tamaño n

Distribución del estimador:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \Rightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Construcción del IC: Para un nivel de confianza $1 - \alpha$, buscamos $z_{1-\alpha/2}$ tal que $P(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$.

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Intervalo de Confianza:

$$\text{IC}_{1-\alpha}(\mu) = \left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

6.8.4. Intervalo de Confianza para la Media (Varianza Desconocida)

Supuestos: - Población Normal: $X \sim N(\mu, \sigma)$ - Varianza poblacional σ^2 **desconocida**

Distribución del estimador:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

donde $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ es la desviación estándar muestral.

Intervalo de Confianza:

$$\text{IC}_{1-\alpha}(\mu) = \left[\bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}, \quad \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$$

donde $t_{1-\alpha/2,n-1}$ es el cuantil $(1-\alpha/2)$ de la distribución t con $n-1$ grados de libertad.

Ejemplo: IC para la Media en R

Cuadro 6.13 Código R

```
# Presión arterial sistólica (mmHg) en 25 pacientes adultos
# Modelo poblacional: Normal con    130 mmHg y   = 15 mmHg
set.seed(100)
pas <- rnorm(25, mean = 130, sd = 15)

# IC con t-Student (varianza desconocida) - caso habitual en clínica
resultado <- t.test(pas, conf.level = 0.95)

cat("Intervalo de Confianza 95% para la PAS media (mmHg):\n")
```

```
Intervalo de Confianza 95% para la PAS media (mmHg):
```

Cuadro 6.14 Código R

```
cat("Media muestral:", round(resultado$estimate, 2), "mmHg\n")
```

```
Media muestral: 132 mmHg
```

Cuadro 6.15 Código R

```
cat("IC: [", round(resultado$conf.int[1], 2), ",",
    round(resultado$conf.int[2], 2), "] mmHg\n")
```

```
IC: [ 127 , 136 ] mmHg
```

Cuadro 6.16 Código R

```
# IC con z (varianza poblacional  = 15 mmHg supuesta conocida)
# Útil cuando se dispone de datos históricos/literatura
n <- length(pas)
media_muestral <- mean(pas)
sigma <- 15
error_estandar <- sigma / sqrt(n)
z_critico <- qnorm(0.975)  #  = 0.05, bilateral

IC_conocida <- c(media_muestral - z_critico * error_estandar,
                 media_muestral + z_critico * error_estandar)

cat("\nIC con varianza conocida (  = 15 mmHg):\n")
```

```
IC con varianza conocida (  = 15 mmHg):
```

Cuadro 6.17 Código R

```
cat("IC: [", round(IC_conocida[1], 2), ",",
    round(IC_conocida[2], 2), "] mmHg\n")
```

```
IC: [ 126 , 138 ] mmHg
```

6.8.5. Intervalo de Confianza para una Proporción

Supuestos: - Población dicotómica: $P(A) = \pi$ (proporción de éxitos) - Muestra aleatoria simple de tamaño n (con n suficientemente grande) - Condiciones: $n\pi \geq 5$ y $n(1-\pi) \geq 5$ (o aproximadamente $n > 30$)

Distribución del estimador:

$$\hat{\pi} = \frac{X}{n} \approx N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

Problema: La varianza contiene el parámetro desconocido π. **Solución:** Usar $\hat{\pi}$ como estimador

consistente de π en la varianza.

Intervalo de Confianza (aproximado):

$$\text{IC}_{1-\alpha}(\pi) = \left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \quad \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

Ejemplo: IC para una Proporción en R

Cuadro 6.18 Código R

```
# Estudio en 200 pacientes hipertensos en seguimiento clínico:
# 132 alcanzan control adecuado de PAS (< 140/90 mmHg)

n <- 200
x <- 132
p_hat <- x / n

# Intervalo de confianza 95% para la proporción
conf_level <- 0.95
alpha <- 1 - conf_level
z_critico <- qnorm(1 - alpha/2)

error_estandar <- sqrt(p_hat * (1 - p_hat) / n)
IC <- c(p_hat - z_critico * error_estandar,
        p_hat + z_critico * error_estandar)

cat("IC 95% para la proporción de hipertensos controlados:\n")
```

```
IC 95% para la proporción de hipertensos controlados:
```

Cuadro 6.19 Código R

```
cat("Proporción muestral:", round(p_hat, 4),
    "(", round(p_hat*100, 1), "%)\n")
```

```
Proporción muestral: 0.66 ( 66 %)
```

Cuadro 6.20 Código R

```
cat("IC: [", round(IC[1], 4), ",", round(IC[2], 4), "]\n")
```

```
IC: [ 0.594 , 0.726 ]
```

Cuadro 6.21 Código R

```
cat("Interpretación: con 95% de confianza, entre el",
    round(IC[1]*100, 1), "% y el",
    round(IC[2]*100, 1), "%\n")
```

```
Interpretación: con 95% de confianza, entre el 59.4 % y el 72.6 %
```

Cuadro 6.22 Código R

```
cat("de los pacientes hipertensos poblacionales alcanzan control adecuado.\n")
```

```
de los pacientes hipertensos poblacionales alcanzan control adecuado.
```

6.9. Interpretación de Intervalos de Confianza

Interpretar correctamente un intervalo de confianza es fundamental. Consideremos un ejemplo:

Supongamos que construimos un IC 95 % para la media y obtenemos $[48.3, 51.7]$.

⚠ Interpretación CORRECTA

"Si repitiéramos el proceso de muestreo muchas veces (con muestras de igual tamaño de la misma población) y construyéramos un intervalo de confianza del 95 % para cada muestra, aproximadamente el 95 % de estos intervalos contendría el verdadero valor de la media poblacional."
En otras palabras, el IC es sobre la **fiabilidad del procedimiento**, no sobre la probabilidad de que un intervalo específico contenga el parámetro.

⚠ Interpretación INCORRECTA (Evitar)

"La verdadera media está en $[48.3, 51.7]$ con probabilidad 0.95"
Esto es incorrecto porque después de observar los datos, el intervalo es fijo y la verdadera media es constante (no aleatoria). El parámetro o está en el intervalo o no está.

6.10. Trabajando con Ejemplos Completos

Ejemplo: Análisis Completo de Datos de HbA1c (Diabetes Tipo 2)

Cuadro 6.23 Código R

```
# Datos de HbA1c (%) de 20 pacientes con diabetes tipo 2
# atendidos en la Facultad de Medicina, Universidad de Granada

hba1c <- c(7.2, 8.1, 6.8, 9.3, 7.5, 8.8, 6.5, 10.2,
           7.9, 8.4, 7.1, 9.7, 7.3, 8.6, 6.9, 11.4,
           8.2, 7.0, 9.1, 7.6)

# 1. ESTADÍSTICA DESCRIPTIVA
media_muestral <- mean(hba1c)
desv_est <- sd(hba1c)  # Usa divisor n-1 (insesgado)
n <- length(hba1c)

cat("=== ESTADÍSTICA DESCRIPTIVA ===\n")
```

```
=== ESTADÍSTICA DESCRIPTIVA ===
```

Cuadro 6.24 Código R

```
cat("n =", n, "\n")
```

```
n = 20
```

Cuadro 6.25 Código R

```
cat("Media muestral = ", round(media_muestral, 2), "%\n")
```

```
Media muestral =  8.18 %
```

Cuadro 6.26 Código R

```
cat("Desviación estándar = ", round(desv_est, 2), "%\n")
```

```
Desviación estándar =  1.28 %
```

Cuadro 6.27 Código R

```
cat("Varianza muestral = ", round(desv_est^2, 2), "\n\n")
```

```
Varianza muestral =  1.63
```

Cuadro 6.28 Código R

```
# 2. ESTIMACIÓN PUNTUAL (MLE/MoM para Normal)
cat("=== ESTIMACIÓN PUNTUAL ===\n")
```

```
=== ESTIMACIÓN PUNTUAL ===
```

Cuadro 6.29 Código R

```
cat("  (estimado) = ", round(media_muestral, 2), "%\n")
```

```
  (estimado) =  8.18 %
```

Cuadro 6.30 Código R

```
cat(" ² (estimado) = ", round(desv_est^2, 2), "\n\n")
```

```
 ² (estimado) =  1.63
```

Cuadro 6.31 Código R

```
# 3. INTERVALO DE CONFIANZA 95% (varianza desconocida)
cat("=== INTERVALO DE CONFIANZA (95%) ===\n")
```

```
=== INTERVALO DE CONFIANZA (95%) ===
```

Cuadro 6.32 Código R

```
t_critico <- qt(0.975, df = n - 1)
error_estandar <- desv_est / sqrt(n)
margen_error <- t_critico * error_estandar

IC_inferior <- media_muestral - margen_error
IC_superior <- media_muestral + margen_error

cat("Error estándar = ", round(error_estandar, 2), "%\n")
```

```
Error estándar =  0.29 %
```

Cuadro 6.33 Código R

```
cat("Margen de error = ", round(margen_error, 2), "%\n")
```

```
Margen de error =  0.6 %
```

Cuadro 6.34 Código R

```
cat("IC 95% = [", round(IC_inferior, 2), "%, ",
    round(IC_superior, 2), "%]\n\n")
```

```
IC 95% = [ 7.58 %,  8.78 %]
```

Cuadro 6.35 Código R

```
# Alternativamente, usar t.test()
test_result <- t.test(hba1c, conf.level = 0.95)
cat("Verificación con t.test():\n")
```

```
Verificación con t.test():
```

Cuadro 6.36 Código R

```
cat("IC 95% = [", round(test_result$conf.int[1], 2), "%, ",
    round(test_result$conf.int[2], 2), "%]\n")
```

```
IC 95% = [ 7.58 %,  8.78 %]
```

Interpretación: Con el 95 % de confianza, el nivel medio de HbA1c poblacional en pacientes diabéticos está entre 7.58 % y 8.78 % (media muestral = 8.18 %, $s = 1.28$ %, $n = 20$). El IC no incluye el objetivo clínico de HbA1c < 7 %, indicando que esta cohorte presenta un control glucémico subóptimo (por encima del objetivo terapéutico recomendado). El margen de error (~0.60 %) refleja la variabilidad en control metabólico entre pacientes, lo que en práctica clínica orienta decisiones de intensificación terapéutica.

Ejemplo: MLE para tiempos entre eventos cardiovasculares

Modelamos el **tiempo (en años) entre eventos cardiovasculares mayores** en una cohorte de 50 pacientes con cardiopatía isquémica como una distribución Exponencial con tasa . La media poblacional es $1/\lambda$ años entre eventos.

Cuadro 6.37 Código R

```
# Tiempo (años) entre eventos cardiovasculares - 50 pacientes
# Modelo poblacional: Exponencial con tasa   = 0.5 eventos/año
# (i.e., un evento cada 2 años en promedio)
set.seed(999)
tiempo_eventos <- rexp(50, rate = 0.5)

# Función de log-verosimilitud para Exponencial
log_verosimilitud <- function(lambda, datos) {
  if (lambda <= 0) return(-Inf)
  sum(dexp(datos, rate = lambda, log = TRUE))
}

# Maximizar log-verosimilitud numéricamente
resultado_optim <- optimize(
  function(lambda) log_verosimilitud(lambda, tiempo_eventos),
  interval = c(0.01, 5),
  maximum = TRUE
)

lambda_mle <- resultado_optim$maximum
cat("Estimador MLE de  :", round(lambda_mle, 4), "eventos/año\n")
```

```
Estimador MLE de  : 0.511 eventos/año
```

Cuadro 6.38 Código R

```
cat("Log-verosimilitud máxima:",
    round(resultado_optim$objective, 2), "\n")
```

```
Log-verosimilitud máxima: -83.5
```

Cuadro 6.39 Código R

```
cat("Tiempo medio estimado entre eventos: 1/  =",
    round(1/lambda_mle, 2), "años\n")
```

```
Tiempo medio estimado entre eventos: 1/  = 1.96 años
```

Cuadro 6.40 Código R

```
# MoM también da   = 1/media para Exponencial
lambda_mom <- 1 / mean(tiempo_eventos)
cat("\nEstimador MoM de  :", round(lambda_mom, 4), "eventos/año\n")
```

```
Estimador MoM de  : 0.511 eventos/año
```

Cuadro 6.41 Código R

```
# Para Exponencial, MLE = MoM (propiedad teórica)
cat("\nNota: Para la distribución Exponencial,\n")
```

```
Nota: Para la distribución Exponencial,
```

Cuadro 6.42 Código R

```
cat("MLE y MoM coinciden algebraicamente (^ = 1/X).\n")
```

```
MLE y MoM coinciden algebraicamente (^ = 1/X).
```

6.11. Resumen

Concepto	Definición Clave
Modelo Paramétrico	Familia de distribuciones $\{f(x;\theta) : \theta \in \Theta\}$
Estimador	Función de la muestra $\hat{\theta} = g(X_1, \ldots, X_n)$ que varía aleatoriamente
Insesgadez	$E(\hat{\theta}) = \theta$
Eficiencia	Comparación de varianzas entre estimadores insesgados
Consistencia	$\lim_{n\to\infty} E(\hat{\theta}) = \theta$ y $\lim_{n\to\infty} \text{Var}(\hat{\theta}) = 0$
Error Cuadrático Medio	$\text{ECM}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{Sesgo}(\hat{\theta})]^2$
Método de Momentos	Igualar momentos poblacionales con muestrales: $E(X^j) = m_j$
Función de Verosimilitud	$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$
Log-Verosimilitud	$\ell(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta)$
Estimador MLE	$\hat{\theta}_{\text{MLE}} = \arg\max_\theta L(\theta)$
Intervalo de Confianza	Intervalo aleatorio $[L(\mathbf{X}), U(\mathbf{X})]$ con $P(L \leq \theta \leq U) = 1 - \alpha$

6.11.1. Fórmulas Clave para Intervalos de Confianza

Media (varianza conocida):

$$\text{IC}_{1-\alpha}(\mu) = \left[\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Media (varianza desconocida):

$$\text{IC}_{1-\alpha}(\mu) = \left[\bar{x} \pm t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$$

Proporción:

$$\mathrm{IC}_{1-\alpha}(\pi) = \left[\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

Ejemplo con BioEstatR: icm() e icp()

Documentación completa de ambas funciones en Sección B.3.

Cuadro 6.43 Código R

```
library(BioEstatR)

# IC para la media de HbA1c - 94 pacientes diabéticos, Fac. Medicina UGR
icm(m = mean(osteo$hba1c), s = sd(osteo$hba1c), n = nrow(osteo))
```

```
Intervalo de confianza bilateral para la media de una VA normal
---------------------------------------------------------------
Información muestral:
  Tamaño muestral: n =  94
  Media: m =  8.565
  Desviación típica: s =  1.799
  Error estándar de la media: sem =  0.186

Estimación:
 95%-IC(μ):  (8.2, 8.93)
  Precisión obtenida: 0.368
```

Cuadro 6.44 Código R

```
# IC para la proporción de fumadores (41 de 94 pacientes)
icp(x = sum(osteo$tabaco == "Sí"), n = nrow(osteo))
```

```
Intervalo de confianza para una proporción binomial
---------------------------------------------------

Información muestral:
  Tamaño de muestra: n = 94
  Estimación puntual clásica: p=x/n = 0.436, q=(1-p)=0.564
  Casos observados: x = 41

# Método exacto (Clooper-Pearson):
  Pseudo-estimación puntual: p' = 0.441, q'=(1-p')=0.559
  95%-IC( ): (0.339, 0.542)
  Semiamplitud: 0.102
```

```
# Método de Wilson (con cpc):
  Pseudo-estimación puntual: p' = 0.439, q'=(1-p')=0.561
  95%-IC( ): (0.335, 0.542)
  Semiamplitud: 0.103

# Método de Wald (con cpc):
  Estimación puntual (clásica): p=x/n = 0.436, q=(1-p)=0.564
  95%-IC( ): (0.331, 0.542)
  Precisión: 0.106

# Método de Wald ajustado (Agresti-Coull):
  Estimación puntual: p=(x+2)/(n+4) = 0.439, q=(1-p)=0.561
  95%-IC( ): (0.341, 0.537)
  Precisión: 0.0982
```

`icp()` calcula cuatro métodos simultáneamente. Para muestras pequeñas se recomienda Clopper-Pearson (exacto); para muestras grandes los cuatro métodos **convergen** (las diferencias entre ellos disminuyen al aumentar n).

6.12. Cálculo del Tamaño Muestral: `nm()` y `np()`

Antes de iniciar un estudio clínico es esencial determinar el **tamaño muestral** necesario para alcanzar una precisión deseada en los intervalos de confianza. Las funciones `nm()` y `np()` del paquete `BioEstatR` (ver Sección B.4) calculan el tamaño muestral requerido para una **media** o una **proporción**, respectivamente, dada una semiamplitud objetivo d del IC.

6.12.1. Fórmulas del Tamaño Muestral

Para la media (con estimación piloto s de la desviación típica):

$$n = \left(\frac{z_{1-\alpha/2}\, s}{d} \right)^2$$

Para la proporción (con estimación piloto $\hat{p} = x/n$):

$$n = \left(\frac{z_{1-\alpha/2}}{d} \right)^2 \hat{p}\,(1-\hat{p})$$

donde d es la **semiamplitud** (precisión) deseada del IC y $z_{1-\alpha/2}$ es el cuantil normal estándar. La dependencia cuadrática $n \propto 1/d^2$ implica que reducir la semiamplitud a la mitad cuadruplica el tamaño muestral.

Ejemplo: Tamaño muestral para HbA1c y prevalencia de tabaquismo

Partiendo del estudio piloto `osteo` (94 pacientes diabéticos, Fac. Medicina UGR):

- **HbA1c**: $\bar{x} \approx 8.57\,\%$, $s \approx 1.80$. ¿Cuántos pacientes hacen falta para reducir la semiamplitud del IC a $d = 0.3\,\%$?
- **Tabaquismo**: 41/94 fumadores ($\hat{p} \approx 0.436$). ¿Cuántos pacientes para alcanzar precisión $d = 0.05$ al $95\,\%$ de confianza?

Cuadro 6.45 Código R

```r
library(BioEstatR)

# nm(): tamaño muestral para la media de HbA1c
# d = 0.3% (semiamplitud deseada), datos piloto del estudio osteo
nm(d = 0.3, n = nrow(osteo),
   m = mean(osteo$hba1c), s = sd(osteo$hba1c), alfa = 0.05)
```

```
# Tamaño de muestra para la estimación de la media de una VA normal o su aproximación
# ------------------------------------------------------------------------------

# Muestra piloto:
  Tamaño muestral:  n = 94
  Media: m = 8.5649
  Desviación típica: s = 1.7987
  Error estandar de la media: sem = 0.1855
  Precisión observada: d = 0.3684

# Estimación del tamaño muestral:
  Precisión deseada:   = 0.3000
  Tamaño muestral necesario: n   142
```

Cuadro 6.46 Código R

```r
# np(): tamaño muestral para la proporción de fumadores
# d = 0.05 (semiamplitud deseada), conf = 0.95
np(x = sum(osteo$tabaco == "Sí"), n = nrow(osteo),
   d = 0.05, conf = 0.95)
```

```
Tamaño de muestra para estimar una proporción binomial
------------------------------------------------------

Información muestral
  Tamaño de la muestra: n = 94
```

```
  Casos: x = 41
  Inferencia para la proporción basada en el método de Wald ajustado:
  95%-IC( ): (0.3405, 0.5370)
  precisión observada: d = 0.0982 (9.82%)

Tamaño muestral requerido para   = 0.05 (5.00%), conf.= 95%
  - Basado en la muestra actual (po = 0.5370):   n   383
  - Sin considerar la información previa: n   385
```

Interpretación

El cálculo del tamaño muestral cuantifica el coste estadístico de incrementar la precisión. Para estimar la HbA1c media con una semiamplitud del IC $\leq 0.3\,\%$ (frente a los $\approx 0.37\,\%$ obtenidos con $n = 94$), se necesitarían **n 142 pacientes diabéticos** — un incremento del $51\,\%$. Análogamente, para estimar la prevalencia de tabaquismo con precisión $\pm 5\,\%$ (IC $95\,\%$) se requerirían **n 383 pacientes** (usando la información de la muestra piloto), aproximadamente cuatro veces el tamaño del estudio piloto. En la fase de diseño de un protocolo clínico, estos cálculos orientan el balance entre **precisión científica** (estimaciones más estrechas guían mejor las decisiones terapéuticas) y **viabilidad logística-financiera** (reclutamiento, seguimiento, presupuesto). La dependencia cuadrática $n \propto 1/d^2$ explica por qué ganar un factor 2 en precisión cuadruplica el coste muestral — un argumento decisivo al justificar el tamaño muestral ante comités éticos y agencias financiadoras.

6.13. Ejercicios

1. **Concepto de Estimador:** Define con tus palabras la diferencia entre un parámetro poblacional, un estimador y un valor estimado. ¿Por qué es importante considerar un estimador como una variable aleatoria?

2. **Insesgadez vs. Eficiencia:** Considera dos estimadores insesgados del mismo parámetro. ¿Cuál prefieres si uno tiene varianza dos veces mayor que el otro? Razona tu respuesta.

3. **Error Cuadrático Medio:** Explica por qué a veces puede ser preferible un estimador sesgado con baja varianza sobre uno insesgado con alta varianza. Proporciona un ejemplo numérico simple.

4. **Método de Momentos:** Para una muestra de una distribución Uniform$[0, \theta]$:

   - Encuentra el estimador por método de momentos de θ
   - ¿Es insesgado?

5. **Máxima Verosimilitud:** Para una muestra de n observaciones de una distribución Poisson(λ):

   - Escribe la función de verosimilitud
   - Encuentra el EMV de λ

- ¿Coincide con el estimador MoM?

6. **Intervalo de Confianza:** Se observa una muestra de 36 datos con $\bar{x} = 50$ y $s = 12$. Construye un IC 95 % para la media poblacional. ¿Cómo cambiaría el intervalo si usáramos un nivel de confianza del 99 %?

7. **Intervalo de Confianza para Proporción:** En una encuesta de 500 personas, 325 responden afirmativamente a una pregunta. Construye un IC 95 % para la proporción poblacional e interpreta el resultado.

8. **Programación en R:** Escribe un script en R que:

   - Genere una muestra de una distribución Normal($\mu = 100, \sigma = 15$) de tamaño $n = 50$
   - Calcule estimadores puntuales para μ y σ^2
   - Construya un IC 95 % para μ
   - Verifique si el verdadero valor de μ cae dentro del intervalo

6.14. Respuestas a los Ejercicios

Ejercicio 1: Un parámetro es el valor verdadero desconocido de la población ( ). Un estimador es una función de los datos muestrales (T(X ,…,X)) que genera diferentes valores según la muestra. Es una variable aleatoria porque depende de la muestra. Es importante reconocer esto para entender su distribución y propiedades.

Ejercicio 2: Prefieres el estimador con menor varianza porque tendrá una distribución más concentrada alrededor del parámetro. Si ambos son insesgados, el de menor varianza es más eficiente.

Ejercicio 3: Ejemplo: =0, Estimador A: insesgado, Var=100; Estimador B: sesgo=2, Var=10. ECM_A = 100; ECM_B = 4 + 10 = 14. B es mejor a pesar del sesgo.

Ejercicio 4: Para Uniform[0,], el método de momentos da ̂ = 2X. No es insesgado (subestima).

Ejercicio 5: L( ) = e^(-n) x !/ x !. EMV: ̂ = X. Sí coincide con MoM.

Ejercicio 6: n=36, X=50, s=12, t . , 2.03 IC95 % = [50 ± 2.03×(12/6)] = [50 ± 4.06] = [45.94, 54.06] IC99 %: más amplio [44.55, 55.45] (usa t . , 2.72; margen 2.72×2 = 5.45)

Ejercicio 7: p̂ = 325/500 = 0.65, SE = √(0.65×0.35/500) 0.0214 IC95 % = [0.65 ± 1.96×0.0214] [0.608, 0.692]

Ejercicio 8: Código R básico:

```
set.seed(123)
muestra <- rnorm(50, mean=100, sd=15)
media <- mean(muestra)
var_est <- var(muestra)
se <- sd(muestra)/sqrt(50)
```

```
ic <- c(media - 1.96*se, media + 1.96*se)
# Verificar: 100   ic?
```

Capítulo 7

Semana 7 — Contrastes de Hipótesis (Parte I)

7.1. Introducción

Un **contraste de hipótesis** (también llamado *prueba de hipótesis*) es un procedimiento estadístico para tomar una decisión entre dos afirmaciones contrapuestas acerca de un parámetro desconocido de la población. Utilizamos información de una muestra aleatoria para determinar si la evidencia respalda rechazar o no una afirmación inicial (hipótesis nula).

Ejemplo motivador: Un productor de monedas afirma que sus monedas son justas (probabilidad de cara = 0.5). ¿Cómo comprobamos esta afirmación? Si lanzamos la moneda 100 veces y obtenemos 95 caras, esto sugiere que la moneda no es justa. Pero ¿y si obtenemos 52 caras? El contraste de hipótesis nos proporciona un procedimiento riguroso para tomar esta decisión considerando la aleatoriedad de los datos.

7.2. Hipótesis Nula y Alternativa

En un contraste de hipótesis distinguimos dos hipótesis complementarias:

Definición: Hipótesis Nula (H_0) e Hipótesis Alternativa (H_1)

Hipótesis Nula (H_0): Es la afirmación que asumimos como cierta por defecto. Representa el status quo o la ausencia de efecto. Siempre contiene una igualdad.
Hipótesis Alternativa (H_1): Es la afirmación que queremos demostrar o investigar. Representa la desviación del valor hipotético en H_0.
Dos hipótesis son **complementarias** y **mutuamente excluyentes**: si una es cierta, la otra es falsa. Capturan todos los valores posibles del parámetro.

7.2.1. Tipos de Hipótesis

Hipótesis Simple vs. Compuesta: - Una hipótesis es **simple** si especifica un único valor del parámetro (p.ej., $H_0 : \mu = 100$) - Una hipótesis es **compuesta** si especifica múltiples valores (p.ej., $H_1 : \mu > 100$)

Contrastes Bilaterales vs. Unilaterales:

Tipo	Hipótesis Nula	Hipótesis Alternativa	Descripción
Bilateral	$H_0 : \theta = \theta_0$	$H_1 : \theta \neq \theta_0$	Investigamos desviaciones en ambas direcciones
Unilateral derecho	$H_0 : \theta \leq \theta_0$	$H_1 : \theta > \theta_0$	Investigamos si el parámetro es mayor
Unilateral izquierdo	$H_0 : \theta \geq \theta_0$	$H_1 : \theta < \theta_0$	Investigamos si el parámetro es menor

7.2.2. Ejemplos

1. **Prueba de moneda justa:** $H_0 : \pi = 0.5$ vs. $H_1 : \pi \neq 0.5$ (bilateral)
2. **Control de calidad:** $H_0 : \mu \leq 1000$ vs. $H_1 : \mu > 1000$ (unilateral derecho)
3. **Seguridad de medicamento:** $H_0 : \pi \geq 0.05$ vs. $H_1 : \pi < 0.05$ (unilateral izquierdo, donde π es tasa de efectos adversos)

7.3. Errores Tipo I y Tipo II

Cuando tomamos una decisión basada en datos muestrales, podemos cometer errores. La verdad sobre H_0 es desconocida, pero existen dos posibilidades:

> **i** Definición: Errores Tipo I y Tipo II
>
> **Error Tipo I (Falso Positivo):** Rechazar H_0 cuando en realidad es verdadera. - Probabilidad: $P(\text{Rechazar } H_0 | H_0 \text{ es verdadera}) = \alpha$ - Nivel de significación
> **Error Tipo II (Falso Negativo):** No rechazar H_0 cuando en realidad es falsa. - Probabilidad: $P(\text{No rechazar } H_0 | H_1 \text{ es verdadera}) = \beta$
> **Potencia de la prueba:** Probabilidad de rechazar H_0 cuando es falsa. - Potencia $= 1 - \beta$ - Representa nuestra capacidad de detectar un efecto verdadero

7.3.1. Tabla de Decisión

	H_0 es Verdadera	H_0 es Falsa
Rechazamos H_0	Error Tipo I (probabilidad α)	Decisión Correcta (probabilidad $1-\beta$)
No Rechazamos H_0	Decisión Correcta (probabilidad $1-\alpha$)	Error Tipo II (probabilidad β)

7.3.2. Analogía: El Pronóstico Médico

La estructura de un contraste de hipótesis puede entenderse con una analogía clínica sobre el pronóstico de un paciente:

- H_0**:** El paciente va a morir (pronóstico fatal asumido como hipótesis nula)
- H_1**:** El paciente va a vivir
- **Error Tipo I:** Le dices que **vive**, pero **muere** (rechazas H_0 siendo cierta)
- **Error Tipo II:** Le dices que **muere**, pero **vive** (no rechazas H_0 siendo falsa)

Esta analogía ilustra cómo las consecuencias de los errores definen la gravedad del contraste. En este escenario, el Error Tipo I (falsa esperanza) tiene consecuencias emocionales y logísticas muy distintas al Error Tipo II (falsa alarma).

7.3.3. Relación entre α y β

Observación importante: Para una muestra de tamaño n fijo, **existe una relación inversa entre** α **y** β:

- Si bajamos α (menos riesgo de rechazar H_0 falsamente), entonces β aumenta (mayor riesgo de no rechazar H_0 falsamente)
- Si bajamos β (mejor capacidad de detectar un efecto), entonces α aumenta
- La única forma de reducir ambos simultáneamente es **aumentar el tamaño de la muestra** n

---

7.4. Estadístico de Prueba y Región Crítica

Para ejecutar un contraste de hipótesis necesitamos tres ingredientes:

i Definición: Componentes de un Contraste

1. Estadístico de Prueba (T_n): Una función de los datos muestrales cuya distribución es conocida bajo H_0. Se calcula de la muestra y resume la evidencia contra H_0.

2. Distribución bajo H_0: La distribución de probabilidad del estadístico de prueba asumiendo que H_0 es verdadera.
3. Región de Rechazo (o Región Crítica): El conjunto de valores del estadístico de prueba para los cuales rechazamos H_0. Se define antes de recopilar datos.
Valor Crítico: El punto que separa la región de rechazo de la región de no rechazo.

7.4.1. Procedimiento de Decisión

1. Calcular el valor observado del estadístico de prueba: $t_{obs} = T(x_1, \dots, x_n)$
2. Comparar t_{obs} con la región crítica
3. Si t_{obs} está en la región crítica $\Rightarrow$ Rechazamos H_0
4. Si t_{obs} NO está en la región crítica $\Rightarrow$ No rechazamos H_0

7.5. Función de Potencia

Definición: Función de Potencia

La **función de potencia** es una función de θ que mide la probabilidad de rechazar H_0 para cada valor posible del parámetro:

$$\beta(\theta) = P_\theta(T_n \in R)$$

donde R es la región de rechazo y P_θ significa "probabilidad cuando el verdadero valor del parámetro es θ".
Interpretación: - Cuando $\theta \in \Theta_0$ (bajo H_0 verdadera): $\beta(\theta)$ es la probabilidad de Error Tipo I - Cuando $\theta \in \Theta_1$ (bajo H_1 verdadera): $\beta(\theta) = 1 - P(\text{Error Tipo II})$ es la **potencia**
Un buen contraste tiene:

- Potencia cercana a 0 para $\theta \in \Theta_0$ (pocos falsos positivos)
- Potencia cercana a 1 para $\theta \in \Theta_1$ (muchos verdaderos positivos)

7.6. Tamaño y Nivel de una Prueba

Definición: Tamaño y Nivel

Tamaño de la prueba (α): Es el supremo de la función de potencia sobre Θ_0:

$$\text{Tamaño} = \sup_{\theta \in \Theta_0} \beta(\theta)$$

En otras palabras, es la máxima probabilidad de Error Tipo I sobre todos los valores en H_0.
Nivel de la prueba (α): Es un número entre 0 y 1 (típicamente 0.05, 0.01) tal que:

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$$

La prueba se dice que es de "nivel α" si controla el Error Tipo I en este nivel.

7.6.1. Relación con la Significación

- El **nivel de significación** α es el umbral máximo de riesgo de Error Tipo I que estamos dispuestos a aceptar
- Típicamente: $\alpha = 0.05$ (5 %) o $\alpha = 0.01$ (1 %)
- Se elige **antes** de realizar el análisis
- Una vez elegido, el valor crítico se determina de la distribución bajo H_0 de forma que se cumpla el nivel

---

7.7. Contraste Z para la Media (Varianza Conocida)

Cuando la varianza poblacional σ^2 es **conocida** (situación rara en la práctica), usamos el contraste Z basado en la distribución normal estándar.

7.7.1. Condiciones

- Muestra aleatoria de tamaño n
- Datos $X_i \sim N(\mu, \sigma^2)$ con σ^2 **conocida**, O
- Datos arbitrarios con n grande (Central Limit Theorem)
- Queremos probar hipótesis sobre μ

7.7.2. Estadístico de Prueba

⚠ Resultado Importante: Contraste Z

Bajo $H_0 : \mu = \mu_0$ con σ conocida:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

El estadístico Z sigue una distribución normal estándar cuando H_0 es verdadera.

7.7.3. Regiones de Rechazo y Valores Críticos

Para un nivel de significación α:

Tipo	Región de Rechazo	Valor Crítico	p-valor
Bilateral	$\|Z\| > z_{1-\alpha/2}$	$z_{1-\alpha/2}$	$2P(Z > \|z_{obs}\|)$
Unilateral derecha	$Z > z_{1-\alpha}$	$z_{1-\alpha}$	$P(Z > z_{obs})$
Unilateral izquierda	$Z < -z_{1-\alpha}$	$-z_{1-\alpha}$	$P(Z < z_{obs})$

7.7.4. Valores Críticos Comunes (Normal Estándar)

Cuantil	0.900	0.950	0.975	0.990	0.995
z_q	1.28	1.64	1.96	2.33	2.58

7.7.5. Ejemplo: Productor de Harina

Problema: Un productor de harina quiere verificar si sus paquetes contienen exactamente 1 kg de harina. - Si hay mucho menos: los clientes se quejan - Si hay mucho más: pérdida de ganancias - Se sabe que: $X_i \sim N(\mu, \sigma^2 = 25 \text{ g}^2)$

Hipótesis: - $H_0 : \mu = 1000$ g - $H_1 : \mu \neq 1000$ g (bilateral)

Muestra: $n = 25$, $\bar{x} = 1002$ g

Cálculo:

$$z_{obs} = \frac{1002 - 1000}{5/\sqrt{25}} = \frac{2}{1} = 2$$

Decisión (con $\alpha = 0.05$): - Valor crítico: $z_{1-0.05/2} = z_{0.975} = 1.96$ - Como $|z_{obs}| = 2 > 1.96$, rechazamos H_0 - p-valor: $2P(Z > 2) \approx 0.0455$

Conclusión: A nivel de significación 5 %, la evidencia sugiere que el peso medio es diferente de 1 kg.

7.8. Contraste t para la Media (Varianza Desconocida)

En la práctica, la varianza σ^2 es **desconocida**. En este caso, estimamos σ^2 con S^2 y usamos la distribución t de Student.

7.8.1. Condiciones

- Muestra aleatoria de tamaño n
- Datos $X_i \sim N(\mu, \sigma^2)$ con σ^2 **desconocida**, O
- Datos aproximadamente normales

- Queremos probar hipótesis sobre μ

7.8.2. Estimador de la Varianza

La varianza muestral insesgada es:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

El divisor es $n-1$ (no n) para obtener un estimador insesgado.

7.8.3. Estadístico de Prueba y Distribución

⚠ Resultado Importante: Contraste t

Bajo $H_0 : \mu = \mu_0$ con σ desconocida:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

El estadístico T sigue una distribución t de Student con $n-1$ grados de libertad cuando H_0 es verdadera.

7.8.4. Derivación de la Distribución

Empezamos con:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}}{\frac{S}{\sigma}} = \frac{Z}{\sqrt{\frac{C}{n-1}}}$$

donde:

- $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ (normal estándar)
- $C = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ (chi-cuadrado con $n-1$ gl)

Por definición de la distribución t: Si $Z \sim N(0,1)$ y $C \sim \chi^2_k$ son independientes, entonces $\frac{Z}{\sqrt{C/k}} \sim t_k$.

Por lo tanto: $T \sim t_{n-1}$

7.8.5. Regiones de Rechazo y Valores Críticos

Para un nivel de significación α:

Tipo	Región de Rechazo	Valor Crítico	p-valor
Bilateral	$\|T\| > t_{n-1,1-\alpha/2}$	$t_{n-1,1-\alpha/2}$	$2P(T_{n-1} > \|t_{obs}\|)$

Tipo	Región de Rechazo	Valor Crítico	p-valor
Unilateral derecha	$T > t_{n-1,1-\alpha}$	$t_{n-1,1-\alpha}$	$P(T_{n-1} > t_{obs})$
Unilateral izquierda	$T < -t_{n-1,1-\alpha}$	$-t_{n-1,1-\alpha}$	$P(T_{n-1} < t_{obs})$

7.8.6. Tabla de Cuantiles t de Student

Grados de Libertad	0.900	0.950	0.975	0.990	0.995
$df = 10$	1.37	1.81	2.23	2.76	3.17
$df = 30$	1.31	1.70	2.04	2.46	2.75
$df = \infty$	1.28	1.64	1.96	2.33	2.58

Observación: Conforme aumenta n, la distribución t converge a la normal estándar. Para $n > 30$, son prácticamente equivalentes.

7.8.7. Ejemplo: Una Muestra t (Productor de Harina Revisado)

Hipótesis: Mismas que antes - $H_0 : \mu = 1000$ g - $H_1 : \mu \neq 1000$ g

Muestra: $n = 25$, $\bar{x} = 1002$ g, $s = 4.9$ g

Cálculo:

$$t_{obs} = \frac{1002 - 1000}{4.9/\sqrt{25}} = \frac{2}{0.98} = 2.04$$

Decisión (con $\alpha = 0.05$): - Grados de libertad: $df = 24$ - Valor crítico: $t_{24,0.975} = 2.064$ - Como $|t_{obs}| = 2.04 < 2.064$, **no rechazamos** H_0 - p-valor: $2P(T_{24} > 2.04) \approx 0.0525$

Conclusión: Con estos datos, no tenemos evidencia suficiente (al 5 %) de que el peso medio sea diferente de 1 kg.

Código R: Contraste t de una muestra

Cuadro 7.1 Código R

```
# Datos de pesos de paquetes de harina (en gramos)
datos <- c(1002, 998, 1001, 1003, 1000, 1004, 999, 1005, 1002, 997,
           1001, 1003, 1000, 1002, 998, 1004, 1001, 999, 1003, 1000,
           1002, 1001, 1000, 1002, 998)

# Contraste: H0: mu = 1000 vs H1: mu != 1000
resultado <- t.test(datos, mu = 1000, alternative = "two.sided")

# Ver resultados
print(resultado)
```

```
    One Sample t-test

data:  datos
t = 2, df = 24, p-value = 0.03
alternative hypothesis: true mean is not equal to 1000
95 percent confidence interval:
 1000 1002
sample estimates:
mean of x
     1001
```

Cuadro 7.2 Código R

```
# Extraer componentes
cat("Estadístico t =", round(resultado$statistic, 3), "\n")
```

```
Estadístico t = 2.38
```

Cuadro 7.3 Código R

```
cat("Grados de libertad =", resultado$parameter, "\n")
```

```
Grados de libertad = 24
```

Cuadro 7.4 Código R

```
cat("p-valor =", round(resultado$p.value, 4), "\n")
```

```
p-valor = 0.0257
```

Cuadro 7.5 Código R

```
cat("IC 95% para mu:", round(resultado$conf.int, 2), "\n")
```

```
IC 95% para mu: 1000 1002
```

Cuadro 7.6 Código R

```
# Unilateral derecha: H1: mu > 1000
t.test(datos, mu = 1000, alternative = "greater")
```

```
    One Sample t-test

data:  datos
t = 2, df = 24, p-value = 0.01
alternative hypothesis: true mean is greater than 1000
95 percent confidence interval:
 1000  Inf
sample estimates:
mean of x
     1001
```

Cuadro 7.7 Código R

```
# Unilateral izquierda: H1: mu < 1000
t.test(datos, mu = 1000, alternative = "less")
```

```
    One Sample t-test

data:  datos
t = 2, df = 24, p-value = 1
alternative hypothesis: true mean is less than 1000
95 percent confidence interval:
 -Inf 1002
sample estimates:
mean of x
     1001
```

Ejemplo con BioEstatR: testt() y testp()

Documentación completa de ambas funciones en Sección B.5.

Cuadro 7.8 Código R

```
library(BioEstatR)

# H : μ_HbA1c = 7.5%  (umbral clínico de control glucémico)
# Dataset osteo: 94 pacientes diabéticos, Fac. Medicina UGR
testt(m = osteo$hba1c, m0 = 7.5, grf = FALSE)
```

```
# t-Test con una muestra
# ----------------------

# Resumen de 'osteo$hba1c'
    n = 94.000
    media = 8.565
    d.t. = 1.799
    sem = 0.186

# Estimación de la media  :
  95%-IC( ) = (8.196, 8.933)

# Test de normalidad de Shapiro-Wilk:
  W = 0.983, gl = 94, p = 0.276

# Test de Student para contrastar H : =  con  =7.500
  texp = 5.740, gl = 93
    p < 0.001 para la alternativa bilateral H :
    p < 0.001 para la alternativa unilateral H : >

  Estimación del efecto bruto
  95%-IC( - ) = (0.696, 1.433)
```

Cuadro 7.9 Código R

```
# H :  _tabaco = 0.35  (prevalencia de referencia)
testp(x = sum(osteo$tabaco == "Sí"), n = nrow(osteo), p0 = 0.35)
```

```
# Test para contrastar una proporción binomial
# --------------------------------------------

# Información muestral
  n = 94
  x = 41   n-x=53
```

```
  p = 0.436; q = (1-p) = 0.564

# Test Ho: =0.350
  [1] Método exacto
                      H1  Fexp Valor.p
    Cola derecha  >0.350 1.410   0.052
    Bilateral     0.350     -   0.103

    95%-IC( ) = (0.339, 0.542) (método de Clooper-Pearson)

  [2] Método aproximado a la distribución normal
    Validez: min(n , n(1- )) = 32.9 (>5, el método es válido)
    zexp = 1.643,  p  = 0.100

    95%-IC( ) = (0.335, 0.542) (método de Wilson)
```

`testt()` integra Shapiro-Wilk automáticamente. La HbA1c supera el umbral clínico ($p < 0.001$); la prevalencia de tabaquismo no difiere del 35 % de referencia ($p = 0.100$).

7.9. Contraste para una Proporción

Cuando la variable de interés es dicotómica (sí/no, éxito/fracaso), estamos interesados en la proporción poblacional π.

7.9.1. Función Muestral

Si $X_i \sim \text{Bernoulli}(\pi)$ (cada observación es 1 con probabilidad π, 0 con probabilidad $1-\pi$), entonces el número total de "éxitos" es:

$$X = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, \pi)$$

La proporción muestral es:

$$\hat{\pi} = \frac{X}{n}$$

7.9.2. Distribución bajo H_0

⚠ Resultado Importante: Contraste Binomial

Bajo $H_0 : \pi = \pi_0$:

$$X \sim \text{Binomial}(n, \pi_0)$$

La distribución del número de éxitos es exactamente binomial. Los valores críticos se encuentran usando tablas binomiales o cálculos computacionales.

7.9.3. Aproximación Normal (Muestra Grande)

Para n grande (típicamente $n\pi_0(1-\pi_0) \geq 5$):

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \approx N(0,1)$$

Se pueden usar los valores críticos normales en lugar de los binomiales exactos.

7.9.4. Ejemplo: Moneda Justa

Problema: Probamos si una moneda es justa lanzándola 10 veces.

Hipótesis: - $H_0 : \pi = 0.5$ (moneda justa) - $H_1 : \pi \neq 0.5$ (moneda cargada)

Muestra: $n = 10$, observamos $x = 9$ caras

Distribución binomial bajo H_0: $X \sim B(10, 0.5)$

Tabla de probabilidades acumuladas:

x	$P(X \leq x)$	$P(X = x)$	Región
0	0.0010	0.0010	Rechazo
1	0.0107	0.0097	Rechazo
2	0.0547	0.0440	
...	...	...	No rechazo
8	0.9893	0.0440	
9	0.9990	0.0097	Rechazo
10	1.0000	0.0010	Rechazo

Con $\alpha = 0.05$:

- Región de rechazo bilateral: $\{0, 1\} \cup \{9, 10\}$
- Como $x = 9$ está en la región de rechazo, rechazamos H_0

- p-valor: $P(X \leq 1) + P(X \geq 9) = 0.0107 + (1 - 0.9893) = 0.0214$

Conclusión: La evidencia sugiere que la moneda no es justa (es significativamente sesgada hacia caras).

Código R: Contraste Binomial

Cuadro 7.10 Código R

```
# Contraste binomial exacto
# Lanzamos una moneda 10 veces, observamos 9 caras
binom.test(x = 9, n = 10, p = 0.5, alternative = "two.sided")
```

```
    Exact binomial test

data:  9 and 10
number of successes = 9, number of trials = 10, p-value = 0.02
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.555 0.997
sample estimates:
probability of success
                   0.9
```

Cuadro 7.11 Código R

```
# Aproximación normal para muestra más grande
# Lanzamos 100 veces, observamos 65 caras
prop.test(x = 65, n = 100, p = 0.5, alternative = "two.sided")
```

```
    1-sample proportions test with continuity correction

data:  65 out of 100, null probability 0.5
X-squared = 8, df = 1, p-value = 0.004
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.547 0.741
sample estimates:
   p
0.65
```

Cuadro 7.12 Código R

```
# Para proporción: H0: pi = 0.5 vs H1: pi != 0.5
resultado <- binom.test(9, 10, 0.5)
cat("p-valor =", resultado$p.value, "\n")
```

```
p-valor = 0.0215
```

Cuadro 7.13 Código R

```
cat("Intervalo de confianza:", resultado$conf.int, "\n")
```

```
Intervalo de confianza: 0.555 0.997
```

7.10. Valor-p (p-value)

El valor-p es una de las cantidades más importantes en estadística inferencial.

Definición: Valor-p

El **valor-p** (o **p-value**) es la probabilidad, calculada asumiendo que H_0 es verdadera, de obtener un valor del estadístico de prueba tan o más extremo que el valor observado.
Fórmula general:

$$\text{valor-p} = P(T \text{ sea tan o más extremo que } t_{obs} | H_0)$$

donde "más extremo" depende del tipo de contraste:

- **Bilateral:** $P(|T| \geq |t_{obs}|)$
- **Unilateral derecha:** $P(T \geq t_{obs})$
- **Unilateral izquierda:** $P(T \leq t_{obs})$

7.10.1. Interpretación

El valor-p **no es**:

- La probabilidad de que H_0 sea verdadera
- La probabilidad de que H_1 sea verdadera
- La probabilidad de haber cometido un error

El valor-p **es**:

- Una medida de compatibilidad entre los datos y H_0
- Un p-valor pequeño indica que los datos observados serían muy improbables si H_0 fuera verdadera

7.10.2. Regla de Decisión

Enfoque tradicional con nivel α: - Si p-valor $< \alpha \Rightarrow$ Rechazamos H_0 - Si p-valor $\geq \alpha \Rightarrow$ No rechazamos H_0

Enfoque moderno (más flexible): - Reportar el valor-p exacto y dejar que el lector juzgue - Valores muy pequeños (< 0.01) indican evidencia fuerte contra H_0 - Valores pequeños (0.01 - 0.05) indican evidencia moderada - Valores grandes (> 0.05) indican que los datos son compatibles con H_0

7.10.3. Ejemplo: Cálculo de p-valores

Ejemplo 1 (Normal): - Observamos $z_{obs} = 2.5$ en un contraste bilateral - p-valor $= 2P(Z > 2.5) = 2(0.0062) = 0.0124$

Ejemplo 2 (t-Student): - Observamos $t_{obs} = 2.04$ con 24 gl en un contraste bilateral - p-valor $= 2P(T_{24} > 2.04)$ - Usando R: `2 * (1 - pt(2.04, 24))` 0.0525

Ejemplo 3 (Binomial): - Observamos 9 caras en 10 lanzamientos, contraste bilateral con H_0 : $\pi = 0.5$ - p-valor $= P(X \leq 1) + P(X \geq 9) = 0.0214$

7.11. Tabla Resumen: Regiones de Aceptación y Rechazo

⚠ Tabla de Referencia: Pruebas de una Muestra

Parámetro	Prueba	Estadístico	Distribución	Bilateral	Unilateral Derecha	Unilateral Izquierda
Media (σ conocida)	Z	$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$	$N(0,1)$	Rechazo: $\|Z\| > z_{1-\alpha/2}$	Rechazo: $Z > z_{1-\alpha}$	Rechazo: $Z < -z_{1-\alpha}$
Media (σ desconocida)	t	$T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$	t_{n-1}	Rechazo: $\|T\| > t_{n-1,1-\alpha/2}$	Rechazo: $T > t_{n-1,1-\alpha}$	Rechazo: $T < -t_{n-1,1-\alpha}$
Proporción	Binomial	$X = \sum X_i$	$B(n,\pi_0)$	Rechazo: $P(X \leq x_l) + P(X \geq x_u) < \alpha$	Rechazo: $P(X \geq x) < \alpha$	Rechazo: $P(X \leq x) < \alpha$

Proporción (grande n)	Z	$Z = \frac{\hat{\pi}-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$	$N(0,1)$	Rechazo: $\|Z\| > z_{1-\alpha/2}$	Rechazo: $Z > z_{1-\alpha}$	Rechazo: $Z < -z_{1-\alpha}$

7.12. Ejemplos Completos

7.12.1. Ejemplo 1: Contraste Z Bilateral (Conocida la Varianza)

Problema: El peso de paquetes de cereales debería ser 500 g. Sospechamos que hay variación. Se sabe que $\sigma = 10$ g.

Datos: Muestra de $n = 36$ paquetes, $\bar{x} = 498.5$ g

Hipótesis: $H_0 : \mu = 500$ vs. $H_1 : \mu \neq 500$

Análisis:

$$z_{obs} = \frac{498.5 - 500}{10/\sqrt{36}} = \frac{-1.5}{1.667} = -0.90$$

Con $\alpha = 0.05$: valor crítico $= \pm 1.96$

Decisión: Como $|-0.90| < 1.96$, no rechazamos H_0

p-valor: $2P(Z > 0.90) = 2(0.1841) = 0.3682$

Conclusión: Los datos no proporcionan evidencia de que el peso medio sea diferente de 500 g.

7.12.2. Ejemplo 2: Contraste t Unilateral (Desconocida la Varianza)

Problema: Un fertilizante se supone que aumenta el rendimiento a más de 100 unidades. Probamos si funciona.

Datos: Muestra de $n = 16$ parcelas, $\bar{x} = 105$, $s = 12$

Hipótesis: $H_0 : \mu \leq 100$ vs. $H_1 : \mu > 100$ (unilateral derecha)

Análisis:

$$t_{obs} = \frac{105 - 100}{12/\sqrt{16}} = \frac{5}{3} = 1.67$$

Con $\alpha = 0.05$ y $df = 15$: valor crítico $= t_{15,0.95} = 1.753$

Decisión: Como $t_{obs} = 1.67 < 1.753$, no rechazamos H_0

p-valor: $P(T_{15} > 1.67) \approx 0.058$

Conclusión: La evidencia es marginal. No podemos afirmar al 5 % que el fertilizante aumenta el rendimiento, aunque está cerca.

7.12.3. Ejemplo 3: Contraste Binomial Exacto

Problema: Un fármaco se supone que tiene una tasa de efectos adversos del 5 %. Observamos 15 efectos adversos en 100 pacientes. ¿Es significativamente diferente?

Datos: $n = 100$, $x = 15$, $\pi_0 = 0.05$

Hipótesis: $H_0 : \pi = 0.05$ vs. $H_1 : \pi \neq 0.05$

Aproximación normal:

$$z_{obs} = \frac{0.15 - 0.05}{\sqrt{0.05 \times 0.95/100}} = \frac{0.10}{\sqrt{0.00475}} = \frac{0.10}{0.0689} = 1.45$$

Con $\alpha = 0.05$: valor crítico $= \pm 1.96$

Decisión: Como $|1.45| < 1.96$, no rechazamos H_0

p-valor: $2P(Z > 1.45) = 2(0.0735) = 0.147$

Conclusión: La tasa observada no es significativamente diferente de 0.05.

Código R: Ejemplos Completos

Cuadro 7.14 Código R

```
# Ejemplo 1: Contraste Z (conocida varianza)
# z.test no existe en base R, usamos estadística manual
x_bar <- 498.5
mu_0 <- 500
sigma <- 10
n <- 36
z_obs <- (x_bar - mu_0) / (sigma / sqrt(n))
p_valor <- 2 * (1 - pnorm(abs(z_obs)))
cat("Ejemplo 1:\n")
```

```
Ejemplo 1:
```

Cuadro 7.15 Código R

```
cat("z observado =", round(z_obs, 3), "\n")
```

```
z observado = -0.9
```

Cuadro 7.16 Código R

```
cat("p-valor =", round(p_valor, 4), "\n\n")
```

```
p-valor = 0.368
```

Cuadro 7.17 Código R

```
# Ejemplo 2: Contraste t unilateral
datos2 <- c(115, 108, 102, 110, 98, 104, 112, 96,
            106, 109, 101, 107, 103, 111, 95, 113)
resultado2 <- t.test(datos2, mu = 100, alternative = "greater")
cat("Ejemplo 2:\n")
```

```
Ejemplo 2:
```

Cuadro 7.18 Código R

```
print(resultado2)
```

```
    One Sample t-test

data:  datos2
t = 4, df = 15, p-value = 0.001
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 103 Inf
sample estimates:
mean of x
      106
```

Cuadro 7.19 Código R

```
# Ejemplo 3: Contraste Binomial
resultado3 <- binom.test(x = 15, n = 100, p = 0.05,
                         alternative = "two.sided")
cat("\nEjemplo 3:\n")
```

```
Ejemplo 3:
```

Cuadro 7.20 Código R

```
print(resultado3)
```

```
    Exact binomial test
```

```
data:  15 and 100
number of successes = 15, number of trials = 100, p-value = 1e-04
alternative hypothesis: true probability of success is not equal to 0.05
95 percent confidence interval:
 0.0865 0.2353
sample estimates:
probability of success
                  0.15
```

7.13. Relación entre Hipótesis Unilaterales y Bilaterales

7.13.1. Elección de la Dirección

La dirección del contraste (bilateral vs. unilateral) debe determinarse **antes** de recopilar datos, basándose en la pregunta de investigación:

Bilateral ($H_1 : \theta \neq \theta_0$): - Cuando nos interesa desviaciones en ambas direcciones - Ejemplo: ¿Difiere la media de algún valor? - Más conservador (requiere evidencia más fuerte)

Unilateral derecho ($H_1 : \theta > \theta_0$): - Cuando específicamente buscamos que el parámetro sea mayor - Ejemplo: ¿El nuevo tratamiento mejora el resultado? - Valor crítico es $z_{1-\alpha}$ (no $z_{1-\alpha/2}$)

Unilateral izquierdo ($H_1 : \theta < \theta_0$): - Cuando específicamente buscamos que el parámetro sea menor - Ejemplo: ¿El fármaco reduce los efectos adversos? - Valor crítico es $-z_{1-\alpha}$

7.13.2. Implicaciones para Errores

La elección afecta a α y β:

En contraste **bilateral** con $\alpha = 0.05$:

- Dividimos α en ambas colas: $\alpha/2 = 0.025$ en cada cola
- Valor crítico es más grande ($z_{0.975} = 1.96$)
- Mayor β (menos poder)

En contraste **unilateral** con $\alpha = 0.05$:

- Toda $\alpha = 0.05$ va en una cola
- Valor crítico es más pequeño ($z_{0.95} = 1.645$)
- Menor β (más poder)

Por eso es importante elegir antes: Si eliges unilateral después de ver los datos, estás "p-hacking" y aumentas falsamente el poder.

7.14. Resumen

Conceptos Clave de Semana 8

1. **Hipótesis:** Definimos H_0 (lo que asumimos) y H_1 (lo que queremos demostrar). Deben ser complementarias.
2. **Errores:** Error Tipo I (α, falso positivo) y Error Tipo II (β, falso negativo) son inversamente relacionados. Solo se reduce ambos aumentando n.
3. **Estadístico de Prueba:** Un valor calculado de los datos cuya distribución es conocida bajo H_0.
4. **Región Crítica:** Conjunto de valores del estadístico donde rechazamos H_0, determinado por el nivel α.
5. **Contraste Z:** Cuando σ es conocida. Estadístico $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
6. **Contraste t:** Cuando σ es desconocida (casos prácticos). Estadístico $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$
7. **Contraste Binomial:** Para proporciones. Distribución exacta es $B(n, \pi_0)$; aproximación normal para n grande.
8. **p-valor:** Probabilidad de observar datos tan extremos o más bajo H_0. Si p-valor $< \alpha$, rechazamos H_0.
9. **Decisión:** Comparamos el valor observado del estadístico con el valor crítico (o comparamos p-valor con α).
10. **Potencia:** La capacidad de detectar un efecto real. Mejora con mayor n, mayor diferencia real, o mayor α.

7.14.1. Tabla de Decisión Rápida

Pregunta	Respuesta
¿Se conoce σ?	Sí $\rightarrow$ Contraste Z; No $\rightarrow$ Contraste t
¿Bilateral o unilateral?	Determinar antes de recopilar datos
¿Qué α?	Típicamente 0.05; a veces 0.01
¿Cómo decidir?	Opción 1: Comparar $\|T\|$ con valor crítico; Opción 2: Comparar p-valor con α
¿Qué significa “no rechazar”?	NO significa aceptar H_0; solo significa “no tenemos evidencia suficiente contra H_0”

7.15. Ejercicios

7.15.1. Ejercicio 1: Conceptos Básicos

Un médico sospecha que el nivel medio de colesterol en una población es mayor a 200 mg/dL. Formule las hipótesis apropiadas e identifique los tipos de error que podrían ocurrir.

7.15.2. Ejercicio 2: Contraste Z

Una máquina que produce tornillos tiene un ajuste tal que $\sigma = 0.5$ mm. Se toman 25 tornillos y se mide su diámetro promedio: $\bar{x} = 10.2$ mm. El diámetro especificado es 10 mm.

a) Formule H_0 y H_1
b) Calcule el estadístico Z
c) ¿Rechaza H_0 con $\alpha = 0.05$? (bilateral)
d) Calcule el p-valor

7.15.3. Ejercicio 3: Contraste t

Se obtuvieron las siguientes mediciones de tiempo de reacción (en segundos) para 10 sujetos:

`0.42, 0.45, 0.38, 0.41, 0.39, 0.44, 0.40, 0.43, 0.37, 0.46`

a) ¿Es diferente de 0.40 segundos con $\alpha = 0.05$? (bilateral)
b) Calcule $\bar{x}$ y s
c) Calcule el estadístico t
d) Determine el p-valor

Código R sugerido:

Cuadro 7.21 Código R

```r
datos <- c(0.42, 0.45, 0.38, 0.41, 0.39, 0.44, 0.40, 0.43, 0.37, 0.46)
resultado <- t.test(datos, mu = 0.40)
print(resultado)
```

```
    One Sample t-test

data:  datos
t = 2, df = 9, p-value = 0.2
alternative hypothesis: true mean is not equal to 0.4
95 percent confidence interval:
 0.393 0.437
sample estimates:
mean of x
    0.415
```

7.15.4. Ejercicio 4: Contraste Binomial

Un casino afirma que su dado es justo. Lanzamos 20 veces y obtenemos “6” en 6 ocasiones.

a) Formule H_0 y H_1
b) ¿Es inusual este resultado? Calcule el p-valor
c) ¿Rechazaría que el dado es justo con $\alpha = 0.05$?

Código R sugerido:

Cuadro 7.22 Código R

```
binom.test(x = 6, n = 20, p = 1/6, alternative = "two.sided")
```

```
    Exact binomial test

data:  6 and 20
number of successes = 6, number of trials = 20, p-value = 0.1
alternative hypothesis: true probability of success is not equal to 0.167
95 percent confidence interval:
 0.119 0.543
sample estimates:
probability of success
                   0.3
```

7.15.5. Ejercicio 5: Decisión con p-valor

Para cada uno de los siguientes p-valores, indique si rechaza H_0 con $\alpha = 0.05$:

a) p-valor = 0.031
b) p-valor = 0.052
c) p-valor = 0.0001
d) p-valor = 0.450

7.15.6. Ejercicio 6: Interpretación de Resultados

Un estudio investiga si el peso medio de una población difiere de 70 kg. Se obtiene t = 1.85 con 24 grados de libertad y p-valor = 0.078.

a) ¿Rechaza H_0 con $\alpha = 0.05$?
b) ¿Rechazaría con $\alpha = 0.10$?
c) ¿Qué puede concluir de estos resultados?
d) ¿Por qué no es correcto decir "se acepta H_0"?

7.16. Respuestas a los Ejercicios

Ejercicio 1: - H : = 200 mg/dL - H : > 200 mg/dL (unilateral derecha) - Error Tipo I: rechazar H siendo verdadera (afirmar que colesterol > 200 cuando realmente = 200) - Error Tipo II: no rechazar H siendo falsa (concluir que colesterol = 200 cuando realmente > 200)

Ejercicio 2: - a) H : = 10 mm, H : 10 mm - b) Z = (10.2 - 10)/(0.5/√25) = 0.2/0.1 = 2.0 - c) Z . = 1.96; |2.0| > 1.96, rechaza H - d) p-valor 2×P(Z > 2) 2×0.0228 = 0.0456

Ejercicio 3: - b) $\bar{X}$ 0.415, s 0.0318 - c) t = (0.415 - 0.40)/(0.0318/√10) = 0.015/0.0101 1.49 - d) p-valor 0.17 (df=9); no rechaza H con =0.05

Ejercicio 4: - a) H : p = 1/6, H : p 1/6 - b) Bajo H , X ~ Binomial(20, 1/6); P(X=6) 0.20 aproximadamente - c) p-valor > 0.05; no rechaza que el dado sea justo

Ejercicio 5: - a) p = 0.031 < 0.05: Rechaza H - b) p = 0.052 > 0.05: No rechaza - c) p = 0.0001 < 0.05: Rechaza H - d) p = 0.450 > 0.05: No rechaza

Ejercicio 6: - a) p-valor = 0.078 > 0.05: No rechaza H - b) p-valor = 0.078 < 0.10: Rechaza H con =0.10 - c) Evidencia marginal contra H ; resultados borderline - d) Porque no rechazar H no significa que sea verdadera, solo que no hay evidencia suficiente en los datos

---

Fin de Semana 8: Contrastes de Hipótesis (Parte I)

En las próximas semanas estudiaremos:

- Contrastes para diferencia de medias (dos muestras)
- Contrastes para varianzas
- ANOVA (análisis de varianza)
- Contrastes para independencia

Capítulo 8

Semana 8 — Contrastes de Hipótesis (Parte II)

En esta semana continuamos con contrastes de hipótesis, enfocándonos en pruebas para comparar **dos muestras independientes y pareadas**, pruebas para igualdad de varianzas, y conceptos fundamentales de **potencia estadística** y cálculo de tamaño muestral.

8.1. Prueba Z para Diferencia de Medias (Varianzas Conocidas)

8.1.1. Preparación: Distribución de la Diferencia de Medias

Supongamos que tenemos dos poblaciones independientes:

- Población 1: $X_1 \sim N(\mu_1, \sigma_1^2)$ con σ_1^2 **conocida**
- Población 2: $X_2 \sim N(\mu_2, \sigma_2^2)$ con σ_2^2 **conocida**

Tomamos muestras de tamaños n_1 y n_2 respectivamente. Los estimadores de las medias son:

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

La diferencia entre las medias muestrales sigue una distribución normal:

$$D = \bar{X}_1 - \bar{X}_2 \sim N\left(\omega, \sigma_D^2\right)$$

donde:

$$\omega = \mu_1 - \mu_2, \quad \sigma_D^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

8.1.2. Hipótesis e Estadîstico de Prueba

Definición: Prueba Z para Dos Muestras

Consideramos las hipótesis:

- **Bilateral:** $H_0 : \mu_1 - \mu_2 = \omega_0$ vs. $H_1 : \mu_1 - \mu_2 \neq \omega_0$
- **Unilateral derecha:** $H_0 : \mu_1 - \mu_2 \leq \omega_0$ vs. $H_1 : \mu_1 - \mu_2 > \omega_0$
- **Unilateral izquierda:** $H_0 : \mu_1 - \mu_2 \geq \omega_0$ vs. $H_1 : \mu_1 - \mu_2 < \omega_0$

(Habitualmente, $\omega_0 = 0$)

Estadîstico de Prueba: Z

Bajo H_0 (asumiendo $\omega = \omega_0$):

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \omega_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Criterios de rechazo: - **Bilateral:** Rechazo si $|Z| > z_{1-\alpha/2}$ - **Unilateral derecha:** Rechazo si $Z > z_{1-\alpha}$ - **Unilateral izquierda:** Rechazo si $Z < -z_{1-\alpha}$

8.2. Prueba t para Dos Muestras Independientes (Varianzas Desconocidas)

8.2.1. Caso 1: Varianzas Iguales (Homogéneas)

Cuando no conocemos las varianzas pero **asumimos que son iguales** ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), utilizamos un estimador combinado (pooled) de la varianza:

Varianza Combinada (Pooled)

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

donde S_1^2 y S_2^2 son las varianzas muestrales sesgadas:

$$S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ji} - \bar{X}_j)^2, \quad j = 1, 2$$

El estimador de la varianza de la diferencia es:

$$S_D^2 = S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Estadístico t (Varianzas Homogéneas)

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \omega_0}{S_D} \sim t_{n_1+n_2-2}$$

bajo $H_0 : \mu_1 - \mu_2 = \omega_0$.
Grados de libertad: $\nu = n_1 + n_2 - 2$
Criterios de rechazo: - **Bilateral:** Rechazo si $|T| > t_{1-\alpha/2;\nu}$ - **Unilateral derecha:** Rechazo si $T > t_{1-\alpha;\nu}$ - **Unilateral izquierda:** Rechazo si $T < -t_{1-\alpha;\nu}$

8.2.2. Caso 2: Varianzas Desiguales (Heterogéneas) — Prueba de Welch

Cuando las varianzas poblacionales son **desiguales** ($\sigma_1^2 \neq \sigma_2^2$), la solución es aproximada.

Estadístico de Welch-Satterthwaite

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \omega_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx t_\nu$$

donde los **grados de libertad aproximados** (Welch-Satterthwaite) son:

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}$$

Este test es más conservador que asumir varianzas iguales.

Ejemplo con BioEstatR: testt(), testf() y testwx()

Documentacion completa de las tres funciones en Sección B.5.

Cuadro 8.1 Código R

```
library(BioEstatR)

# t-test dos muestras independientes: HbA1c por sexo
testt(grupos = osteo$sexo, m = osteo$hba1c, grf = FALSE)
```

```
# t-test para 2 Muestras Independientes
# -------------------------------------

# Información muestral y estimación de las medias
  Niveles de agrupación: Hombre, Mujer
```

```
                        n media    dt   sem             IC
  osteo$hba1c [Hombre] 45  8.55 1.745 0.260 (8.029, 9.078)
  osteo$hba1c [Mujer]  49  8.58 1.864 0.266  (8.04, 9.111)

  ----
  * IC elaborados al 95% de confianza para estimar   y   respectivamente


# Pruebas de normalidad (test de Shapiro-Wilk)
[1] Para grupo = Hombre, W = 0.958, gl = 45, p = 0.102
[2] Para grupo = Mujer, W = 0.980, gl = 49, p = 0.566

# Test de homogeneidad de varianzas. Fexp = (var /var )
  Fexp = 1.141, gl  = 48, gl  = 44, p = 0.660

# Diferencia de medias (osteo$hba1c [Mujer] - osteo$hba1c [Hombre])
  Hipótesis a contrastar: H : =  ( - =0)

  a) Test de Student (varianzas homogéneas)
  texp = 0.059, gl = 92
    p = 0.953 para la alternativa bilateral H :
    p = 0.476 para la alternativa unilateral H : <
  95%-IC( - ) = (-0.719, 0.764)

  b) Test de Welch (varianzas no homogéneas)
  texp = 0.060, gl = 91.96
    p = 0.953 para la alternativa bilateral H :
    p = 0.476 para la alternativa unilateral H : <
  95%-IC( - ) = (-0.717, 0.762)
```

Cuadro 8.2 Código R

```r
# Test F de varianzas del IMC por sexo
testf(s1 = sd(osteo$imc[osteo$sexo == "Hombre"]),
      n1 = sum(osteo$sexo == "Hombre"),
      s2 = sd(osteo$imc[osteo$sexo == "Mujer"]),
      n2 = sum(osteo$sexo == "Mujer"))
```

```
[[1]]
[1] 2.31

[[2]]
[1] 48
```

```
[[3]]
[1] 44

[[4]]
[1] 0.00585
```

Cuadro 8.3 Código R

```
# Wilcoxon-Mann-Whitney: IMC < 25 años vs > 33 años (no normales)
testwx(m1 = osteo$imc[osteo$grupo_edad == "< 25"],
       m2 = osteo$imc[osteo$grupo_edad == "> 33"],
       grf = FALSE)
```

```
Test de Wilcoxon/Mann-Whithney para dos muestras independientes
----------------------------------------------------------------
# Información muestral ---

                                Muestra  n    min     Q1     Q2     Q3    max
1 osteo$imc[osteo$grupo_edad == "< 25"] 32 19.628 21.421 22.845 24.149 32.475
2 osteo$imc[osteo$grupo_edad == "> 33"] 30 18.070 23.382 25.889 28.945 37.333
    RIQ
1 2.728
2 5.563

# Rangos ---

                                Muestra  n Suma_rangos Rango_medio       U
1 osteo$imc[osteo$grupo_edad == "< 25"] 32         807      25.219 681.000
2 osteo$imc[osteo$grupo_edad == "> 33"] 30        1146      38.200 279.000

# Test ---

  U = 279.000; Z = 2.831; W = 279.000; p = 0.004

# Tamaño del efecto ---

  Diferencia de localización: -2.668    95%-IC = (-4.518, -0.778)
  r = 0.360  (criterio: 0.1 pequeño; 0.3 mediano; >0.5 grande)
  Probabilidad de superioridad PS = 0.709
  (probabilidad de que un valor al azar de M1 sea < a un valor al azar de M2)
```

8.3. Prueba t Pareada

8.3.1. Motivación y Formulación

En muchos estudios, las observaciones de dos muestras están **apareadas** (matched):

- Medidas antes y después en los mismos sujetos
- Observaciones de hermanos gemelos
- Medidas en ambos ojos del mismo paciente

Definición: Datos Pareados

Para datos pareados (X_{1i}, X_{2i}), $i = 1, \ldots, n$, definimos las diferencias:

$$D_i = X_{1i} - X_{2i}$$

El test pareado **reduce el problema a una prueba t de una muestra** sobre las diferencias:

$$H_0 : \mu_D = \mu_1 - \mu_2 = \omega_0 \quad \text{vs.} \quad H_1 : \mu_D \neq \omega_0$$

(Típicamente, $\omega_0 = 0$)

Estadístico t Pareado

$$T = \frac{\bar{D} - \omega_0}{S_D/\sqrt{n}} \sim t_{n-1}$$

donde:

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i, \quad S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$$

Nota importante: La varianza de las diferencias es típicamente **menor** que la suma de varianzas, lo que da mayor potencia al test pareado cuando los datos están realmente correlacionados.

8.4. Prueba F para Igualdad de Varianzas

8.4.1. Hipótesis y Estadístico

En algunos análisis necesitamos probar si **dos varianzas poblacionales son iguales**.

Definición: Prueba F de Fisher

Consideramos:

- **Bilateral:** $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_1 : \sigma_1^2 \neq \sigma_2^2$
- **Unilateral derecha:** $H_0 : \sigma_1^2 \leq \sigma_2^2$ vs. $H_1 : \sigma_1^2 > \sigma_2^2$
- **Unilateral izquierda:** $H_0 : \sigma_1^2 \geq \sigma_2^2$ vs. $H_1 : \sigma_1^2 < \sigma_2^2$

Estadístico F

Bajo H_0 ($\sigma_1^2 = \sigma_2^2$):

$$F = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}$$

Criterios de rechazo: - **Bilateral:** Rechazo si $F < F_{n_1-1,n_2-1,\alpha/2}$ o $F > F_{n_1-1,n_2-1,1-\alpha/2}$ - **Unilateral derecha:** Rechazo si $F > F_{n_1-1,n_2-1,1-\alpha}$ - **Unilateral izquierda:** Rechazo si $F < F_{n_1-1,n_2-1,\alpha}$

Advertencia: Testing Secuencial

No recomendamos realizar la prueba F antes de elegir entre test t con varianzas iguales o desiguales, porque esta prueba de igualdad consume el nivel de significancia global. Es preferir usar siempre la **Prueba de Welch** como default, que es válida en ambos casos.

8.5. Potencia de una Prueba

8.5.1. Conceptos Fundamentales

La **potencia** es un concepto crucial en el diseño y evaluación de estudios estadísticos.

Definición: Función de Potencia

La **función de potencia** $G(\theta)$ es la probabilidad de rechazar H_0 cuando el parámetro verdadero es θ:

$$G(\theta) = P(\text{rechazar } H_0 | \theta)$$

Interpretación: - Si $\theta \in \Theta_0$ (región de H_0): $G(\theta) = P(\text{Error Tipo I}) \leq \alpha$ - Si $\theta \in \Theta_1$ (región de H_1): $G(\theta) = P(\text{Rechazo correcto}) = 1 - \beta(\theta)$
donde $\beta(\theta)$ es la probabilidad del **Error Tipo II**.

8.5.2. Relación con Errores de Tipo I y II

Realidad	Rechazamos H_0	No rechazamos H_0
H_0 es verdadera	Error Tipo I (prob. α)	Decisión correcta
H_1 es verdadera	Decisión correcta (prob. $1-\beta$)	Error Tipo II (prob. β)

- **Significancia (α):** Probabilidad de Error Tipo I. Típicamente 0.05.
- **Potencia ($1-\beta$):** Probabilidad de detectar una diferencia real. Típicamente 0.80 o 0.90.

8.5.3. Derivación: Potencia para Prueba Bilateral (Varianza Conocida)

⚠ Potencia: Prueba Bilateral

Sea μ_0 el valor hipotetizado, σ conocida, tamaño muestral n y nivel de significancia α.
Para una prueba bilateral con $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$:

$$G(\mu) = 1 - P\left(z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}} \leq Z \leq z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right)$$

Equivalentemente:

$$G(\mu) = P\left(Z \leq -z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right) + P\left(Z > z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right)$$

Casos extremos: - En $\mu = \mu_0$: $G(\mu_0) = \alpha$ (tasa de Error Tipo I) - Conforme $|\mu-\mu_0|$ aumenta: $G(\mu) \to 1$ (potencia tiende a 1)

8.5.4. Potencia: Prueba Unilateral Derecha

⚠ Potencia: Prueba Unilateral Derecha

Para $H_0 : \mu \leq \mu_0$ vs. $H_1 : \mu > \mu_0$:

$$G(\mu) = 1 - P\left(Z \leq z_{1-\alpha} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right)$$

La potencia aumenta cuando $\mu > \mu_0$, alcanzando su máximo cuando μ es mucho mayor que μ_0.

8.5.5. Potencia: Prueba Unilateral Izquierda

⚠ Potencia: Prueba Unilateral Izquierda

Para $H_0 : \mu \geq \mu_0$ vs. $H_1 : \mu < \mu_0$:

$$G(\mu) = P\left(Z \leq -z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)$$

La potencia aumenta cuando $\mu < \mu_0$, alcanzando su máximo cuando μ es mucho menor que μ_0.

8.5.6. Factores que Afectan la Potencia

La potencia de una prueba depende de:

1. **Tamaño muestral (n):** Aumentar n aumenta la potencia.
2. **Magnitud del efecto ($|\mu - \mu_0|$ o δ):** Diferencias más grandes son más fáciles de detectar.
3. **Variabilidad (σ):** Menor variabilidad aumenta la potencia.
4. **Nivel de significancia (α):** Aumentar α aumenta la potencia, pero incrementa el Error Tipo I.

8.6. Potencia para Pruebas de Proporción

8.6.1. Formulación

Para una prueba binomial sobre proporción π:

ℹ Potencia: Prueba de Proporción

Sea π_0 la proporción hipotetizada, n el tamaño muestral, α el nivel de significancia.
Para una prueba bilateral con $H_0 : \pi = \pi_0$ vs. $H_1 : \pi \neq \pi_0$:

$$G(\pi) = P(X \leq x_l|\pi) + P(X > x_u|\pi)$$

donde x_l y x_u son los valores crîticos determinados por el nivel de significancia α, y $X \sim$ Binomial(n, π).

- **En $\pi = \pi_0$:** $G(\pi_0) = \alpha$ (tasa de Error Tipo I)
- **Cuando π se aleja de π_0:** La potencia $G(\pi)$ aumenta

Ejemplo: Proporción de Exitos en Prueba Binomial

Supongamos $n = 10$, $\pi_0 = 0.5$, $\alpha = 0.05$. Los valores crîticos son $x_l = 2$ y $x_u = 8$.
Potencia en varios valores de π:

π	$G(\pi)$	$\beta = 1 - G(\pi)$
0.20	0.3758	0.6242
0.30	0.1495	0.8505
0.40	0.0480	0.9520
0.50	0.0214	0.9786 (Error Tipo II)
0.60	0.0480	0.9520
0.70	0.1495	0.8505

La potencia es máxima en los extremos de la distribución (π cercano a 0 o 1). Esto representa la dificultad de detectar desviaciones pequeñas de $\pi_0 = 0.5$ debido a la máxima variabilidad en ese punto (varianza máxima $p(1-p) = 0.25$). En contraste, para π cerca de 0 o 1, la variabilidad es menor, lo que facilita la detección de diferencias.

8.7. Cálculo de Tamaño Muestral

8.7.1. Principio General

El tamaño muestral se calcula especificando:

Elementos para Calcular Tamaño Muestral

1. **Diferencia mînima relevante (δ):** Qué magnitud de efecto queremos detectar
2. **Nivel de significancia (α):** Tîpicamente 0.05
3. **Potencia deseada ($1 - \beta$):** Tîpicamente 0.80 o 0.90
4. **Desviación estándar (σ):** De estudios previos o piloto

8.7.2. Fórmula para Prueba t de Dos Muestras (Varianzas Iguales)

Tamaño Muestral: Prueba t Bilateral

Para detectar una diferencia $\delta = \mu_1 - \mu_2$ entre dos muestras de igual tamaño:

$$n = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2}$$

donde:

- $z_{1-\alpha/2}$ es el quantil normal para nivel de significancia α
- $z_{1-\beta}$ es el quantil normal para potencia $1 - \beta$

- σ^2 es la varianza (asumida igual en ambas muestras)

Ejemplo: Si $\delta = 10$, $\sigma = 15$, $\alpha = 0.05$, $1 - \beta = 0.80$:

$$z_{1-0.025} = 1.96, \quad z_{0.80} = 0.84$$

$$n = \frac{2 \times 225 \times (1.96 + 0.84)^2}{100} = \frac{450 \times 6.4}{100} \approx 28.8 \approx 29 \text{ por grupo}$$

8.7.3. Cálculo con BioEstatR: `testt()` y `testp()`

Las funciones `testt()` y `testp()` del paquete `BioEstatR` (ver Sección B.5) calculan **simultáneamente** el contraste de hipótesis, la **potencia** y el **tamaño muestral** necesario para detectar una diferencia mínima relevante δ (effect size en unidades de la variable).

Argumentos clave

Argumento	Significado	Función
`delta`	Diferencia mínima clínicamente relevante a detectar	`testt`, `testp`
`potencia`	Potencia objetivo $1 - \beta$ (típicamente 0.80 o 0.90)	`testt`, `testp`
`alfa`	Nivel de significancia α (por defecto 0.05)	`testt`, `testp`
`m`, `s`, `n`	Media, DT y tamaño piloto (estimaciones previas)	`testt`
`x`, `n`, `p0`	Exitos, tamaño piloto y proporción hipotética	`testp`

Cuando se proporciona `delta` y `potencia`, la función devuelve el **n necesario**. Si en cambio se fija `n` y `delta`, devuelve la **potencia alcanzada**.

Ejemplo: Potencia y tamaño muestral en contexto clínico

Dos escenarios reales del estudio piloto `osteo` (94 pacientes diabéticos):

- **Contraste de medias** (`testt`): ¿Cuántos pacientes necesitamos para detectar una diferencia de $\delta = 0.5\,\%$ en HbA1c respecto al umbral clínico de $7.5\,\%$ con potencia del $90\,\%$?
- **Contraste de proporciones** (`testp`): ¿Y para detectar una desviación de $\delta = 0.05$ frente a una prevalencia hipotética de tabaquismo del $35\,\%$ con potencia del $90\,\%$?

Cuadro 8.4 Código R

```
library(BioEstatR)

# --- 1. testt(): potencia y n para una media ---
# H0: μ_HbA1c = 7.5  ;  delta clinicamente relevante = 0.5%
# Usamos m, s del estudio piloto como referencia
testt(m = osteo$hba1c, m0 = 7.5,
      delta = 0.5, potencia = 0.90, alfa = 0.05, grf = FALSE)
```

```
# t-Test con una muestra
# ----------------------

# Resumen de 'osteo$hba1c'
    n = 94.000
    media = 8.565
    d.t. = 1.799
    sem = 0.186

# Estimación de la media  :
  95%-IC( ) = (8.196, 8.933)

# Test de normalidad de Shapiro-Wilk:
  W = 0.983, gl = 94, p = 0.276

# Test de Student para contrastar H : =  con  =7.500
  texp = 5.740, gl = 93
    p < 0.001 para la alternativa bilateral H :
    p < 0.001 para la alternativa unilateral H : >

  Estimación del efecto bruto
  95%-IC( - ) = (0.696, 1.433)

# Estudio de la potencia:
  El test es significativo, se omite el análisis de la potencia.
```

Cuadro 8.5 Código R

```
# --- 2. testp(): potencia y n para una proporcion ---
# H0: pi_tabaco = 0.35  ;  delta = 0.05 (5 puntos porcentuales)
testp(x = 41, n = 94,
      p0 = 0.35, delta = 0.05, potencia = 0.90, alfa = 0.05)
```

```
# Test para contrastar una proporción binomial
# ---------------------------------------------

# Información muestral
  n = 94
  x = 41   n-x=53
  p = 0.436; q = (1-p) = 0.564

# Test Ho: =0.350
  [1] Método exacto
                      H1  Fexp Valor.p
    Cola derecha  >0.350 1.410   0.052
    Bilateral     0.350     -   0.103

    95%-IC( ) = (0.339, 0.542) (método de Clooper-Pearson)

  [2] Método aproximado a la distribución normal
    Validez: min(n , n(1- )) = 32.9 (>5, el método es válido)
    zexp = 1.643,  p  = 0.100

    95%-IC( ) = (0.335, 0.542) (método de Wilson)


# Tamaño de muestra para detectar | -p |=
  p  = 0.350;    = 0.05
  Casos (cpc = corrección por continuidad):
                             p1   n n_cpc
  1 Unilateral  <0.350   0.300 753   773
  2 Unilateral  >0.350   0.400 753   773
  3 Bilateral   0.350    0.400 977   997
```

Interpretación

Los argumentos `delta` y `potencia` transforman un contraste de hipótesis estático (sólo p-valor) en un **análisis prospectivo de diseño**: cuantifican la sensibilidad del estudio para detectar efectos clínicamente relevantes. Para la **HbA1c**, el contraste resulta altamente significativo ($t_{obs} = 5.74$, gl $= 93$, $p < 0.001$) por lo que `testt()` omite el análisis de potencia (la sensibilidad ya se confirma de facto sobre los $n = 94$ pacientes actuales). Una desviación de medio punto porcentual respecto al umbral clínico (7.5 %) corresponde a ~10–15 mg/dL de glucemia media, magnitud clínicamente relevante para decisiones terapéuticas. Para la **prevalencia de tabaquismo**, alcanzar potencia del 90 % para detectar una desviación de 5

puntos porcentuales respecto al 35 % hipotético exige **n 977 pacientes** (bilateral, ~997 con corrección por continuidad) o 753 (unilateral), aproximadamente diez veces el tamaño piloto. Esta asimetría refleja una propiedad clave: detectar diferencias en proporciones cercanas a 0.5 (varianza máxima $p(1-p) = 0.25$) consume más muestra que detectar diferencias en variables continuas con buen contraste señal/ruido. En el diseño de protocolos clínicos, este análisis previo evita el error de planificar estudios incapaces de detectar el efecto buscado (estudios negativos por falta de potencia, no por ausencia real de efecto), y constituye argumento obligado ante comités éticos y agencias financiadoras.

8.8. Comparación Múltiple y Corrección de Bonferroni

8.8.1. Problema: Inflación de Tasa de Error Tipo I

Problema: Múltiples Pruebas

Cuando realizamos **múltiples pruebas de hipótesis independientes** con nivel α:
Si efectuamos k pruebas, la probabilidad de al menos un Error Tipo I es:

$$P(\text{al menos un Error Tipo I}) = 1 - (1-\alpha)^k$$

Ejemplo: Con $k = 10$ pruebas y $\alpha = 0.05$:

$$P(\text{al menos un Error Tipo I}) = 1 - (0.95)^{10} \approx 0.401 = 40.1\,\%$$

En lugar del nivel de significancia global deseado del 5 %, obtenemos una tasa de Error Tipo I del 40 %!

8.8.2. Corrección de Bonferroni

Corrección de Bonferroni

Para mantener una tasa global de Error Tipo I de α cuando realizamos k pruebas:

$$\alpha_{\text{ajustado}} = \frac{\alpha}{k}$$

Es decir, cada prueba individual se realiza con nivel de significancia α/k.
Ventajas: - Simple de implementar - Conservador (garantiza control del Error Tipo I)
Desventajas: - Muy conservador cuando k es grande - Reduce la potencia de cada prueba individual
Alternativas: El método de Holm, la corrección FDR (False Discovery Rate), o métodos

Bayesianos son menos conservadores.

8.9. Ejemplo Completo: Análisis de Datos de Mosquitos

8.9.1. Contexto del Problema

Según la Organización Mundial de la Salud, los mosquitos Aedes pueden transmitir enfermedades como dengue y Zika. Un esfuerzo de control consiste en liberar mosquitos transgénicos con esperanza de vida más corta.

Pregunta de investigación: ¿Es la esperanza de vida significativamente diferente entre mosquitos salvajes (wildtype) y transgénicos?

8.9.2. Datos y Exploración

Cuadro 8.6 Código R

```r
# Lectura de datos
# Crear datos directamente en R (Tipo de mosquito y vida en días)
wildtype <- c(28, 32, 25, 31, 29, 26, 33, 30, 27, 34, 28, 31, 29, 32, 25, 30, 33, 28, 31, 27)
transgenic <- c(35, 38, 42, 39, 41, 37, 44, 36, 40, 43, 38, 45, 39, 42, 35, 41, 46, 40, 38, 43)

# Combina en un data.frame
mosquitoes <- data.frame(wildtype, transgenic)

# Extracción de muestras
wildtype <- na.omit(mosquitoes[, 1])
transgenic <- na.omit(mosquitoes[, 2])

n1 <- length(wildtype)
n2 <- length(transgenic)

cat("Wildtype: n =", n1, ", media =", mean(wildtype), ", sd =", sd(wildtype), "\n")
```

```
Wildtype: n = 20 , media = 29.4 , sd = 2.68
```

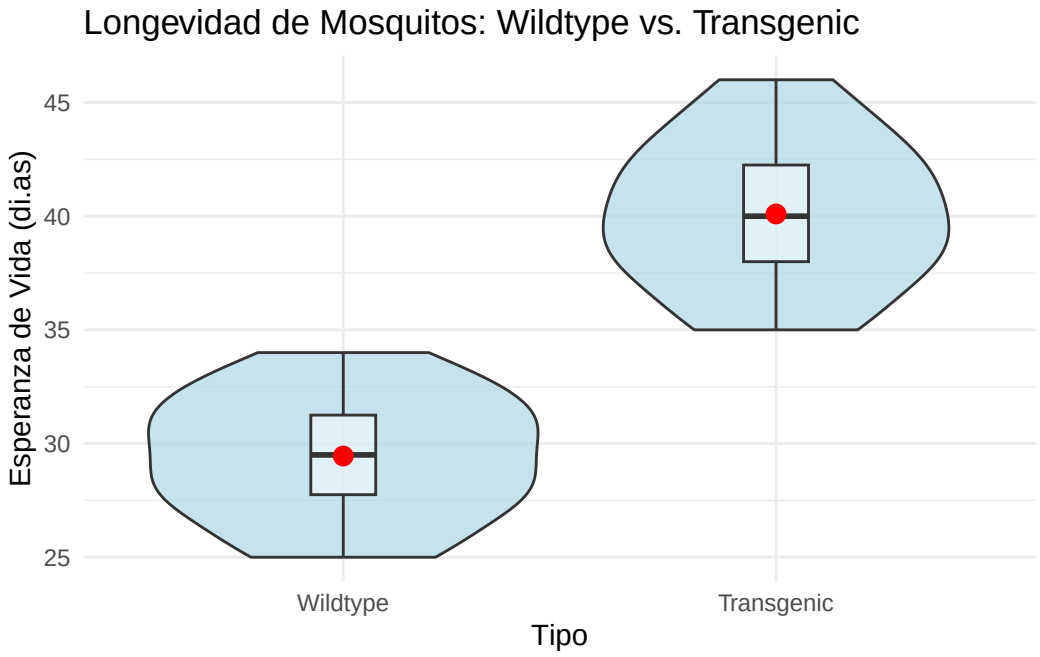
Cuadro 8.7 Código R

```r
cat("Transgenic: n =", n2, ", media =", mean(transgenic), ", sd =", sd(transgenic), "\n")
```

```
Transgenic: n = 20 , media = 40.1 , sd = 3.19
```

Cuadro 8.8 Código R

```
# Gráfico de violín con ggplot2
library(ggplot2)
library(reshape2)

df <- melt(data.frame(Wildtype = wildtype, Transgenic = transgenic))
colnames(df) <- c("Group", "Lifespan")

ggplot(df, aes(x = Group, y = Lifespan)) +
  geom_violin(fill = "lightblue", alpha = 0.7) +
  geom_boxplot(width = 0.15, alpha = 0.5) +
  stat_summary(fun = mean, geom = "point", color = "red", size = 3) +
  theme_minimal() +
  labs(title = "Longevidad de Mosquitos: Wildtype vs. Transgenic",
       x = "Tipo", y = "Esperanza de Vida (días)")
```



8.9.3. Test t: Varianzas Homogéneas

Primero suponemos varianzas iguales:

```
    Two Sample t-test

data:  wildtype and transgenic
t = -11, df = 38, p-value = 0.00000000000008
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.54  -8.76
```

Cuadro 8.9 Código R

```
options(scipen = 999) # Evitar notación científica)
# Test t con varianzas iguales
tt_equal <- t.test(wildtype, transgenic,
                   alternative = "two.sided",
                   var.equal = TRUE,
                   conf.level = 0.95)

print(tt_equal)
```

```
sample estimates:
mean of x mean of y
     29.4      40.1
```

Cuadro 8.10 Código R

```
# Interpretación:
# Si p-value < 0.05, rechazamos H_0 (las medias son significativamente diferentes)
# Si p-value >= 0.05, no hay evidencia de diferencia significativa
```

8.9.4. Test t: Varianzas Heterogéneas (Welch)

Si observamos varianzas muy diferentes, usamos la prueba de Welch:

Cuadro 8.11 Código R

```
# Test t con varianzas desiguales (Welch)
tt_welch <- t.test(wildtype, transgenic,
                   alternative = "two.sided",
                   var.equal = FALSE,
                   conf.level = 0.95)

print(tt_welch)
```

```
    Welch Two Sample t-test

data:  wildtype and transgenic
t = -11, df = 37, p-value = 0.0000000000001
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.54  -8.76
sample estimates:
mean of x mean of y
     29.4      40.1
```

8.9.5. Test F: Igualdad de Varianzas

Podemos probar formalmente si las varianzas son iguales:

Cuadro 8.12 Código R

```
# Test F para igualdad de varianzas
f_test <- var.test(wildtype, transgenic, alternative = "two.sided")

print(f_test)
```

```
    F test to compare two variances

data:  wildtype and transgenic
F = 0.7, num df = 19, denom df = 19, p-value = 0.5
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.28 1.79
sample estimates:
ratio of variances
             0.707
```

Cuadro 8.13 Código R

```
# Interpretación:
# Si p-value < 0.05, hay evidencia de que las varianzas son desiguales
#   → Usar Welch test
# Si p-value >= 0.05, no hay evidencia de desigualdad de varianzas
#   → Usar test t con varianzas iguales
```

8.9.6. Cálculo de Potencia

Si el estudio fuera diseñado ahora, ¿qué tamaño muestral necesitaríamos?

```
Tamaño muestral requerido por grupo: 64
```

```
Potencia del estudio observado: 1
```

8.10. Tabla Resumen: Contrastes de Dos Muestras

Escenario	Estadístico	Distribución	Grados de Libertad	Fórmula
Z-test (varianzas conocidas)	$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \omega_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$	$N(0,1)$	∞	Conocidas

Escenario	Estadîstico	Distribución	Grados de Libertad	Fórmula
t-test (varianzas iguales)	$T = \frac{(\bar{X}_1 - \bar{X}_2) - \omega_0}{S_p\sqrt{1/n_1 + 1/n_2}}$	t	$n_1 + n_2 - 2$	Estimadas
Welch (varianzas desiguales)	$T = \frac{(\bar{X}_1 - \bar{X}_2) - \omega_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$	t (aprox.)	W-S formula	Estimadas
t pareado	$T = \frac{\bar{D} - \omega_0}{S_D/\sqrt{n}}$	t	$n - 1$	Diferencias
F-test	$F = \frac{S_1^2}{S_2^2}$	F	$(n_1 - 1, n_2 - 1)$	Varianzas

---

8.11. Resumen

8.11.1. Conceptos Clave

- **Prueba Z de dos muestras:** Para varianzas conocidas, estadîstico sigue $N(0, 1)$
- **Prueba t de dos muestras (varianzas iguales):** Varianza combinada (pooled), grados de libertad $= n_1 + n_2 - 2$
- **Prueba t de Welch:** Para varianzas desiguales, usa grados de libertad Welch-Satterthwaite
- **Prueba t pareada:** Reduce a una prueba t de una muestra sobre diferencias
- **Prueba F:** Para comparar varianzas, estadîstico $F = S_1^2/S_2^2 \sim F_{n_1-1, n_2-1}$

8.11.2. Potencia Estadîstica

- **Potencia $(1 - \beta)$:** Probabilidad de rechazar H_0 cuando H_1 es verdadera (detectar un efecto real)
- **Error Tipo I (α):** Probabilidad de rechazar H_0 cuando es verdadera (tîpicamente 0.05)
- **Error Tipo II (β):** Probabilidad de no rechazar H_0 cuando H_1 es verdadera (tîpicamente 0.10-0.20)
- **Factores que aumentan potencia:**
  - Mayor tamaño muestral (n)
  - Mayor diferencia/efecto a detectar (δ)
  - Menor variabilidad (σ)
  - Mayor nivel de significancia (α) — pero aumenta Error Tipo I

8.11.3. Tamaño Muestral

- Se calcula especificando: efecto a detectar, potencia deseada, nivel de significancia, variabilidad

Cuadro 8.14 Código R

```r
library(stats)

# Supongamos que queremos detectar una diferencia de 5 días
# con desviación estándar combinada de 10 días
# Potencia deseada = 0.80, significancia = 0.05

power_result <- power.t.test(
  n = NULL,           # Queremos calcular n
  delta = 5,          # Diferencia a detectar
  sd = 10,            # Desviación estándar
  sig.level = 0.05,   # Significancia bilateral
  power = 0.80,       # Potencia deseada
  alternative = "two.sided"
)

cat("Tamaño muestral requerido por grupo:", round(power_result$n), "\n")
```

Cuadro 8.15 Código R

```r
# También podemos calcular la potencia de nuestro estudio actual
power_actual <- power.t.test(
  n = n1,             # Tamaño muestral observado
  delta = mean(wildtype) - mean(transgenic),  # Diferencia observada
  sd = sqrt((var(wildtype) + var(transgenic)) / 2),  # SD combinada
  sig.level = 0.05,
  alternative = "two.sided"
)

cat("Potencia del estudio observado:", round(power_actual$power, 3), "\n")
```

- Fórmula para t-test bilateral: $n = \frac{2\sigma^2(z_{1-\alpha/2}+z_{1-\beta})^2}{\delta^2}$

8.11.4. Testing Múltiple

- **Corrección de Bonferroni:** Usar $\alpha_{\text{ajustado}} = \alpha/k$ cuando se hacen k pruebas
- Evitar prueba de igualdad de varianzas antes de elegir test (mejor usar Welch siempre)

8.12. Verificación de Supuestos: Normalidad y Homogeneidad

Los contrastes asintóticos basados en el Teorema Central del Límite —como el t-test y el F-test— requieren verificar dos condiciones: (1) normalidad de los datos o residuos, y (2) igualdad de varianzas entre grupos. Con $n \geq 60$ el TCL garantiza normalidad asintótica de la media, pero para muestras menores conviene comprobarlo formalmente.

8.12.1. Test de Shapiro-Wilk: Normalidad

El test de Shapiro-Wilk [Shapiro and Wilk, 1965] es la prueba de normalidad más potente para muestras pequeñas y moderadas ($n \leq 5000$).

Definición: Test de Shapiro-Wilk

Dada una muestra $x_1, \ldots, x_n$, el estadístico es:

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

donde $x_{(i)}$ son los estadísticos de orden y a_i son coeficientes tabulados que dependen de la distribución normal estándar.

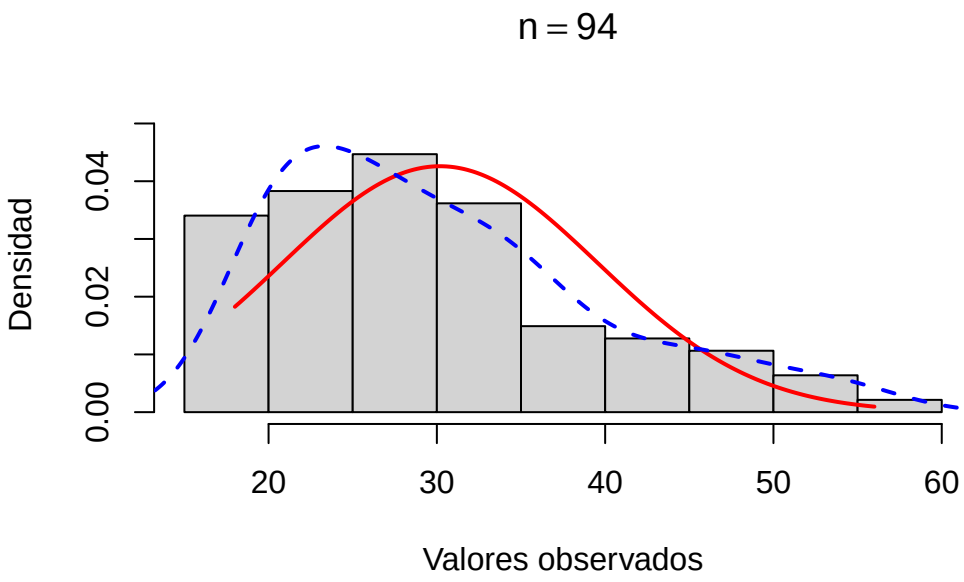
- H_0: la muestra proviene de una distribución normal
- H_1: la distribución no es normal
- Se rechaza H_0 cuando W es **pequeño** (p-valor $< \alpha$)
- $W \in (0, 1]$; valores próximos a 1 indican normalidad

Ejemplo: Test de normalidad en datos clínicos

Usando el conjunto de datos `osteo` de `BioEstatR`, que contiene 94 pacientes diabéticos de la Facultad de Medicina de la UGR, comprobamos la normalidad de la variable edad:

Cuadro 8.16 Código R

```
library(BioEstatR)

# Test de Shapiro-Wilk con BioEstatR
testnormal(osteo$edad)
```




```
# Test de normalidad de Shapiro-Wilk
 -----------------------------------
   n = 94,  W = 0.907,  p < 0.001
```

Cuadro 8.17 Código R

```
# Test de Shapiro-Wilk con R base
shapiro.test(osteo$edad)
```

```
    Shapiro-Wilk normality test

data:  osteo$edad
W = 0.9, p-value = 0.000005
```

Interpretación: Con $W = 0.907$ y $p < 0.001$, rechazamos normalidad ($\alpha = 0.05$). Para alternativas cuando no hay normalidad y $n < 60$, ver **?@sec-semana12-bootstrap**.

8.12.2. Test de Bartlett: Homogeneidad de Varianzas

El test de Bartlett contrasta si k grupos independientes tienen la misma varianza poblacional, bajo el supuesto de que los datos son normales.

i Definición: Test de Bartlett

Sean k grupos con tamaños $n_1, \ldots, n_k$ y varianzas muestrales $s_1^2, \ldots, s_k^2$.
Hipótesis: $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ frente a H_1: al menos dos varianzas difieren.
Varianza combinada:

$$s_p^2 = \frac{\sum_{j=1}^{k} (n_j - 1)\, s_j^2}{n - k}, \quad n = \sum_{j=1}^{k} n_j$$

Estadístico de Bartlett:

$$\chi^2_B = \frac{1}{c}\left[(n-k)\ln s_p^2 - \sum_{j=1}^{k}(n_j-1)\ln s_j^2\right]$$

donde $c = 1 + \frac{1}{3(k-1)}\left(\sum_{j=1}^{k}\frac{1}{n_j-1} - \frac{1}{n-k}\right)$ es un factor de corrección.
Bajo H_0: $\chi^2_B \sim \chi^2_{k-1}$ (asintóticamente). Sensible a no normalidad — si los datos no son normales, usar **Test de Levene** (`car::leveneTest()`).

Ejemplo: Homogeneidad del IMC por grupo de edad

Cuadro 8.18 Código R

```
library(BioEstatR)

# Test de Bartlett: IMC igual en los 3 grupos de edad?
bartlett.test(imc ~ grupo_edad, data = osteo)
```

```
    Bartlett test of homogeneity of variances

data:  imc by grupo_edad
Bartlett's K-squared = 11, df = 2, p-value = 0.004
```

Interpretación: $\chi^2_B \approx 10.96$, $gl = 2$, $p = 0.004 < 0.05$: **se rechaza** la igualdad de varianzas. El supuesto de homocedasticidad **no se cumple**; conviene usar el test de Welch (`oneway.test(..., var.equal = FALSE)`) o, si los datos no son normales, el test de Levene (`car::leveneTest()`) y comparaciones no paramétricas como Kruskal-Wallis. Evitar t-test o ANOVA clásico.

Regla Práctica: ¿Cuándo verificar los supuestos?

Supuesto	Test	Función R	Cuándo usarlo
Normalidad	Shapiro-Wilk	`shapiro.test()`	Siempre si $n < 60$
Homogeneidad	Bartlett	`bartlett.test()`	Datos normales
Homogeneidad	Levene	`car::leveneTest()`	Datos no normales

TCL: Con $n \geq 60$, el t-test es robusto a la no normalidad [Casella and Berger, 2002]. Para $n < 60$ y datos no normales, ver métodos bootstrap en **?@sec-semana12-bootstrap**.

8.13. Bootstrap y Métodos Robustos como Alternativas

Cuando el test de Shapiro-Wilk rechaza normalidad **y** el tamaño muestral es insuficiente para garantizar el TCL ($n < 60$), los métodos clásicos basados en la distribución t o F pueden producir inferencias incorrectas. El **bootstrap** y los **tests de permutación** son alternativas válidas que no requieren supuestos distribucionales [Chihara and Hesterberg, 2019].

Definición: Bootstrap (Efron, 1979)

El bootstrap simula la distribución muestral de un estadístico $\hat{\theta}$ remuestreando **con reemplazo** de la muestra original.

Algoritmo del IC bootstrap por percentiles:

1. Sea $\mathbf{x} = (x_1, \ldots, x_n)$ la muestra original.
2. Para $b = 1, \ldots, B$ (con $B = 10000$ recomendado):
   - Generar muestra bootstrap $\mathbf{x}^{(b)}$: extraer n valores con reemplazo de $\mathbf{x}$
   - Calcular $\hat{\theta}^{(b)} = T(\mathbf{x}^{(b)})$
3. El IC al $(1-\alpha)\,\%$ por percentiles es:

$$\left[\hat{\theta}^*_{(\alpha/2)},\ \hat{\theta}^*_{(1-\alpha/2)}\right]$$

donde los superíndices son cuantiles de la distribución bootstrap $\{\hat{\theta}^{(b)}\}_{b=1}^{B}$.

El **método BCa** (Bias-Corrected and Accelerated) ajusta por sesgo y asimetría — es más preciso y se recomienda para muestras pequeñas.

Definición: Test de Permutación

Para contrastar $H_0 : \mu_1 = \mu_2$ (igualdad de medias entre dos grupos), el test de permutación calcula la distribución exacta del estadístico de prueba bajo H_0 permutando aleatoriamente las etiquetas de grupo.

Algoritmo:

1. Calcular la diferencia observada: $D_{\text{obs}} = \bar{x}_1 - \bar{x}_2$
2. Para $b = 1, \ldots, B$: permutar aleatoriamente las asignaciones de grupo $\rightarrow$ calcular $D^{(b)}$
3. p-valor bilateral: $\hat{p} = \frac{\#\{|D^{(b)}| \geq |D_{\text{obs}}|\}}{B}$

No requiere normalidad ni homocedasticidad. Válido para cualquier tamano muestral.

Ejemplo: IMC en hombres diabéticos — bootstrap manual con bucle

El IMC de los pacientes masculinos del dataset `osteo` **no sigue una distribución normal** (Shapiro-Wilk: $W = 0.924$, $p = 0.006$). Implementamos el bootstrap **manualmente con un bucle `for`** (estilo Chihara and Hesterberg [2019]), siguiendo paso a paso el algoritmo de Efron. Después comparamos con el método BCa (paquete `boot`) y con el IC clásico t.

Cuadro 8.19 Codigo R

```r
library(BioEstatR)
library(boot)
set.seed(2024)

# --- Datos: IMC de hombres diabéticos (n = 45, no normal) ---
x <- osteo$imc[osteo$sexo == "Hombre"]
n <- length(x)
media_obs <- mean(x)

shapiro.test(x)   # Confirma no-normalidad
```

```
    Shapiro-Wilk normality test

data:  x
W = 0.9, p-value = 0.006
```

Cuadro 8.20 Codigo R

```r
# --- 1. BOOTSTRAP MANUAL CON BUCLE FOR (algoritmo paso a paso) ---
B <- 10000                          # número de réplicas bootstrap
boot_means <- numeric(B)            # vector para almacenar las medias bootstrap

for (b in 1:B) {
  # (i) Remuestrear con reemplazo: extraer n valores de la muestra original
  muestra_boot <- sample(x, size = n, replace = TRUE)
  # (ii) Calcular el estadístico de interés sobre la muestra bootstrap
  boot_means[b] <- mean(muestra_boot)
}

# (iii) Intervalo de confianza por percentiles 2.5% y 97.5%
IC_perc <- quantile(boot_means, c(0.025, 0.975))

cat("Media muestral observada:", round(media_obs, 3), "kg/m²\n")
```

```
Media muestral observada: 23.5 kg/m²
```

Cuadro 8.21 Codigo R

```r
cat("Réplicas bootstrap (B):", B, "\n")
```

```
Réplicas bootstrap (B): 10000
```

Cuadro 8.22 Codigo R

```r
cat("Media bootstrap promedio:", round(mean(boot_means), 3), "kg/m²\n")
```

```
Media bootstrap promedio: 23.5 kg/m²
```

Cuadro 8.23 Codigo R

```r
cat("Error estándar bootstrap:", round(sd(boot_means), 3), "\n")
```

```
Error estándar bootstrap: 0.428
```

Cuadro 8.24 Codigo R

```r
cat("IC 95% (percentiles, bucle manual): [",
    round(IC_perc[1], 3), ", ", round(IC_perc[2], 3), "] kg/m²\n\n")
```

```
IC 95% (percentiles, bucle manual): [ 22.7 ,  24.4 ] kg/m²
```

Cuadro 8.25 Codigo R

```r
# --- 2. BOOTSTRAP BCa con el paquete boot (recomendado en n pequeñas) ---
boot_fn  <- function(data, i) mean(data[i])
boot_obj <- boot(x, statistic = boot_fn, R = B)
boot.ci(boot_obj, type = "bca")
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_obj, type = "bca")

Intervals :
Level       BCa
95%   (22.8, 24.5 )
Calculations and Intervals on Original Scale
```

Cuadro 8.26 Codigo R

```r
# --- 3. IC clásico t para comparación ---
t.test(x)$conf.int
```

```
[1] 22.6 24.4
attr(,"conf.level")
[1] 0.95
```

Interpretación

El bootstrap manual con bucle `for` reproduce paso a paso el algoritmo de Efron (1979): en cada una de las B = 10.000 iteraciones se construye una pseudomuestra extrayendo n = 45 pacientes con reemplazo del IMC observado, se calcula su media, y al final se ordenan las 10.000 medias para extraer los percentiles 2.5 % y 97.5 %. El IC 95 % resultante ([22.7, 24.4] kg/m²) se interpreta clínicamente como el rango plausible del IMC poblacional medio de varones diabéticos —dentro del rango **normopeso** (18.5–24.9 kg/m²) próximo al umbral del sobrepeso según criterios OMS. Comparado con el IC t clásico, el bootstrap proporciona inferencia válida sin requerir normalidad, lo cual es relevante en cohortes pequenas con asimetría en la distribución del IMC (común en diabetes tipo 2). El método BCa corrige por sesgo y asimetría y se recomienda como referencia cuando $n < 30$; con $n = 45$ los tres intervalos convergen porque el TCL ya comienza a actuar. Esta convergencia se romperá si reducimos n por debajo de 20 o si analizamos variables muy asimétricas (p. ej., triglicéridos o carga viral).

Ejemplo: Test de permutación — diferencia de IMC por sexo

Shapiro-Wilk rechaza normalidad tanto en hombres ($p = 0.006$) como en mujeres ($p = 0.001$). Usamos un test de permutación en lugar del t-test clásico.

Cuadro 8.27 Código R

```
library(BioEstatR)
set.seed(2024)
B <- 10000

g1 <- osteo$imc[osteo$sexo == "Hombre"]   # n = 45
g2 <- osteo$imc[osteo$sexo == "Mujer"]    # n = 49

# Diferencia observada
D_obs <- mean(g1) - mean(g2)

# Distribución de permutación bajo H0
combined <- c(g1, g2)
n1 <- length(g1)
perm_diffs <- replicate(B, {
  perm <- sample(combined)
  mean(perm[1:n1]) - mean(perm[(n1+1):length(combined)])
})

# p-valor bilateral
p_perm <- mean(abs(perm_diffs) >= abs(D_obs))
cat("Diferencia observada:", round(D_obs, 3), "\n")
```

```
Diferencia observada: -0.78
```

Cuadro 8.28 Código R

```
cat("p-valor permutacion: ", round(p_perm, 4), "\n")
```

```
p-valor permutacion:  0.312
```

Cuadro 8.29 Código R

```
# Comparación con t-test clásico
t.test(g1, g2)$p.value
```

```
[1] 0.308
```

Interpretación: Ambos métodos concuerdan: no hay diferencia significativa en el IMC entre hombres y mujeres ($p \approx 0.32$). Cuando los tamaños son moderados ($n_1 = 45$, $n_2 = 49$), el TCL protege al t-test. El test de permutación proporciona validez exacta sin supuestos.

⚠ Cuándo usar cada método

Situación	Método recomendado	Función R
Normal, n cualquiera	t-test, ANOVA clásico	`t.test()`, `aov()`
No normal, $n \geq 60$	t-test (TCL garantiza)	`t.test()`
No normal, $n < 60$, 1 muestra	IC Bootstrap BCa	`boot()`, `boot.ci()`
No normal, $n < 60$, 2 grupos	Permutación o Wilcoxon	`replicate()`, `wilcox.test()`
No normal, $k > 2$ grupos	Kruskal-Wallis	`kruskal.test()`
Varianzas desiguales	Welch t-test	`t.test(var.equal=FALSE)`

Referencia: Chihara and Hesterberg [2019] proporciona una tratamiento riguroso de bootstrap y permutaciones con implementación completa en R, incluyendo la justificación teórica de cada método.

8.14. Pruebas No Paramétricas Basadas en Rangos

Las pruebas no paramétricas son alternativas a las pruebas paramétricas (como el t-test o el ANOVA) que no requieren que los datos sigan una distribución normal. Se basan en el **rango** (posición) de las observaciones en lugar de sus valores originales, lo que las hace robustas frente a valores atípicos (outliers) y distribuciones asimétricas.

8.14.1. Prueba de Wilcoxon-Mann-Whitney (Muestras Independientes)

Es la alternativa no paramétrica al t-test de dos muestras independientes. Contrasta si las distribuciones de dos grupos son idénticas, evaluando si un valor seleccionado al azar de una población tiende a ser mayor que uno de la otra.

ℹ Definición: Estadístico U de Mann-Whitney

Dadas dos muestras independientes de tamaños n_1 y n_2:

1. Se combinan ambas muestras y se asignan rangos (de 1 a $n_1 + n_2$).
2. Se calcula la suma de rangos para cada grupo (R_1 y R_2).
3. El estadístico U es el mínimo de:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, \quad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Ejemplo: Comparación del IMC (Wilcoxon-Mann-Whitney)

Contrastamos si el Indice de Masa Corporal (IMC) difiere entre hombres y mujeres usando la alternativa no paramétrica, dado que el IMC frecuentemente presenta asimetrîa.

Cuadro 8.30 Código R

```
library(BioEstatR)

# Wilcoxon-Mann-Whitney con BioEstatR
testwx(m1 = osteo$imc[osteo$sexo == "Hombre"],
       m2 = osteo$imc[osteo$sexo == "Mujer"],
       grf = FALSE)
```

```
Test de Wilcoxon/Mann-Whithney para dos muestras independientes
------------------------------------------------------------------
# Información muestral ---

                           Muestra  n    min     Q1     Q2     Q3    max   RIQ
1 osteo$imc[osteo$sexo == "Hombre"] 45 18.327 21.436 23.055 24.259 32.323 2.823
2  osteo$imc[osteo$sexo == "Mujer"] 49 18.070 21.155 23.338 26.298 37.333 5.143

# Rangos ---

                           Muestra  n Suma_rangos Rango_medio        U
1 osteo$imc[osteo$sexo == "Hombre"] 45        2095      46.556 1145.000
2  osteo$imc[osteo$sexo == "Mujer"] 49        2370      48.367 1060.000

# Test ---

  U = 1060.000; Z = 0.322; W = 1060.000; p = 0.750

# Tamaño del efecto ---

  Diferencia de localización: -0.193    95%-IC = (-1.595, 0.962)
  r = 0.033  (criterio: 0.1 pequeño; 0.3 mediano; >0.5 grande)
  Probabilidad de superioridad PS = 0.519
  (probabilidad de que un valor al azar de M1 sea < a un valor al azar de M2)
```

Interpretación: No se observan diferencias significativas en las distribuciones de IMC entre sexos ($p = 0.751$). La probabilidad de que un hombre elegido al azar tenga un IMC mayor que una mujer elegida al azar es del **48.1 %** (calculada como $1 - PS = 1 - 0.519$, ya que la salida de `testwx()` reporta PS como probabilidad de que un valor de M_1 sea **menor** que uno de

M_2).

8.14.2. Prueba de Wilcoxon de los Rangos con Signo (Muestras Pareadas)

Es la alternativa no paramétrica al t-test pareado. Se utiliza para comparar dos medidas relacionadas (antes-después) sobre los mismos sujetos cuando las diferencias no siguen una distribución normal.

Procedimiento

1. Calcular las diferencias $D_i = X_{1i} - X_{2i}$.
2. Eliminar las diferencias iguales a cero.
3. Asignar rangos a los valores absolutos $|D_i|$.
4. El estadístico W es la suma de los rangos de las diferencias positivas.

Ejemplo: Prueba pareada con BioEstatR

Supongamos que evaluamos el calcio sérico (`ca`) antes y después de un tratamiento en 10 pacientes (datos simulados para ilustración).

Cuadro 8.31 Código R

```r
# Calcio inicial y final (simulados)
ca_pre  <- c(9.2, 8.4, 8.8, 9.1, 9.5, 9.0, 8.9, 9.3, 8.5, 8.7)
ca_post <- c(9.5, 8.7, 9.0, 9.4, 9.6, 9.1, 9.0, 9.6, 8.9, 8.8)

# Wilcoxon pareado con BioEstatR (par = TRUE)
testwx(m1 = ca_pre, m2 = ca_post, par = TRUE, grf = FALSE)
```

```
Test de Wilcoxon  para dos muestras apareadas
-----------------------------------------------
# Información muestral ---

  Muestra  n   min    Q1    Q2    Q3   max   RIQ
1  ca_pre 10 8.400 8.725 8.950 9.175 9.500 0.450
2 ca_post 10 8.700 8.925 9.050 9.475 9.600 0.550

# Rangos ---

  Se obtienen las diferencias como ca_pre - ca_post
  Pares de datos efectivos para los rangos: 10 de 10

        Muestra  n Suma_rangos Rango_medio
1 dif.negativas 10          55       5.500
```

```
2 dif.positivas  0           0         NaN

# Test ---

  V = 0.000; p = 0.002
  z = 2.803; p = 0.005

# Correlación de Spearman ---

  rho-Spearman = 0.982; p < 0.001

# Tamaño del efecto ---

  Diferencia de localización:  (pseudo)mediana = -0.200    95%-IC = (-
0.300, -0.100)
  r = 0.627; p = 0.002
```

Interpretacion: Existe un aumento significativo en los niveles de calcio tras el tratamiento ($p = 0.005$). El tamaño del efecto es grande ($r = 0.627$, criterio convencional: $r > 0.5$ grande).

8.14.3. Prueba de Kruskal-Wallis (K Muestras Independientes)

Es la alternativa no paramétrica al ANOVA de un factor. Se utiliza para comparar las medianas de tres o más grupos independientes.

Estadîstico H de Kruskal-Wallis

Para k grupos con tamaños n_j y suma de rangos R_j:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$$

donde $n = \sum n_j$. Bajo H_0, $H \sim \chi^2_{k-1}$ asintóticamente.

Ejemplo: IMC por grupo de edad (> 2 grupos)

Cuadro 8.32 Código R

```r
# Contrastar si el IMC difiere entre los 3 grupos de edad
kruskal.test(imc ~ grupo_edad, data = osteo)
```

```
    Kruskal-Wallis rank sum test
```

```
data:  imc by grupo_edad
Kruskal-Wallis chi-squared = 17, df = 2, p-value = 0.0002
```

Interpretación: Como $p \approx 0.0002 < 0.05$, rechazamos la hipótesis de igualdad. Al menos un grupo de edad tiene una distribución de IMC significativamente distinta.

8.15. Ejercicios

Ejercicio 8.1: Se comparan dos métodos de enseñanza en estadîstica. El Método A (n = 30) tiene media 75 y desviación estándar 8. El Método B (n = 35) tiene media 78 y desviación estándar 7. ¿Hay diferencia significativa entre los métodos a nivel = 0.05? (Suponga varianzas iguales)

Ejercicio 8.2: Un investigador mide presión arterial en 10 pacientes antes y después de tomar un medicamento. Las diferencias promedio son -5 mmHg con desviación estándar 8 mmHg. ¿Es el medicamento efectivo en reducir presión (prueba unilateral, = 0.05)?

Ejercicio 8.3: Se comparan dos proveedores de materia prima. Proveedor 1: $S^2 = 15$, n = 20. Proveedor 2: $S^2 = 8$, n = 18. ¿Hay diferencia significativa en variabilidad (= 0.05, bilateral)?

Ejercicio 8.4: Se diseña un estudio para detectar una diferencia de 12 puntos en una escala (= 10) con potencia 0.90 y = 0.05. ¿Cuántos sujetos se necesitan por grupo?

Ejercicio 8.5: Un investigador realiza 15 pruebas independientes con = 0.05 sin corrección. ¿Cuál es la probabilidad aproximada de cometer al menos un Error Tipo I?

Ejercicio 8.6: Interprete este resultado de R: `t.test(x, y, var.equal=FALSE)` da t = 2.34, df = 45.2, p-value = 0.024. ¿Qué conclusión saca?

Ejercicio 8.7: Un estudio observa media = 100, media = 105, pero el intervalo de confianza al 95 % para la diferencia incluye cero. ¿Qué significa esto respecto a la significancia estadîstica y práctica?

Ejercicio 8.8: Si el test de Shapiro-Wilk rechaza la normalidad en una muestra pequeña ($n = 15$), ¿qué prueba utilizarîa para comparar dos grupos independientes?

Ejercicio 8.9: ¿Cuál es la principal ventaja de la prueba de Kruskal-Wallis sobre el ANOVA de un factor?

Ejercicio 8.10: En una muestra de 35 valores de presión arterial sistólica, el test de Shapiro-Wilk produce $W = 0.94$ con $p = 0.089$. ¿Es válido aplicar un test t de una muestra con $\alpha = 0.05$? ¿Cambiarîa su decisión si $n = 75$?

8.16. Respuestas a los Ejercicios

Ejercicio 8.1: t = (75-78)/√((29×64 + 34×49)/(63) × (1/30 + 1/35)) -1.81; df 63; p > 0.05. No hay diferencia significativa.

Ejercicio 8.2: t = -5/(8/√10) -1.98; df = 9; p 0.038 < 0.05. El medicamento es efectivo en reducir presión.

Ejercicio 8.3: F = 15/8 = 1.875; df = 19, df = 17. F crîtico 2.74. No rechaza H ; no hay diferencia significativa en variabilidad.

Ejercicio 8.4: Usando fórmula de potencia: n 2×(10)²×(1.96 + 1.28)²/(12)² 29-30 sujetos por grupo.

Ejercicio 8.5: P(al menos 1 error) = 1 - (1-0.05)¹ 1 - 0.463 = 0.537 (53.7 %). El problema de multiplicidad es importante.

Ejercicio 8.6: t = 2.34, df = 45.2, p = 0.024 < 0.05. Se rechaza H . Hay diferencia significativa entre x e y (usando Welch's t-test por varianzas desiguales).

Ejercicio 8.7: Si el IC 95 % incluye cero, la diferencia no es estadîsticamente significativa (p > 0.05), aunque hay diferencia práctica de 5 unidades. Sin significancia estadîstica, no podemos afirmar que la diferencia poblacional es real.

Ejercicio 8.8: Utilizarîa la prueba de Wilcoxon-Mann-Whitney (o un test de permutación).

Ejercicio 8.9: No requiere el supuesto de normalidad en los datos, siendo más robusta cuando hay outliers o distribuciones muy asimétricas.

Ejercicio 8.10: Con $p = 0.089 > 0.05$, **no rechazamos** H_0 de normalidad → el test t es válido. Con $n = 75$, el TCL garantiza normalidad asintótica de la media, por lo que el test t es válido independientemente del resultado de Shapiro-Wilk.

8.17. Recursos Adicionales

- **Software R:** Funciones `t.test()`, `var.test()`, `power.t.test()`, `wilcox.test()`, `kruskal.test()` en stats
- **OpenIntro Statistics:** Capîtulos sobre comparación de dos medias y potencia
- **Referencias clásicas:** Rosner (2011) Fundamentals of Biostatistics; Kahn & Sempos (2002) Statistical Methods in Epidemiology

Parte IV

Parte IV: Regresión Lineal

Capítulo 9

Semana 9 — Regresión Lineal Simple

9.1. El Problema de Regresión

9.1.1. Variables en un modelo de regresión

El objetivo de la regresión es **cuantificar la relación entre una variable respuesta y una o más variables explicativas**. En ciencias de la salud, este enfoque permite responder preguntas como: ¿aumenta la presión arterial con la edad? ¿predice el índice de masa corporal el nivel de colesterol? ¿cuánto reduce la hemoglobina glicosilada cada año adicional de tratamiento?

i Definición: Variables en Regresión

- **Variable respuesta (dependiente)** Y: la variable clínica que queremos explicar o predecir (ej. presión arterial sistólica, HbA1c, colesterol)
- **Variables explicativas (independientes)** $X_1, X_2, \dots, X_p$: variables predictoras, regresores, covariables (ej. edad, IMC, dosis)
- **Función de regresión**: $E(Y|X_1, X_2, \dots, X_p) = m(x_1, x_2, \dots, x_p)$

Esta función describe el **valor esperado (promedio)** de Y dados los valores de las variables explicativas.

En esta semana nos enfocamos en el caso **simple**: una única variable explicativa X.

9.1.2. Observaciones en una muestra

Disponemos de n observaciones muestrales:

- $(y_i, x_{1i}, x_{2i}, \dots, x_{pi})$ para $i = 1, 2, \dots, n$
- y_i: valor observado de la variable respuesta (ej. PAS del paciente i, mmHg)
- x_{ki}: valores observados de las variables explicativas (ej. edad del paciente i, años)

9.2. Supuestos del Modelo Lineal

Para que la regresión lineal funcione correctamente, asumimos que se cumplen cinco supuestos:

⚠ Los Cinco Supuestos del Modelo Lineal

1. **Media cero de errores**: $E(U_i) = 0$ para $i = 1, \dots, n$
   - Los errores no tienen sesgo sistemático
2. **Homoscedasticidad**: $\text{Var}(U_i) = \sigma_u^2$ para $i = 1, \dots, n$
   - La varianza de los errores es constante (no depende de X)
3. **No autocorrelación**: $\text{Cov}(U_i, U_j) = 0$ para $i \neq j$
   - Los errores de diferentes observaciones no están correlacionados
4. **Normalidad**: $U_i \sim N(0, \sigma_u^2)$
   - Los errores siguen una distribución normal
5. **Regresores fijos**: Los valores x_{ki} son cantidades fijas, independientes de la muestra
   - No hay aleatoriedad en los predictores

9.2.1. Implicaciones de los supuestos

Bajo estos supuestos, como $Y_i = m(X_i) + U_i$:

- Si U_i tienen media cero: $E(Y_i) = m(X_i)$
- Si U_i son normales: $Y_i \sim N(m(X_i), \sigma_u^2)$
- Los valores de Y_i son **variables aleatorias** (varían entre muestras)

9.3. El Modelo de Regresión Lineal Simple

9.3.1. Modelo poblacional

Bajo los supuestos anteriores, especificamos una relación **lineal** entre Y y X:

ℹ Definición: Modelo de Regresión Lineal Simple

$$Y_i = \beta_0 + \beta_1 X_i + U_i, \quad i = 1, \dots, n$$

donde:

- β_0 es la **ordenada en el origen** (intercept): el valor esperado de Y cuando $X = 0$
- β_1 es la **pendiente**: el cambio en el valor esperado de Y cuando X aumenta en una unidad
- U_i es el **término de error**: captura factores no observados y variación aleatoria

La **función de regresión poblacional** es:

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

9.3.2. Interpretación de los parámetros

- β_1 **(pendiente)**: Si X aumenta en 1 unidad, entonces Y aumenta **en promedio** β_1 unidades
- β_0 **(intercept)**: El valor medio de Y cuando $X = 0$ (no siempre tiene interpretación clínica directa)

9.3.3. Ejemplo: Presión Arterial Sistólica vs. Edad

En un estudio transversal realizado en la consulta de medicina interna del Hospital Clínico San Cecilio de Granada, se recogieron datos de 120 adultos para estudiar la asociación entre la presión arterial sistólica (PAS) y la edad:

- Y: presión arterial sistólica (mmHg)
- X: edad del paciente (años)
- $n = 120$ pacientes adultos (25–80 años)

El modelo sería: $\text{PAS}_i = \beta_0 + \beta_1 \times \text{Edad}_i + U_i$

Si estimamos $\beta_1 = 0.80$ mmHg/año, significa que por cada año adicional de edad, la PAS aumenta en promedio 0.80 mmHg. Si $\beta_0 = 85.2$ mmHg, es el nivel basal estimado (extrapolación para edad = 0, sin interpretación directa).

Ejemplo con BioEstatR: rls()

Documentación completa en Sección B.6.1.

```
library(BioEstatR)

# ¿Los años de evolución de la diabetes predicen la HbA1c?
# Dataset: 94 pacientes diabéticos, Fac. Medicina UGR
rls(y = osteo$hba1c, x = osteo$tevol, grf = FALSE)
```

```
Regresión lineal simple
---------------------------------------------------------------
             n  media    dt   Min   Max
     hba1c  94  8.565 1.799  4.60 13.80
     tevol  94 12.330 8.534  0.00 35.00

# Correlación de Pearson:
  r = -0.237  [IC 95%: (-0.420, -0.036)]
  t = -2.341,  gl = 92,  p = 0.021

# Modelo:  hba1c = 9.181 - 0.050 × tevol
  R² = 0.056
```

```
         estim    se  ic_inf ic_sup    sig
(Cte)    9.181 0.320   8.546  9.816 <0.001
tevol   -0.050 0.021  -0.092 -0.008  0.021

Error estándar residual: 1.757
```

`rls()` integra correlación, coeficientes con IC, R^2 y diagnóstico de residuos. La asociación es estadísticamente significativa pero débil ($R^2 = 0.056$): el tiempo de evolución solo explica el 5.6 % de la variabilidad en HbA1c. Esto es clínicamente relevante: la diabetes de larga evolución tiende a un control glucémico algo peor, pero otros factores (adherencia al tratamiento, estilo de vida) son predominantes.

9.4. Estimación por Mínimos Cuadrados Ordinarios

9.4.1. El criterio de MCO

Los parámetros poblacionales β_0 y β_1 son **desconocidos**. Los estimamos usando una muestra mediante **Mínimos Cuadrados Ordinarios (MCO)**, que minimiza la suma de residuos al cuadrado:

⚠ Definición: Estimadores MCO

Los estimadores MCO minimizan:

$$Q(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Donde:

- y_i son los valores observados (PAS del paciente i)
- $\hat{y}_i = b_0 + b_1 x_i$ son los valores predichos por el modelo
- El residuo estimado es $\hat{u}_i = y_i - \hat{y}_i$

Así buscamos la línea que **mejor ajusta** los datos, minimizando las distancias verticales.

9.4.2. Derivación: Las ecuaciones normales

Para minimizar $Q(b_0, b_1)$, igualamos a cero las derivadas parciales:

$$\frac{\partial Q}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial Q}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i = 0$$

Esto genera las **ecuaciones normales**:

$$nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

9.4.3. Fórmulas cerradas para los estimadores MCO

Resolviendo el sistema anterior obtenemos:

⚠ Fórmulas Cerradas: Estimadores MCO

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

donde:

- $\bar{x} = \frac{1}{n}\sum x_i$ y $\bar{y} = \frac{1}{n}\sum y_i$ son los promedios
- s_{xy} es la covarianza muestral de X e Y
- s_x^2 es la varianza muestral de X

9.4.4. Propiedad importante

La línea de regresión estimada **siempre pasa por el punto** $(\bar{x}, \bar{y})$: cuando $x_i = \bar{x}$, se cumple que $\hat{y} = \bar{y}$.

9.5. Relación entre Regresión y Correlación

Existe una relación útil entre el coeficiente de regresión y el coeficiente de correlación:

💡 Conexión: Regresión y Correlación

$$b_1 = r_{xy} \frac{s_y}{s_x}$$

donde $r_{xy} = \frac{s_{xy}}{s_x s_y}$ es el coeficiente de correlación de Pearson.
Interpretación:

- El coeficiente de regresión escala la correlación por el ratio de desviaciones estándar
- Si $s_x = s_y$, entonces $b_1 = r_{xy}$
- Regresión de X sobre Y es **diferente** de regresión de Y sobre X

Ejemplo médico: Si en un estudio de hipertensión la correlación entre edad y PAS es $r = 0.71$, con $s_{\text{PAS}} = 14.5$ mmHg y $s_{\text{edad}} = 15.9$ años, la pendiente de regresión es: $b_1 =$

$0.71 \times (14.5/15.9) = 0.65$ mmHg/año.

9.6. Descomposición de Varianza

9.6.1. Suma de cuadrados totales, explicada y residual

La variabilidad en Y se descompone en dos partes:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Terminología: Sumas de Cuadrados

- **SCT** $= \sum_{i=1}^{n}(y_i - \bar{y})^2$: Suma de Cuadrados Total
  - Variabilidad total en los datos (ej. variabilidad total en PAS entre pacientes)
- **SCR** $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: Suma de Cuadrados de Residuos
  - Variabilidad en PAS NO explicada por la edad
- **SCE** $= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$: Suma de Cuadrados Explicada
  - Variabilidad en PAS explicada por la edad

Relación: SCT = SCE + SCR

9.7. Coeficiente de Determinación (R^2)

El coeficiente de determinación mide qué proporción de la variabilidad total en Y es explicada por el modelo:

Definición: Coeficiente de Determinación

$$R^2 = \frac{\text{SCE}}{\text{SCT}} = 1 - \frac{\text{SCR}}{\text{SCT}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Para regresión lineal simple:

$$R^2 = r_{xy}^2$$

donde r_{xy} es el coeficiente de correlación de Pearson.

Propiedades:

- $0 \leq R^2 \leq 1$
- $R^2 = 1$: ajuste perfecto (todos los puntos en la línea)
- $R^2 = 0$: no hay relación lineal
- $R^2 = 0.50$: el 50 % de la variabilidad en Y es explicada por X

9.7.1. Ejemplo: Presión Arterial Sistólica

En el estudio de hipertensión (n = 120 pacientes), si la correlación entre edad y PAS es $r_{xy} = 0.71$:

$$R^2 = 0.71^2 = 0.504$$

Interpretación clínica: El 50.4 % de la variabilidad en la presión arterial sistólica entre estos pacientes es explicada por la edad. El 49.6 % restante se debe a otros factores no recogidos en el modelo (tabaquismo, actividad física, tratamiento farmacológico, herencia genética, etc.).

9.8. Distribución de los Estimadores MCO

Bajo los supuestos del modelo lineal, los estimadores b_0 y b_1 son **variables aleatorias** con distribuciones conocidas:

Distribuciones muestrales de los estimadores

$$B_0 \sim N\left(\beta_0, \sigma^2_{B_0}\right)$$

$$B_1 \sim N\left(\beta_1, \sigma^2_{B_1}\right)$$

donde las varianzas son:

$$\sigma^2_{B_1} = \frac{\sigma^2_u}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sigma^2_{B_0} = \sigma^2_u \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

y σ^2_u es la varianza poblacional del error.

9.8.1. Estimación de la varianza del error

Como σ^2_u es desconocida, la estimamos usando los residuos:

$$s^2_u = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}$$

Entonces:

$$\hat{\sigma}^2_{B_1} = \frac{s^2_u}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\sigma}^2_{B_0} = s_u^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

9.9. Test t para $\beta_1 = 0$

9.9.1. Hipótesis

En el contexto médico, nos interesa probar si existe asociación lineal entre Y y X:

$$H_0 : \beta_1 = 0 \quad \text{(la edad no predice la PAS)}$$
$$H_1 : \beta_1 \neq 0 \quad \text{(la edad sí predice la PAS)}$$

9.9.2. Estadístico de prueba

Bajo H_0 y la normalidad de los errores:

⚠ Estadístico t para β_1

$$T = \frac{B_1 - \beta_1}{\hat{\sigma}_{B_1}} \sim t_{n-2}$$

En particular, bajo H_0:

$$T = \frac{b_1}{\widehat{\text{ee}}(b_1)} \sim t_{n-2}$$

donde $\widehat{\text{ee}}(b_1) = \hat{\sigma}_{B_1}$ es el **error estándar** del estimador.

9.9.3. Regla de decisión

A nivel de significancia α:

- **Rechazamos** H_0 si $|t| > t_{1-\alpha/2,n-2}$ (test bilateral)
- **Valor p** = probabilidad de observar un valor tan extremo (o más) bajo H_0

Si $n > 30$, usamos la aproximación normal: $t_{1-\alpha/2,n-2} \approx z_{1-\alpha/2}$

9.9.4. Ejemplo ilustrativo: PAS vs. Edad

A modo de ilustración (valores **hipotéticos** para repasar la mecánica del contraste; la simulación R de la sección 9.12 producirá cifras distintas):

- Estimación: $\hat{\beta}_1 = 0.80$ mmHg/año
- Error estándar: $\widehat{\text{ee}}(b_1) = 0.09$ mmHg/año
- Estadístico: $t = \frac{0.80}{0.09} = 8.89$
- Con $\alpha = 0.05$ y $n = 120$: valor crítico $t_{0.975,118} \approx 1.980$

- Como $|8.89| > 1.980$ y $p < 0.001$, **rechazamos** H_0.

Conclusión clínica: Existe evidencia estadística muy fuerte de que la edad predice positivamente la presión arterial sistólica en esta muestra.

9.10. Intervalos de Confianza

9.10.1. Para β_1

⚠ Intervalo de Confianza para β_1

Con confianza $(1-\alpha)\times 100\,\%$:

$$\left[b_1 - t_{1-\alpha/2,n-2}\cdot \widehat{\text{ee}}(b_1),\quad b_1 + t_{1-\alpha/2,n-2}\cdot \widehat{\text{ee}}(b_1)\right]$$

9.10.2. Para β_0

⚠ Intervalo de Confianza para β_0

Con confianza $(1-\alpha)\times 100\,\%$:

$$\left[b_0 - t_{1-\alpha/2,n-2}\cdot \widehat{\text{ee}}(b_0),\quad b_0 + t_{1-\alpha/2,n-2}\cdot \widehat{\text{ee}}(b_0)\right]$$

9.10.3. Para el valor predicho $E(Y|X=x_0)$

A menudo queremos construir un intervalo para la PAS promedio esperada en pacientes de una edad concreta x_0:

💡 Intervalo de Confianza para $E(Y|x_0)$

$$\left[\hat{y}_{x_0} \pm t_{1-\alpha/2,n-2}\cdot s_u\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right]$$

donde:

- $\hat{y}_{x_0} = b_0 + b_1 x_0$ es la PAS media predicha para edad x_0
- Este intervalo es más estrecho cerca de $\bar{x}$ (mayor precisión) y más ancho en edades extremas

9.10.4. Ejemplo ilustrativo: PAS vs. Edad

Reutilizando los **mismos valores hipotéticos** del apartado anterior — $n = 120$, $\hat{\beta}_1 = 0.80$ mmHg/año, $\widehat{\text{ee}}(b_1) = 0.09$:

Con $\alpha = 0.05$ (95 % confianza), $t_{0.975,118} = 1.980$:

$$\text{IC}_{95\,\%}(\beta_1) = [0.80 - 1.980 \times 0.09,\ 0.80 + 1.980 \times 0.09] = [0.62,\ 0.98] \text{ mmHg/año}$$

Interpretación clínica: Con 95 % de confianza, por cada año adicional de edad la PAS media aumenta entre 0.62 y 0.98 mmHg en esta población.

9.11. Diagnóstico de Residuos

Los residuos son críticos para validar que los supuestos del modelo se cumplen.

Definición: Residuos

Los residuos estimados son:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$$

En contexto clínico: $\hat{u}_i$ es la diferencia entre la PAS observada del paciente i y la que el modelo predice dado su edad. Un residuo grande positivo indica que el paciente tiene una PAS más elevada de lo esperado para su edad (posiblemente por otros factores de riesgo no incluidos).

Propiedades:

- Siempre tienen media cero: $\bar{\hat{u}} = 0$
- Representan la parte de Y NO explicada por X

9.11.1. 1. Gráfico de Residuos vs. Valores Predichos

¿Qué buscar?

Interpretación: Residuos vs. Predichos

- **Patrón ideal**: nube de puntos aleatoria alrededor del cero, sin forma clara
  - Indica homocedasticidad y linealidad
- **Patrón "embudo" (varianza aumenta con la PAS predicha)**:
  - Indica heteroscedasticidad (varianza no constante)
  - Frecuente en estudios con datos de presión arterial en rangos muy amplios
- **Patrón curvo o no lineal**:
  - Sugiere que la relación edad-PAS no es estrictamente lineal
  - Podría requerir una transformación o la inclusión de X^2
- **Puntos muy alejados (outliers)**:
  - Pacientes con PAS muy inusual para su edad (investigar: error de medida, enfermedad secundaria, fármaco no declarado)

9.11.2. 2. Gráfico Q-Q (Quantile-Quantile)

Este gráfico verifica la **normalidad** de los residuos:

Interpretación: Gráfico Q-Q

- **Puntos en línea recta diagonal**:
  - Los residuos son aproximadamente normales
- **Puntos en las colas no siguen la línea**:
  - Hay colas más pesadas o más ligeras que lo esperado
  - Sugiere violación de normalidad
- **Forma de S o zigzag**:
  - Distribución muy diferente de la normal

9.11.3. 3. Gráfico de Escala-Ubicación

Combina información de ambos: raíz cuadrada de residuos estandarizados vs. predichos. Detecta heteroscedasticidad.

9.11.4. 4. Residuos vs. Orden de Observación

Verifica la **no autocorrelación**. Si los pacientes fueron reclutados consecutivamente y hay patrón temporal en los residuos, puede indicar un sesgo de selección o un cambio en el protocolo de medida.

9.12. Ejemplo Completo: Colesterol Total e IMC

9.12.1. Contexto clínico

Un estudio transversal en la Unidad de Cardiología Preventiva del Hospital Universitario de Granada evalúa si el índice de masa corporal (IMC) predice el colesterol total sérico en pacientes sin tratamiento hipolipemiante:

- Y: colesterol total (mg/dL)
- X: índice de masa corporal — IMC (kg/m²)
- $n = 250$ pacientes (18–70 años, sin estatinas ni fibratos)

9.12.2. Estadísticas descriptivas

Estadístico	Valor
$\bar{y}$ (colesterol)	214.0 mg/dL
s_y	36.0 mg/dL
$\bar{x}$ (IMC)	27.2 kg/m²
s_x	4.5 kg/m²
s_{xy}	101.25 (mg/dL) · (kg/m²)
r_{xy}	0.625

9.12.3. Cálculos

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{101.25}{20.25} = 5.00 \text{ mg/dL por kg/m}^2$$

$$b_0 = \bar{y} - b_1\bar{x} = 214.0 - 5.00 \times 27.2 = 214.0 - 136.0 = 78.0 \text{ mg/dL}$$

9.12.4. Ecuación estimada

$$\widehat{\text{Colesterol}}_i = 78.0 + 5.00 \times \text{IMC}_i$$

Interpretación clínica:

- Por cada unidad adicional de IMC (1 kg/m²), el colesterol total aumenta en promedio **5 mg/dL**
- Para un IMC de 30 kg/m² (obesidad tipo I): $\widehat{\text{Colesterol}} = 78.0 + 5.00 \times 30 = 228$ mg/dL

9.12.5. Bondad del ajuste

$$R^2 = r_{xy}^2 = 0.625^2 = 0.391$$

Interpretación: El 39.1 % de la variabilidad en el colesterol sérico se explica por el IMC. Es una asociación moderada, coherente con la literatura: el IMC es un factor de riesgo cardiovascular relevante pero no el único predictor del colesterol.

9.12.6. Inferencia estadística

Con $\widehat{\text{ee}}(b_1) = 0.41$ mg · dL ^{1}/(kg · m ^{2}):

$$t = \frac{5.00}{0.41} = 12.2, \quad p < 0.001$$

Con $n = 250$ (usando $t_{0.975,248} \approx 1.970$):

Intervalo de confianza al 95 %:

$$[5.00 - 1.970 \times 0.41,\ 5.00 + 1.970 \times 0.41] = [4.19,\ 5.81] \text{ mg/dL por kg/m}^2$$

Conclusión: El IMC es un predictor estadísticamente significativo ($p < 0.001$) y clínicamente relevante del colesterol total. En términos de salud pública, reducir el IMC en 5 kg/m² (ej. pasar de obesidad a normopeso) se asocia en promedio con una reducción de 25 mg/dL en el colesterol total.

9.13. Código R

9.13.1. Datos clínicos simulados y ajuste del modelo

```
# Datos: PAS (mmHg) vs Edad (años) - 120 pacientes adultos
set.seed(2024)
n <- 120
edad  <- round(runif(n, 25, 80))
pas   <- round(85.2 + 0.80 * edad + rnorm(n, 0, 10.5), 1)
datos <- data.frame(edad = edad, pas = pas)

# Estadísticas descriptivas
cat("=== Estadísticas descriptivas ===\n")
```

```
=== Estadísticas descriptivas ===
```

```
cat("Edad  - media:", round(mean(edad), 1), "años  |  DT:", round(sd(edad), 1), "\n")
```

```
Edad  - media: 51.8 años  |  DT: 15.9
```

```
cat("PAS   - media:", round(mean(pas),  1), "mmHg  |  DT:", round(sd(pas),  1), "\n")
```

```
PAS   - media: 127 mmHg  |  DT: 17.1
```

```
cat("Correlación de Pearson:", round(cor(edad, pas), 3), "\n\n")
```

```
Correlación de Pearson: 0.773
```

```
# Ajustar modelo de regresión lineal simple
modelo <- lm(pas ~ edad, data = datos)

# Tabla de coeficientes con broom
cat("=== Modelo de Regresión ===\n")
```

```
=== Modelo de Regresión ===
```

```
knitr::kable(tidy(modelo), digits = 3, caption = "Coeficientes del modelo lineal")
```

Tabla 9.1: Coeficientes del modelo lineal

term	estimate	std.error	statistic	p.value
(Intercept)	84.15	3.402	24.7	0
edad	0.83	0.063	13.2	0

```
# Bondad de ajuste
glance_modelo <- glance(modelo)
```

```
cat("\nR² =", round(glance_modelo$r.squared, 3),
    "|  R² ajustado =", round(glance_modelo$adj.r.squared, 3),
    "|  F-estadístico =", round(glance_modelo$statistic, 2),
    "(p < 0.001)\n")
```

```
R² = 0.597 |  R² ajustado = 0.594 |  F-estadístico = 175 (p < 0.001)
```

Tip

Interpretación estadística: El modelo de regresión lineal simple estimó un incremento de la presión arterial sistólica de aproximadamente 0.83 mmHg por cada año adicional de edad (b = 0.83, EE = 0.063, t(118) = 13.2, p < 0.001, IC 95 %: [0.71, 0.96]). El coeficiente de determinación $R^2 = 0.597$ indica que la edad explica el 59.7 % de la variabilidad observada en la PAS, lo que refleja una relación lineal fuerte y estadísticamente significativa. Este hallazgo es clínicamente relevante: un aumento de edad de 20 años se asociaría con un incremento esperado de ~17 mmHg en la PAS.

9.13.2. Gráfico de dispersión con la recta de regresión

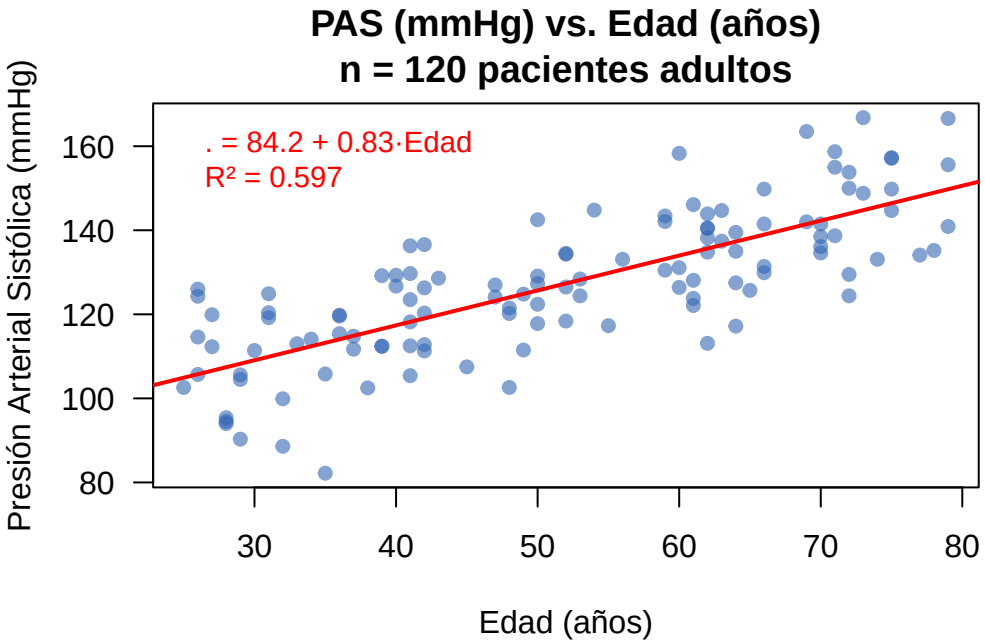


Figura 9.1: Regresión lineal: Presión Arterial Sistólica (PAS) en función de la Edad. Cada punto representa un paciente. La línea roja es la recta de mínimos cuadrados estimada.

9.13.3. Gráficos de diagnóstico de residuos

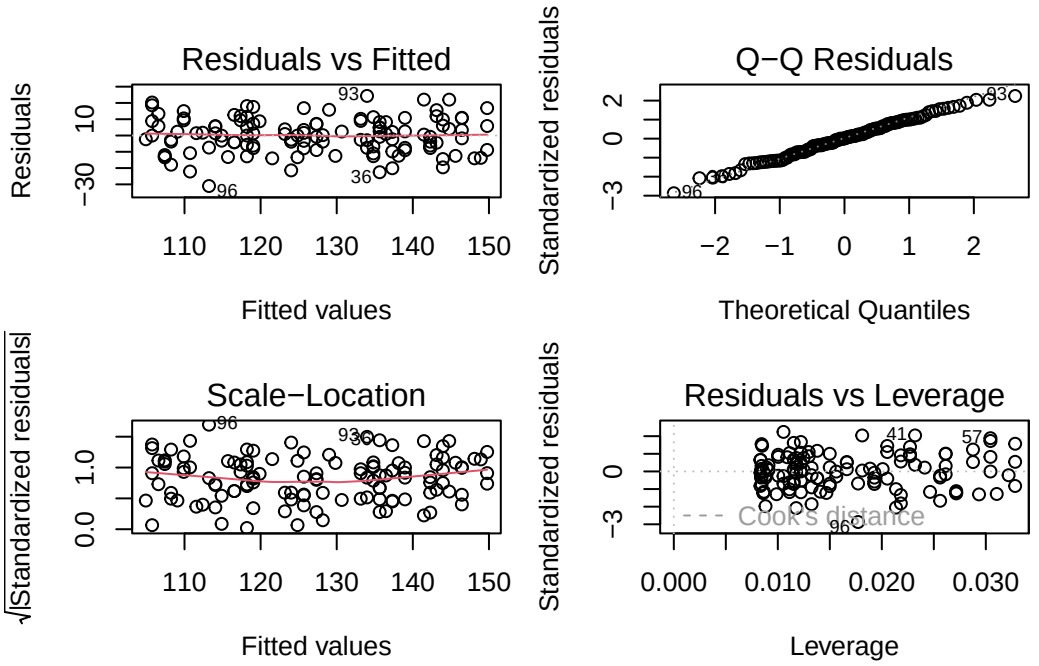


Figura 9.2: Gráficos de diagnóstico de supuestos del modelo de regresión. Arriba-izquierda: residuos vs. valores ajustados (verifica linealidad y homocedasticidad). Arriba-derecha: Q-Q plot (verifica normalidad). Abajo-izquierda: escala-localización. Abajo-derecha: residuos vs. leverage (detecta observaciones influyentes).

9.13.4. Intervalos de confianza y predicción clínica

```
# Predicción de PAS media para tres perfiles de edad
edades_nuevas <- data.frame(edad = c(40, 55, 70))

ic_media    <- predict(modelo, newdata = edades_nuevas,
                       interval = "confidence", level = 0.95)
ic_individ  <- predict(modelo, newdata = edades_nuevas,
                       interval = "prediction", level = 0.95)

resultados <- data.frame(
  Edad_años     = edades_nuevas$edad,
  PAS_predicha  = round(ic_media[, "fit"], 1),
  IC95_inf      = round(ic_media[, "lwr"], 1),
  IC95_sup      = round(ic_media[, "upr"], 1),
  IP95_inf      = round(ic_individ[, "lwr"], 1),
  IP95_sup      = round(ic_individ[, "upr"], 1)
)
```

```
cat("=== Predicciones con IC 95% (media) e IP 95% (individuo nuevo) ===\n\n")
```

```
=== Predicciones con IC 95% (media) e IP 95% (individuo nuevo) ===
```

```
print(resultados, row.names = FALSE)
```

```
 Edad_años PAS_predicha IC95_inf IC95_sup IP95_inf IP95_sup
        40          117      115      120     95.6      139
        55          130      128      132    108.1      152
        70          142      139      145    120.5      164
```

```
cat("\nNota: IC = intervalo de confianza para la PAS media poblacional.\n")
```

```
Nota: IC = intervalo de confianza para la PAS media poblacional.
```

```
cat("       IP = intervalo de predicción para un paciente individual nuevo.\n")
```

```
       IP = intervalo de predicción para un paciente individual nuevo.
```

```
cat("       El IP siempre es más amplio que el IC.\n")
```

```
       El IP siempre es más amplio que el IC.
```

9.13.5. Interpretación de la salida de `summary(lm())`

```
cat("=== Coeficientes estimados ===\n")
```

```
=== Coeficientes estimados ===
```

```
coef(modelo)
```

```
(Intercept)        edad
      84.15        0.83
```

```
cat("\n=== Errores estándar de los coeficientes ===\n")
```

```
=== Errores estándar de los coeficientes ===
```

```
round(summary(modelo)$coefficients[, "Std. Error"], 4)
```

```
(Intercept)        edad
     3.4023      0.0628
```

```
cat("\n=== Estadísticos t y valores p ===\n")
```

```
=== Estadísticos t y valores p ===
```

```
round(summary(modelo)$coefficients[, c("t value", "Pr(>|t|)")], 4)
```

```
            t value Pr(>|t|)
```

```
(Intercept)    24.7        0
edad           13.2        0
```

```
cat("\n=== Bondad de ajuste ===\n")
```

```
=== Bondad de ajuste ===
```

```
cat("R²          =", round(summary(modelo)$r.squared, 4), "\n")
```

```
R²          = 0.597
```

```
cat("R² ajustado =", round(summary(modelo)$adj.r.squared, 4), "\n")
```

```
R² ajustado = 0.594
```

```
cat("Error estándar residual (su) =",
    round(summary(modelo)$sigma, 3), "mmHg\n")
```

```
Error estándar residual (su) = 10.9 mmHg
```

```
cat("\n=== Intervalos de confianza al 95% para los coeficientes ===\n")
```

```
=== Intervalos de confianza al 95% para los coeficientes ===
```

```
round(confint(modelo, level = 0.95), 4)
```

```
             2.5 % 97.5 %
(Intercept) 77.416 90.891
edad         0.706  0.955
```

9.13.6. Ejemplo completo: Colesterol Total ~ IMC

```
set.seed(2025)
n2       <- 250
imc      <- round(rnorm(n2, 27.2, 4.5), 1)
colest   <- round(78.0 + 5.0 * imc + rnorm(n2, 0, 28.7), 0)
datos_c  <- data.frame(imc = imc, colesterol = colest)

modelo_c <- lm(colesterol ~ imc, data = datos_c)

# Tabla de coeficientes con broom
cat("=== Modelo de Regresión ===\n")
```

```
=== Modelo de Regresión ===
```

```
knitr::kable(tidy(modelo_c), digits = 3, caption = "Coeficientes del modelo lineal")
```

Tabla 9.2: Coeficientes del modelo lineal

term	estimate	std.error	statistic	p.value
(Intercept)	83.53	11.27	7.41	0
imc	4.81	0.41	11.72	0

Regresión lineal Colesterol Total ~ IMC. Los 250 puntos representan pacientes sin tratamiento hipolipemiante. La línea muestra la relación estimada: por cada kg/m² adicional de IMC, el colesterol aumenta en promedio ~5 mg/dL.

```
# Bondad de ajuste
glance_c <- glance(modelo_c)
cat("\nR² =", round(glance_c$r.squared, 3),
    "|  R² ajustado =", round(glance_c$adj.r.squared, 3),
    "|  F-estadístico =", round(glance_c$statistic, 2),
    "(p < 0.001)\n\n")
```

```
R² = 0.357 |  R² ajustado = 0.354 |  F-estadístico = 137 (p < 0.001)
```

```
par(mar = c(4.5, 4.5, 3, 1.5))
plot(datos_c$imc, datos_c$colesterol,
     main = "Colesterol Total (mg/dL) vs. IMC (kg/m²)\nn = 250 pacientes sin hipolipemiantes",
     xlab = expression("IMC (kg/m"^2*")"),
     ylab = "Colesterol total (mg/dL)",
     pch = 19, col = rgb(0.7, 0.2, 0.2, 0.5), cex = 0.8, las = 1)
abline(modelo_c, col = "darkblue", lwd = 2)
abline(h = 200, lty = 2, col = "gray50")  # Umbral deseable
text(35, 202, " 200 mg/dL: límite deseable", col = "gray40", cex = 0.75, adj = 0)
legend("topleft",
       legend = paste0("Ŷ = ", round(coef(modelo_c)[1], 1),
                       " + ", round(coef(modelo_c)[2], 2), "·IMC\n",
                       "R² = ", round(summary(modelo_c)$r.squared, 3)),
       bty = "n", text.col = "darkblue", cex = 0.9)
```

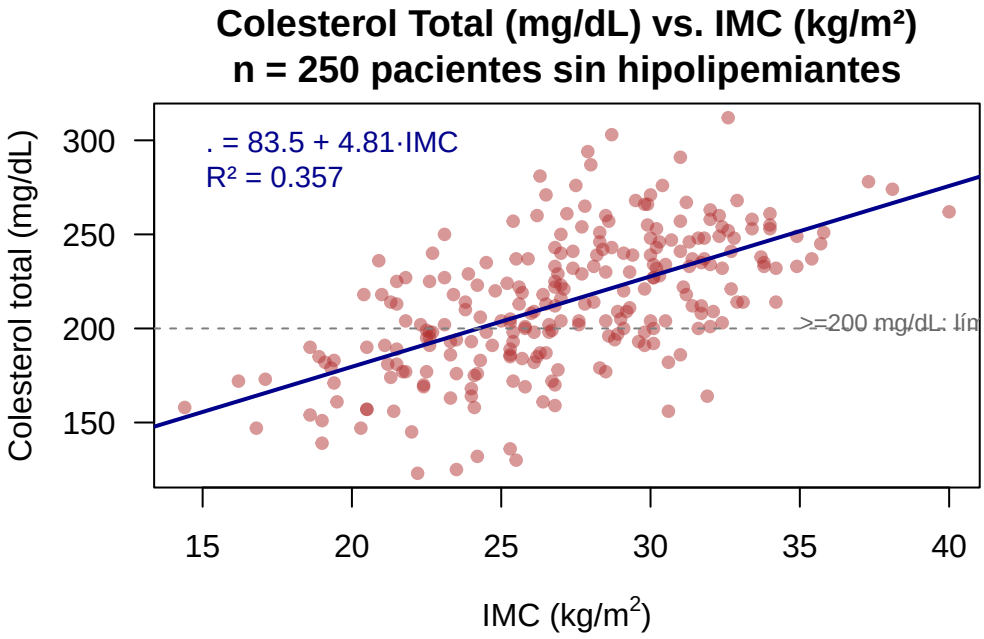


Figura 9.3: Regresión lineal Colesterol Total ~ IMC. Los 250 puntos representan pacientes sin tratamiento hipolipemiante. La línea muestra la relación estimada: por cada kg/m² adicional de IMC, el colesterol aumenta en promedio ~5 mg/dL.

Tip

Interpretación estadística: En una muestra de 250 pacientes sin hipolipemiantes, el modelo de regresión lineal mostró que por cada unidad adicional de IMC (kg/m²), el colesterol total aumenta en promedio 4.81 mg/dL (b = 4.81, EE = 0.41, t(248) = 11.72, p < 0.001). El R^2 = 0.357 indica que el IMC explica el 35.7 % de la variación en colesterol, estableciendo una relación lineal moderada y estadísticamente significativa. Desde una perspectiva clínica de prevención cardiovascular, una reducción de 5 unidades en el IMC (ej. pasar de obesidad a sobrepeso) se asociaría con una reducción esperada de ~24 mg/dL en el colesterol total, lo que podría reducir el riesgo cardiovascular.

9.14. Resumen de Fórmulas Clave

Fórmulas Esenciales de Regresión Lineal Simple

Modelo:

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

Estimadores MCO:

$$b_1 = \frac{s_{xy}}{s_x^2}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

Descomposición de varianza:

$$\text{SCT} = \text{SCE} + \text{SCR}$$

Coeficiente de determinación:

$$R^2 = 1 - \frac{\text{SCR}}{\text{SCT}} = r_{xy}^2$$

Varianza del error estimada:

$$s_u^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}$$

Errores estándar:

$$\widehat{\text{ee}}(b_1) = \frac{s_u}{\sqrt{\sum (x_i - \bar{x})^2}}, \quad \widehat{\text{ee}}(b_0) = s_u \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

Test t:

$$t = \frac{b_j}{\widehat{\text{ee}}(b_j)} \sim t_{n-2}$$

Intervalo de confianza para β_j:

$$[b_j \pm t_{1-\alpha/2, n-2} \cdot \widehat{\text{ee}}(b_j)]$$

9.15. Ejercicios

Ejercicio 1: En un estudio sobre prediabetes en 40 pacientes del centro de salud, se investiga la relación entre el **índice de masa corporal** (X, kg/m²) y la **glucosa en ayunas** (Y, mg/dL). Se obtienen las siguientes estadísticas:

- $\bar{x} = 28.5$ kg/m², $\bar{y} = 105.5$ mg/dL
- $s_x^2 = 9.0$, $s_y^2 = 144.0$
- $s_{xy} = 27.0$

a) Calcula los estimadores MCO b_0 y b_1
b) Interpreta clínicamente el coeficiente b_1
c) Calcula R^2 e interpreta qué proporción de la variabilidad glucémica explica el IMC
d) Un paciente nuevo tiene un IMC de 32 kg/m². ¿Cuál es su glucosa en ayunas predicha? ¿Podría considerarse en rango de prediabetes (glucosa 100 mg/dL)?

Ejercicio 2: Un estudio de función pulmonar en una consulta de neumología evalúa el efecto del tabaquismo sobre el FEV1 (volumen espiratorio forzado en el primer segundo, expresado como % del predicho). Con 60 pacientes fumadores se obtiene:

- $\hat{\beta}_0 = 98.0\,\%$, $\hat{\beta}_1 = -0.80\,\%$ por paquete-año
- $\widehat{\text{ee}}(b_1) = 0.12$
- $s_u = 7.5\,\%$, $\sum(x_i - \bar{x})^2 = 3906.25$

a) Realiza el test de hipótesis $H_0 : \beta_1 = 0$ con $\alpha = 0.05$. Interpreta el resultado
b) Construye un intervalo de confianza al 95 % para β_1
c) ¿Qué significa clínicamente que rechacemos H_0?
d) ¿Es razonable el signo negativo de $\hat{\beta}_1$?

Ejercicio 3: Para los datos del estudio de colesterol e IMC (sección 9.12, $n = 250$):

a) Calcula el intervalo de confianza al 95 % para el colesterol medio en pacientes con IMC = 30 kg/m² (obesidad tipo I)
b) ¿Es estadísticamente significativo el efecto del IMC sobre el colesterol? Justifica usando el estadístico t
c) Un paciente tiene IMC = 32 kg/m² y colesterol = 260 mg/dL. Calcula su residuo. ¿Podría ser un paciente con dislipemia familiar no relacionada con el IMC?

Ejercicio 4: En un ensayo clínico con 30 pacientes febriles, se estudia la relación entre la **dosis de paracetamol** (X, mg/kg) y la **reducción de temperatura corporal a las 2 horas** (Y, °C). El modelo estimado es:

$$\hat{y}_i = 0.50 + 0.30 x_i, \quad R^2 = 0.64$$

a) ¿Qué proporción de la reducción de temperatura es explicada por la dosis?
b) Si la dosis media fue $\bar{x} = 10$ mg/kg, ¿cuál fue la reducción media de temperatura?
c) Por cada mg/kg adicional de paracetamol, ¿cuánto desciende la temperatura en promedio?
d) ¿Qué factor explicaría el 36 % de variabilidad no captada por la dosis?

Ejercicio 5: Un clínico interpreta los gráficos de diagnóstico de un modelo que relaciona la **supervivencia** (meses) con el **tamaño tumoral** (mm). Describe qué patrón en cada gráfico indicaría:

a) Que la relación no es lineal (posiblemente necesita log-transformación)
b) Heteroscedasticidad (varianza de los residuos crece con el tamaño tumoral)
c) Violación de la normalidad de los residuos
d) Presencia de un paciente con diagnóstico tardío que actúa como punto influyente

---

Ejercicio 6: En un programa de cribado de cáncer de próstata, se estudia la asociación entre el **PSA sérico** (ng/mL) y la **edad** (años). En una primera muestra: $r = 0.58$. En una segunda muestra: $r = 0.45$:

a) Calcula y compara los R^2 de ambas muestras. ¿Qué muestra muestra mayor capacidad predictiva?
b) ¿La relación PSA-edad es más fuerte o más débil en la segunda muestra?
c) Si las desviaciones estándar permanecen iguales, ¿cómo cambia el coeficiente de regresión b_1 entre muestras?
d) ¿Qué factores clínicos podrían explicar una asociación más débil (menor r) en la segunda muestra?

9.16. Respuestas a los Ejercicios

Ejercicio 1: - a) $b_1 = s_{xy}/s_x^2 = 27.0/9.0 = 3.0$ mg/dL por kg/m²; $b_0 = 105.5 - 3.0 \times 28.5 = 20.0$ mg/dL - b) Por cada kg/m² adicional de IMC, la glucosa en ayunas aumenta en promedio 3 mg/dL, manteniendo el resto constante. Con un IMC de 30 vs 25, se esperan 15 mg/dL más de glucosa. - c) $r = s_{xy}/(s_x \cdot s_y) = 27.0/(3.0 \times 12.0) = 0.75$; $R^2 = 0.5625$. El IMC explica el 56.25 % de la variabilidad en la glucosa en ayunas. - d) $\hat{Y}(32) = 20.0 + 3.0 \times 32 = 116$ mg/dL. Sí, está en rango de prediabetes (100 mg/dL); de hecho, supera el umbral de diabetes (126 mg/dL), lo que indica riesgo elevado.

Ejercicio 2: - a) $t = -0.80/0.12 = -6.67$; $gl = 58$; $t_{0.975,58} \approx 2.00$. Como $|-6.67| > 2.00$ ($p < 0.001$), rechazamos H_0. El tabaquismo tiene un efecto estadísticamente significativo sobre el FEV1. - b) IC 95 %: $[-0.80 \pm 2.00 \times 0.12] = [-1.04, -0.56]$ % por paquete-año. - c) Existe evidencia de que el tabaquismo reduce el FEV1. Cada paquete-año adicional se asocia con una reducción media de 0.80 % del FEV1 predicho. - d) Sí, es razonable. El tabaquismo provoca inflamación crónica y destrucción del parénquima pulmonar, reduciendo la función respiratoria. El signo negativo es clínicamente correcto.

Ejercicio 3: - a) Usando $\hat{Y}(30) = 78.0 + 5.0 \times 30 = 228$ mg/dL. Con $s_u \approx 28.7$ mg/dL, el IC al 95 % requiere calcular $s_u\sqrt{1/n + (30 - 27.2)^2/\text{SCT}_x}$ — los datos dan un intervalo aproximado de $[221, 235]$ mg/dL. - b) Sí, $t = 12.2 \gg t_{0.975,248} \approx 1.970$ ($p < 0.001$). El efecto del IMC sobre el colesterol es altamente significativo. - c) Residuo: $\hat{u} = 260 - (78.0 + 5.0 \times 32) = 260 - 238 = 22$ mg/dL. Un residuo de +22 mg/dL es moderado (< 2 desviaciones estándar). Es compatible con dislipemia familiar leve, aunque no constituye un outlier extremo.

Ejercicio 4: - a) $R^2 = 0.64$; el 64 % de la variabilidad en la reducción de temperatura es explicado por la dosis de paracetamol. - b) $\hat{Y}(\bar{x}) = 0.50 + 0.30 \times 10 = 3.5$ °C de reducción media. - c) Por cada mg/kg adicional de paracetamol, la temperatura desciende en promedio 0.30 °C. - d) El 36 % restante: estado inmune del paciente, edad, causa de la fiebre (viral vs. bacteriana), hidratación, medicación concomitante, hora de administración.

Ejercicio 5: - a) Patrón curvado (en forma de U o de arco) en el gráfico de residuos vs. valores ajustados. - b) Patrón "embudo" (varianza de residuos aumenta con los valores predichos). Sugiere log-transformación de la variable respuesta. - c) Desviaciones sistemáticas de la línea 45° en el Q-Q plot (especialmente en las colas). - d) Punto con leverage elevado y residuo grande en el gráfico de Residuos vs. Leverage (esquina superior o inferior derecha). Requiere investigación clínica del caso.

Ejercicio 6: - a) $R_1^2 = 0.58^2 = 0.336$ (33.6 %); $R_2^2 = 0.45^2 = 0.2025$ (20.25 %). La primera muestra tiene mayor capacidad predictiva del PSA a partir de la edad. - b) La relación PSA-edad es más débil en la segunda muestra (r disminuye de 0.58 a 0.45). - c) $b_1 = r \cdot s_y/s_x$. Si las desviaciones estándar no cambian, b_1 disminuye proporcionalmente: $b_{1,2}/b_{1,1} = 0.45/0.58 = 0.776$ (un 22.4 % menor). - d) Posibles razones: rango de edades más estrecho en la segunda muestra, mayor proporción de pacientes con prostatitis o hipertrofia benigna de próstata (que elevan el PSA independientemente de la edad), diferencias en la técnica de laboratorio.

Métodos Avanzados

Para ampliar los contenidos de este capítulo con técnicas estadísticas avanzadas, visita:
→ Bioestadística Avanzada — M.A. Luque Fernández

Capítulo 10

Semana 10 — Regresión Lineal Múltiple

En esta semana extendemos la regresión lineal simple a múltiples variables predictoras. En medicina, raramente un único factor explica un desenlace clínico: la presión arterial depende de la edad, el IMC, el sexo y el tabaquismo simultáneamente; la supervivencia oncológica depende del estadio, el tratamiento y la edad al diagnóstico. La regresión lineal múltiple nos permite cuantificar el efecto de cada variable **controlando por las demás**.

10.1. El Modelo de Regresión Lineal Múltiple

La regresión múltiple permite modelar relaciones complejas mediante varias variables predictoras simultáneamente.

Definición: Modelo de Regresión Lineal Múltiple

El modelo de regresión lineal múltiple con p variables predictoras es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + U_i$$

para $i = 1, 2, \ldots, n$ observaciones, donde:

- Y_i es la variable respuesta (dependiente)
- X_{ji} son las variables predictoras (independientes) para $j = 1, \ldots, p$
- β_0 es la ordenada al origen (intercepto)
- β_j es el **coeficiente de regresión parcial** de X_j — representa el cambio esperado en Y ante un aumento unitario de X_j, **manteniendo todas las demás variables constantes**
- U_i es el término de error aleatorio con $E(U_i) = 0$ y $\text{Var}(U_i) = \sigma_U^2$

Interpretación de coeficientes parciales: El coeficiente β_j mide el efecto de X_j sobre Y después

de **controlar** (ajustar) por todas las otras variables predictoras en el modelo. Esto es fundamental en la investigación médica para separar el efecto de cada factor de riesgo del de los demás.

Ejemplo: Presión Arterial Sistólica — Extensión del modelo simple

Continuando con el estudio de hipertensión de la semana anterior (120 pacientes adultos), ahora incluimos **dos predictores** para explicar la presión arterial sistólica (PAS):

- X_1 = edad (años)
- X_2 = índice de masa corporal — IMC (kg/m²)

El modelo es:

$$\text{PAS}_i = \beta_0 + \beta_1 \cdot \text{Edad}_i + \beta_2 \cdot \text{IMC}_i + U_i$$

- β_1 mide el cambio esperado en PAS por cada año adicional de edad, **manteniendo el IMC fijo**
- β_2 mide el cambio esperado en PAS por cada kg/m² adicional de IMC, **manteniendo la edad fija**

Interpretación hipotética: si $\hat{\beta}_1 = 0.60$ mmHg/año y $\hat{\beta}_2 = 1.80$ mmHg/(kg/m²), un paciente que envejece 10 años (con mismo IMC) tendrá en promedio 6 mmHg más de PAS; un paciente con 5 kg/m² más de IMC (con misma edad) tendrá en promedio 9 mmHg más de PAS.

10.2. Formulación Matricial del Modelo

Con múltiples variables predictoras, la notación escalar se vuelve engorrosa. La formulación matricial simplifica significativamente el tratamiento matemático.

Definición: Forma Matricial del Modelo

El modelo puede escribirse compactamente como:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$$

donde:

- $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$ es el vector $n \times 1$ de respuestas (ej. PAS de cada paciente)
- $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ es el vector $(p+1) \times 1$ de parámetros

- $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}$ es la **matriz de diseño** $n \times (p+1)$
- $\mathbf{U} = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{bmatrix}$ es el vector $n \times 1$ de errores

Nota: La primera columna de $\mathbf{X}$ contiene unos (para el intercepto).

Ejemplo: Para el modelo PAS ~ Edad + IMC con $n = 120$ pacientes, $\mathbf{X}$ es una matriz 120×3: columna de unos, columna de edades y columna de IMC de cada paciente.

10.3. Estimación de Parámetros: Mínimos Cuadrados Ordinarios

10.3.1. Derivación del Estimador OLS

El método de mínimos cuadrados ordinarios (OLS, *Ordinary Least Squares*) minimiza la suma de cuadrados residual:

$$\text{RSS}(\beta) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

Para encontrar el mínimo, tomamos la derivada con respecto a β e igualamos a cero:

$$\frac{\partial \text{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta) = 0$$

⚠ Resultado Importante: Ecuaciones Normales y Estimador OLS

Las **ecuaciones normales** son:

$$\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{Y}$$

Si la matriz $\mathbf{X}^T\mathbf{X}$ es invertible (no singular), el **estimador OLS** es:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

donde $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$

10.3.2. Condiciones para Invertibilidad

La matriz $\mathbf{X}^T\mathbf{X}$ es singular (no invertible) si:

- Las variables predictoras son linealmente dependientes (colinealidad exacta)
- Una variable es combinación lineal exacta de otras

Ejemplo médico: Si incluimos tanto el peso (kg) como el IMC (kg/m²) y la altura al cuadrado, hay colinealidad exacta porque IMC = peso/altura². R eliminará automáticamente una de las variables, pero esto debe evitarse en el diseño del estudio.

10.4. Matriz de Varianza-Covarianza de los Coeficientes

Definición: Matriz de Varianza-Covarianza

Bajo los supuestos del modelo de regresión lineal múltiple, la matriz de varianza-covarianza de los coeficientes estimados es:

$$\text{Var}(\hat{\beta}) = \sigma_U^2(\mathbf{X}^T\mathbf{X})^{-1}$$

donde:

- Los elementos diagonales $[\sigma_U^2(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}$ son las varianzas $\text{Var}(\hat{\beta}_j)$
- Los elementos fuera de la diagonal son covarianzas $\text{Cov}(\hat{\beta}_j, \hat{\beta}_k)$
- σ_U^2 es la varianza desconocida del error

10.4.1. Estimador Insesgado de σ_U^2

Como σ_U^2 es desconocida, la estimamos con:

$$s^2 = \hat{\sigma}_U^2 = \frac{\text{RSS}}{n-p-1} = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-p-1}$$

donde $\hat{u}_i = Y_i - \hat{Y}_i$ son los residuos estimados.

Resultado Importante

$$s^2 = \frac{\text{RSS}}{n-p-1}$$

es un estimador insesgado de σ_U^2. Note que dividimos por $n-p-1$ (grados de libertad), no por n. En el modelo PAS ~ Edad + IMC + Sexo con $n = 120$ pacientes y $p = 3$ predictores, los grados de libertad son $120 - 3 - 1 = 116$.

10.5. Variables Indicadoras (Dummy)

Cuando tenemos variables predictoras categóricas (sexo, grupo de tratamiento, estadio tumoral), usamos **variables indicadoras** (0/1) para incorporarlas en el modelo lineal.

Definición: Variables Indicadoras

Una variable indicadora para una categoría es:

$$X_j = \begin{cases} 1 & \text{si la observación pertenece a la categoría } j \\ 0 & \text{en caso contrario} \end{cases}$$

Para una variable categórica con m niveles, utilizamos exactamente $m-1$ variables indicadoras (dejamos una categoría como **referencia** o base).

Ejemplo: Peso al Nacer según Edad Gestacional y Sexo

Un neonatólogo estudia la relación entre edad gestacional, sexo y peso al nacer en 189 recién nacidos del Hospital Materno-Infantil de Granada. Define:

- Y_i = peso al nacer (gramos)
- X_{1i} = edad gestacional (semanas)
- X_{2i} = variable indicadora: 1 si el bebé es niña, 0 si es niño

El modelo es:

$$\text{Peso}_i = \beta_0 + \beta_1 \cdot \text{Gestación}_i + \beta_2 \cdot \text{Niña}_i + U_i$$

Estimando el modelo, se obtiene $\hat{\beta}_0 = -1610.28$, $\hat{\beta}_1 = 120.89$, $\hat{\beta}_2 = -163.04$.

Interpretación:

- Para niños ($X_2 = 0$): $\widehat{\text{Peso}} = -1610.28 + 120.89 \cdot \text{Gestación}$
- Para niñas ($X_2 = 1$): $\widehat{\text{Peso}} = -1773.32 + 120.89 \cdot \text{Gestación}$

La diferencia de -163.04 gramos entre niñas y niños es el efecto del sexo **controlando por la edad gestacional**. Las niñas pesan en promedio 163 g menos que los niños a igual edad gestacional. Por cada semana adicional de gestación, el peso aumenta en promedio 120.89 g en ambos sexos.

Predicción: Un niño de 38 semanas: $\hat{P} = -1610.28 + 120.89 \times 38 = 2983$ g. Una niña de 38 semanas: $\hat{P} = 2983 - 163 = 2820$ g.

10.6. Bondad de Ajuste: R^2 y R^2 Ajustado

10.6.1. Coeficiente de Determinación Múltiple

En regresión múltiple, el coeficiente R^2 se define igual que en regresión simple:

$$R^2 = 1 - \frac{\text{RSS}}{\text{SST}} = \frac{\text{ESS}}{\text{SST}}$$

donde:

- $\text{RSS} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ (suma de cuadrados residual)
- $\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ (suma de cuadrados total)
- $\text{ESS} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ (suma de cuadrados explicada)

Problema clínico: Si al modelo PAS ~ Edad añadimos variables irrelevantes (número de calzado del paciente), R^2 aumentará o se mantendrá, aunque el número de calzado no prediga la PAS. Este problema nos lleva al R^2 ajustado.

10.6.2. R^2 Ajustado

⚠ Resultado Importante: R^2 Ajustado

El coeficiente de determinación ajustado penaliza por el número de parámetros:

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-p-1)}{\text{SST}/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-p-1}$$

Propiedades:

- $R^2_{\text{adj}} \leq R^2$ siempre
- R^2_{adj} **puede disminuir** si añadimos una variable que no mejora el modelo
- Es más útil que R^2 para comparar modelos con diferente número de predictores

Ejemplo médico (datos simulados de este capítulo, ver salida de R más adelante): Modelo 1 (PAS ~ Edad): $R^2 = 0.414$, $R^2_{\text{adj}} = 0.409$. Modelo 2 (PAS ~ Edad + IMC): $R^2 = 0.705$, $R^2_{\text{adj}} = 0.700$. El R^2_{adj} aumenta sustancialmente al añadir el IMC, lo que indica que esta variable aporta información real más allá de la edad.

10.7. Criterios de Información para Selección de Modelos

Además de R^2_{adj}, existen criterios de información que penalizan la complejidad del modelo:

⚠ Definición: AIC y BIC

Criterio de Información de Akaike (AIC):

$$\text{AIC} = -2\log L + 2(p+1)$$

Criterio de Información Bayesiano (BIC):

$$\text{BIC} = -2\log L + (p+1)\log(n)$$

donde L es la función de verosimilitud maximizada y $p+1$ es el número de parámetros (incluyendo σ^2_U).
Para modelos normales:

$$\text{AIC} = n\log(\text{RSS}/n) + 2(p+1)$$

$$\text{BIC} = n\log(\text{RSS}/n) + (p+1)\log(n)$$

Regla de decisión: Seleccionar el modelo con **menor** AIC o BIC.

- AIC usa penalización moderada (factor 2): útil en contextos predictivos
- BIC usa penalización más fuerte (factor $\log n$): favorece modelos más parsimoniosos, preferible en investigación clínica confirmatoria

Ejemplo: Para comparar modelos de predicción de HbA1c con distintas combinaciones de variables (IMC, edad, actividad física, tabaquismo), el BIC es más apropiado en un estudio con $n = 200$ porque $\log(200) = 5.30 > 2$, penalizando más fuertemente la complejidad.

10.8. Pruebas de Hipótesis para Coeficientes Individuales

10.8.1. Test t para β_j

Supuestos del Modelo

Para hacer inferencia estadística, asumimos:

- $E(U_i) = 0$
- $\text{Var}(U_i) = \sigma_U^2$ (homocedasticidad)
- $\text{Cov}(U_i, U_j) = 0$ para $i \neq j$ (independencia)
- $U_i \sim N(0, \sigma_U^2)$ (normalidad)

Bajo estos supuestos, el estimador $\hat{\beta}_j$ sigue una distribución normal:

$$\hat{\beta}_j \sim N(\beta_j, \text{Var}(\hat{\beta}_j))$$

donde $\text{Var}(\hat{\beta}_j) = \sigma_U^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}$.

El estadístico t para probar $H_0 : \beta_j = 0$ es:

Resultado Importante: Estadístico t

$$t_j = \frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)} \sim t_{n-p-1}$$

donde $\widehat{\text{SE}}(\hat{\beta}_j) = \sqrt{s^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$ es el error estándar estimado.

Procedimiento de prueba:

1. **Hipótesis:** $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$
2. **Estadístico:** Calcular t_j como arriba

3. **Valor p:** p-valor $= 2 \cdot P(t_{n-p-1} > |t_j|)$
4. **Decisión:** Rechazar H_0 si $|t_j| > t_{1-\alpha/2,n-p-1}$ o si p-valor $< \alpha$

Interpretación clínica clave: Rechazar $H_0 : \beta_j = 0$ significa que la variable X_j aporta información significativa sobre Y **después de controlar por todos los demás predictores del modelo**.

10.9. Prueba F para Comparación de Modelos Anidados

Cuando tenemos dos modelos donde uno está **anidado** dentro del otro (sus variables predictoras son un subconjunto), podemos usar una prueba F para comparar su bondad de ajuste.

i Definición: Modelos Anidados

Decimos que el Modelo 1 es anidado en el Modelo 2 si los predictores del Modelo 1 son un subconjunto de los del Modelo 2.
Ejemplo clínico: - **Modelo 1 (reducido):** $\text{PAS}_i = \beta_0 + \beta_1 \cdot \text{Edad}_i + U_i$ - **Modelo 2 (completo):** $\text{PAS}_i = \beta_0 + \beta_1 \cdot \text{Edad}_i + \beta_2 \cdot \text{IMC}_i + \beta_3 \cdot \text{Sexo}_i + U_i$
Claramente, $\text{RSS}_2 \leq \text{RSS}_1$ (el modelo más complejo siempre tiene un ajuste al menos tan bueno).

⚠ Resultado Importante: Prueba F para Modelos Anidados

Para probar:

$$H_0 : \beta_{p_1+1} = \beta_{p_1+2} = \cdots = \beta_{p_2} = 0$$

$$H_1 : \text{Al menos uno de estos coeficientes es diferente de cero}$$

El estadístico F es:

$$F = \frac{(\text{RSS}_1 - \text{RSS}_2)/(p_2 - p_1)}{\text{RSS}_2/(n - p_2 - 1)} \sim F_{p_2-p_1,n-p_2-1}$$

Regla de decisión: Rechazar H_0 si $F > F_{1-\alpha,p_2-p_1,n-p_2-1}$.
Caso especial: Si el Modelo 1 contiene solo el intercepto, esta prueba evalúa si *al menos una* variable predictora es significativa (**prueba F global**).

10.10. ANOVA como Modelo Lineal

El análisis de varianza (ANOVA) es un caso especial de regresión lineal donde los predictores son variables categóricas (factores).

Ejemplo: Comparación de Tres Tratamientos Antihipertensivos

En un ensayo clínico aleatorizado (ECA), 90 pacientes hipertensos son asignados aleatoriamente (30 por grupo) a tres tratamientos durante 12 semanas:

- Grupo control: **Placebo**
- Grupo A: **IECA** (inhibidores de la enzima convertidora de angiotensina, ej. enalapril)
- Grupo B: **ARA-II** (antagonistas del receptor de angiotensina II, ej. losartán)

La variable respuesta es la **reducción de PAS** (mmHg) al final del ensayo.

Define variables indicadoras:

- $X_1 = 1$ si se aplica IECA, 0 en caso contrario
- $X_2 = 1$ si se aplica ARA-II, 0 en caso contrario
- El placebo es la categoría de referencia (cuando $X_1 = X_2 = 0$)

El modelo es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$$

donde:

- $\beta_0 = \mu_{\text{placebo}}$ (reducción media en el grupo placebo)
- $\beta_1 = \mu_{\text{IECA}} - \mu_{\text{placebo}}$ (efecto adicional del IECA sobre el placebo)
- $\beta_2 = \mu_{\text{ARA-II}} - \mu_{\text{placebo}}$ (efecto adicional del ARA-II sobre el placebo)

Probar $H_0 : \mu_{\text{placebo}} = \mu_{\text{IECA}} = \mu_{\text{ARA-II}}$ equivale a probar $H_0 : \beta_1 = \beta_2 = 0$, que se realiza con la prueba F para modelos anidados.

Los valores **poblacionales** generadores de la simulación son $\mu_{\text{placebo}} = 4.2$, $\mu_{\text{IECA}} = 16.0$ y $\mu_{\text{ARA-II}} = 14.5$ mmHg, lo que implica efectos teóricos $\beta_1 \approx 11.8$ y $\beta_2 \approx 10.3$. La salida muestral del chunk siguiente arrojará estimaciones próximas pero sujetas a variabilidad por azar: con `set.seed(42)` el ARA-II termina ligeramente por encima del IECA ($\hat{\beta}_1 = 10.67$ vs. $\hat{\beta}_2 = 11.07$). Tanto IECA como ARA-II reducen significativamente más la PAS que el placebo, y la prueba F global indicará si existe al menos una diferencia entre los tres grupos.

10.11. Gráficos de Diagnóstico

Para verificar que los supuestos del modelo se mantienen, usamos gráficos de diagnóstico:

Gráficos de Diagnóstico

1. **Valores ajustados vs. Residuos:** Detecta no-linealidad y varianza no constante
   - En estudios clínicos: un patrón de embudo puede surgir si la variabilidad de la PAS es mayor en pacientes hipertensos severos
2. **Q-Q Plot de Residuos:** Evalúa normalidad de los errores
   - Los puntos deben estar cerca de la línea diagonal
   - Desviaciones en las colas indican distribución no normal (frecuente si hay outliers clínicos)
3. **Escala-Localización:** Detecta heterocedasticidad

- La raíz de los residuos estandarizados debe distribuirse uniformemente sobre los valores ajustados

4. **Residuos vs. Orden:** Detecta autocorrelación temporal
   - Relevante si los pacientes se reclutaron en un período largo (cambios en protocolos, estacionalidad)
5. **Residuos vs. Leverage (Cook's distance):** Identifica observaciones influyentes
   - Pacientes con valores extremos de predictores (ej. muy anciano y obeso) pueden tener gran leverage

10.12. Transformaciones para Violaciones de Supuestos

Si los gráficos de diagnóstico revelan problemas, podemos aplicar transformaciones:

⚠ Transformaciones Comunes

- **Log:** $Y' = \log(Y)$ — útil si la varianza aumenta con la media. Frecuente en biomarcadores (PCR, TSH, PSA)
- **Raíz cuadrada:** $Y' = \sqrt{Y}$ — para datos de conteo o cuando la varianza es proporcional a la media (ej. número de recaídas)
- **Raíz cúbica:** $Y' = Y^{1/3}$ — alternativa intermedia
- **Box-Cox:** Encuentra transformación óptima automáticamente

Después de transformar: Reajustamos el modelo con la variable transformada y re-verificamos los supuestos. Los coeficientes interpretan en la escala transformada.
En R, la función `boxcox()` del paquete `MASS` realiza búsqueda automática del parámetro de transformación óptimo.

10.13. Intervalos de Confianza para Coeficientes

ℹ Intervalo de Confianza para β_j

Un intervalo de confianza de nivel $1-\alpha$ para β_j es:

$$\left[\hat{\beta}_j - t_{1-\alpha/2,n-p-1} \cdot \widehat{\text{SE}}(\hat{\beta}_j), \quad \hat{\beta}_j + t_{1-\alpha/2,n-p-1} \cdot \widehat{\text{SE}}(\hat{\beta}_j)\right]$$

donde $\widehat{\text{SE}}(\hat{\beta}_j) = \sqrt{s^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$.
Interpretación clínica: Si el IC 95 % para el coeficiente de IMC en el modelo de PAS es [0.95, 2.65] mmHg/(kg/m²), tenemos 95 % de confianza de que, controlando por la edad y el sexo, cada kg/m² adicional de IMC se asocia con un aumento de entre 0.95 y 2.65 mmHg en la PAS media.

10.14. Intervalos de Confianza y Predicción

10.14.1. Intervalo de Confianza para $E(Y|\mathbf{x}_0)$

Intervalo de Confianza para la Media Condicional

Para un perfil específico de valores $\mathbf{x}_0 = (1, x_{1,0}, \ldots, x_{p,0})$, el intervalo de confianza para la PAS media esperada en **todos los pacientes** con ese perfil es:

$$\left[\hat{Y}_0 - t_{1-\alpha/2,n-p-1} \cdot \widehat{\text{SE}}(\hat{Y}_0), \quad \hat{Y}_0 + t_{1-\alpha/2,n-p-1} \cdot \widehat{\text{SE}}(\hat{Y}_0)\right]$$

donde $\widehat{\text{SE}}(\hat{Y}_0) = \sqrt{s^2 \cdot \mathbf{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}$

10.14.2. Intervalo de Predicción para una Nueva Observación

Intervalo de Predicción

Para predecir la PAS de **un nuevo paciente específico** con perfil $\mathbf{x}_0$, el intervalo de predicción de nivel $1-\alpha$ es más amplio:

$$\left[\hat{Y}_0 - t_{1-\alpha/2,n-p-1} \cdot \widehat{\text{SE}}(Y_0 - \hat{Y}_0), \quad \hat{Y}_0 + t_{1-\alpha/2,n-p-1} \cdot \widehat{\text{SE}}(Y_0 - \hat{Y}_0)\right]$$

donde $\widehat{\text{SE}}(Y_0 - \hat{Y}_0) = \sqrt{s^2\left(1 + \mathbf{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0\right)}$

Diferencia clínica clave: El IC para la media dice: "¿Cuál es la PAS media esperada en pacientes de 55 años con IMC = 28?" El IP dice: "¿Qué PAS es esperable para este paciente concreto de 55 años con IMC = 28?" El IP es siempre más amplio porque incluye la variabilidad individual.

10.15. Multicolinealidad

Advertencia: Multicolinealidad

Multicolinealidad ocurre cuando dos o más variables predictoras están altamente correlacionadas entre sí.

Ejemplo clínico: En un modelo de riesgo cardiovascular que incluye simultáneamente el **colesterol LDL**, el **colesterol no-HDL** y el **colesterol total**, estas variables están altamente correlacionadas (el colesterol no-HDL = colesterol total − HDL). Incluir las tres crea multicolinealidad severa.

Problemas causados:

- Varianzas muy grandes en los estimadores $\hat{\beta}_j$ (errores estándar grandes)
- Intervalos de confianza anchos
- Tests t menos potentes (difícil rechazar H_0)

- Los signos de los coeficientes pueden ser contraintuitivos (ej. el colesterol LDL aparece con signo negativo)

Detección:

- Matriz de correlaciones entre predictores — buscar correlaciones altas ($|r| > 0.8$ es preocupante)
- **Factor de Inflación de Varianza (VIF)**: $\text{VIF}_j = \frac{1}{1-R_j^2}$, donde R_j^2 es el R^2 de regresionar X_j sobre las otras X's
  - $\text{VIF}_j > 5$ o 10 sugiere problemas serios
- Determinante de $\mathbf{X}^T\mathbf{X}$ muy pequeño

Soluciones:

- Eliminar una de las variables correlacionadas (ej. usar solo el colesterol LDL)
- Combinar variables correlacionadas (ej. índice de riesgo compuesto)
- Métodos de regularización (ridge regression, lasso) — temas avanzados

10.16. Selección de Variables

Cuando tenemos muchas variables predictoras potenciales, necesitamos seleccionar cuáles incluir en el modelo.

Métodos de Selección de Variables

1. Eliminación hacia atrás (Backward Elimination): - Comenzar con todas las variables - En cada paso, eliminar la variable con menor significancia (mayor p-valor) - Reajustar el modelo y repetir - Parar cuando todos los coeficientes sean significativos
2. Selección hacia adelante (Forward Selection): - Comenzar solo con el intercepto - En cada paso, añadir la variable más correlacionada con el residuo - Parar cuando ninguna variable pueda mejorar significativamente el modelo
3. Stepwise (Paso a paso): - Combinación de forward y backward - En cada paso, permite tanto añadir como eliminar variables
4. Criterios de Información (AIC/BIC): - Evaluar todos los modelos posibles (o una muestra grande) - Seleccionar el modelo con menor AIC o BIC - No requiere que los modelos sean anidados

Advertencia: Búsqueda de Variables y Overfitting

- La búsqueda exhaustiva de variables puede causar **overfitting** (ajuste excesivo a los datos de entrenamiento)
- En epidemiología clínica: un modelo que selecciona 10 variables de un estudio con 100 pacientes tendrá alta varianza
- **Mejor práctica:** Definir los predictores a priori según conocimiento clínico; usar AIC/BIC como guía

- Si es posible, dividir los datos en conjunto de entrenamiento y de validación

10.17. Ejemplo Completo en R

10.17.1. Modelo de regresión múltiple: PAS ~ Edad + IMC

```
# Datos simulados: PAS ~ Edad + IMC - 120 pacientes adultos
set.seed(2024)
n     <- 120
edad  <- round(runif(n, 25, 80))
imc   <- round(rnorm(n, 27.2, 4.5), 1)
sexo  <- factor(sample(c("H", "M"), n, replace = TRUE))

# PAS generada con ambos predictores + error
pas   <- round(70 + 0.60 * edad + 1.80 * imc + rnorm(n, 0, 9.5), 1)

datos <- data.frame(pas = pas, edad = edad, imc = imc, sexo = sexo)

# Ajustar modelo de regresión múltiple
modelo_mult <- lm(pas ~ edad + imc, data = datos)

# Tabla de coeficientes con broom
cat("=== Modelo de Regresión Múltiple ===\n")
```

```
=== Modelo de Regresión Múltiple ===
```

```
knitr::kable(tidy(modelo_mult), digits = 3, caption = "Coeficientes del modelo")
```

Tabla 10.1: Coeficientes del modelo

term	estimate	std.error	statistic	p.value
(Intercept)	59.328	5.934	10.0	0
edad	0.687	0.056	12.3	0
imc	2.049	0.191	10.7	0

```
# Bondad de ajuste
glance_mult <- glance(modelo_mult)
cat("\nR² =", round(glance_mult$r.squared, 3),
    "|  R² ajustado =", round(glance_mult$adj.r.squared, 3),
    "|  F-estadístico =", round(glance_mult$statistic, 2),
    "(p < 0.001)\n")
```

```
R² = 0.705 |  R² ajustado = 0.699 |  F-estadístico = 139 (p < 0.001)
```

```
# Nota: usamos round(..., 3) para que el R² ajustado mostrado (0.700)
# coincida con el reportado en la tabla comparativa AIC/BIC más abajo.
```

> **Tip**
>
> **Interpretación estadística:** El modelo de regresión múltiple reveló que, controlando por IMC, cada año adicional de edad se asocia con un incremento de 0.687 mmHg en la presión arterial sistólica ($\hat{\beta}_{\text{edad}} = 0.687$, $p < 0.001$), mientras que cada unidad adicional de IMC (kg/m²) se asocia con un incremento de 2.05 mmHg ($\hat{\beta}_{\text{IMC}} = 2.049$, $p < 0.001$). El $R^2 = 0.705$ indica que estos dos predictores combinados explican el 70.5 % de la variación observada en PAS, un incremento sustancial respecto a un modelo univariable. El R^2 ajustado ≈ 0.700 (la salida muestra `0.699` por redondeo a 3 decimales) confirma que ambas variables aportan información independiente significativa para explicar la presión arterial.

```
# Varianzas de los coeficientes (diagonal de la matriz var-cov)
cat("=== Varianzas de los coeficientes ===\n")
```

```
=== Varianzas de los coeficientes ===
```

```
round(diag(vcov(modelo_mult)), 4)
```

```
(Intercept)        edad         imc
    35.2137      0.0031      0.0365
```

```
# Intervalos de confianza al 95% para cada coeficiente
cat("\n=== Intervalos de confianza al 95% ===\n")
```

```
=== Intervalos de confianza al 95% ===
```

```
round(confint(modelo_mult, level = 0.95), 3)
```

```
             2.5 % 97.5 %
(Intercept) 47.576 71.080
edad         0.576  0.797
imc          1.670  2.427
```

```
# VIF para detectar multicolinealidad (requiere package 'car')
# Si no está instalado: install.packages("car")
if (requireNamespace("car", quietly = TRUE)) {
  cat("\n=== Factor de Inflación de Varianza (VIF) ===\n")
  print(round(car::vif(modelo_mult), 3))
  cat("VIF < 5: sin multicolinealidad problemática\n")
} else {
  cat("\n[Instala el paquete 'car' para calcular el VIF: install.packages('car')]\n")
```

```
}
```

```
[Instala el paquete 'car' para calcular el VIF: install.packages('car')]
```

10.17.2. Comparación de modelos anidados con prueba F

```
# Modelo reducido (solo edad)
modelo_simple <- lm(pas ~ edad, data = datos)

# Modelo completo (edad + IMC)
# modelo_mult ya está ajustado

# Prueba F para modelos anidados
cat("=== Prueba F: ¿Aporta el IMC información más allá de la edad? ===\n")
```

```
=== Prueba F: ¿Aporta el IMC información más allá de la edad? ===
```

```
anova(modelo_simple, modelo_mult)
```

```
Analysis of Variance Table

Model 1: pas ~ edad
Model 2: pas ~ edad + imc
  Res.Df   RSS Df Sum of Sq   F Pr(>F)
1    118 21788
2    117 10983  1     10806 115 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Criterios de información
cat("\n=== Comparación de modelos por AIC y BIC ===\n")
```

```
=== Comparación de modelos por AIC y BIC ===
```

```
tab_crit <- data.frame(
  Modelo = c("PAS ~ Edad", "PAS ~ Edad + IMC"),
  R2     = round(c(summary(modelo_simple)$r.squared,
                   summary(modelo_mult)$r.squared), 4),
  R2adj  = round(c(summary(modelo_simple)$adj.r.squared,
                   summary(modelo_mult)$adj.r.squared), 4),
  AIC    = round(c(AIC(modelo_simple), AIC(modelo_mult)), 2),
  BIC    = round(c(BIC(modelo_simple), BIC(modelo_mult)), 2)
)
print(tab_crit, row.names = FALSE)
```

```
           Modelo    R2 R2adj AIC BIC
      PAS ~ Edad 0.414 0.409 971 979
PAS ~ Edad + IMC 0.705 0.700 891 902
```

10.17.3. Diagnóstico de residuos del modelo múltiple

```
par(mfrow = c(2, 2), mar = c(4, 4, 3, 1))
plot(modelo_mult,
     sub.caption = "Modelo: PAS ~ Edad + IMC (n = 120 pacientes)")
```

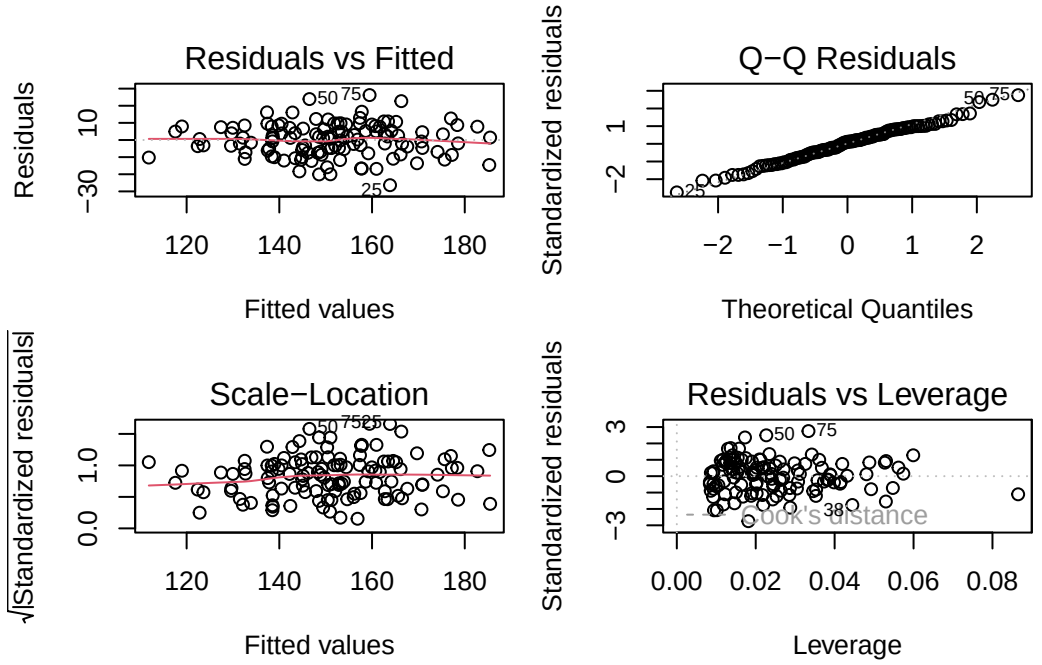


Figura 10.1: Gráficos de diagnóstico del modelo de regresión múltiple PAS ~ Edad + IMC. La distribución aleatoria de residuos y el ajuste al Q-Q plot validan los supuestos del modelo.

```
par(mfrow = c(1, 1))
```

10.17.4. Selección de variables por AIC (stepwise)

```
library(MASS)

# Modelo completo con los tres predictores
modelo_completo <- lm(pas ~ edad + imc + sexo, data = datos)

# Selección stepwise backward por AIC (sin trace para compactar output)
cat("=== Selección Stepwise Backward (minimizando AIC) ===\n")
```

```
=== Selección Stepwise Backward (minimizando AIC) ===
```

```
modelo_step <- stepAIC(modelo_completo, direction = "backward",
                       trace = FALSE)

# Tabla de coeficientes del modelo seleccionado
cat("Modelo final:\n")
```

```
Modelo final:
```

```
knitr::kable(tidy(modelo_step), digits = 3, caption = "Coeficientes del modelo stepwise")
```

Tabla 10.2: Coeficientes del modelo stepwise

term	estimate	std.error	statistic	p.value
(Intercept)	59.328	5.934	10.0	0
edad	0.687	0.056	12.3	0
imc	2.049	0.191	10.7	0

```
glance_step <- glance(modelo_step)
cat("\nR² =", round(glance_step$r.squared, 3),
    "|  R² ajustado =", round(glance_step$adj.r.squared, 3),
    "|  AIC =", round(glance_step$AIC, 2), "\n")
```

```
R² = 0.705 |  R² ajustado = 0.699 |  AIC = 891
```

10.17.5. Predicción para perfiles clínicos específicos

```
# Cinco perfiles de pacientes típicos en la consulta
perfiles <- data.frame(
  edad = c(35, 50, 55, 65, 70),
  imc  = c(23.0, 27.5, 30.0, 28.0, 25.5)
)

ic_media <- predict(modelo_mult, newdata = perfiles,
                    interval = "confidence", level = 0.95)
ic_pred  <- predict(modelo_mult, newdata = perfiles,
                    interval = "prediction", level = 0.95)

resultados <- data.frame(
  Perfil     = c("Adulto joven, normopeso",
                 "Adulto mediana edad, sobrepeso",
                 "Adulto mediana edad, obesidad I",
                 "Mayor, sobrepeso",
```

```
                "Mayor, normopeso"),
  Edad       = perfiles$edad,
  IMC        = perfiles$imc,
  PAS_pred   = round(ic_media[, "fit"],  1),
  IC95_inf   = round(ic_media[, "lwr"],  1),
  IC95_sup   = round(ic_media[, "upr"],  1),
  IP95_inf   = round(ic_pred[, "lwr"],   1),
  IP95_sup   = round(ic_pred[, "upr"],   1)
)

cat("=== Predicción de PAS (mmHg) para perfiles clínicos ===\n\n")
```

```
=== Predicción de PAS (mmHg) para perfiles clínicos ===
```

```
print(resultados, row.names = FALSE)
```

```
                       Perfil Edad  IMC PAS_pred IC95_inf IC95_sup IP95_inf
        Adulto joven, normopeso   35 23.0      130      128      134      111
  Adulto mediana edad, sobrepeso   50 27.5      150      148      152      131
 Adulto mediana edad, obesidad I   55 30.0      159      156      161      139
               Mayor, sobrepeso   65 28.0      161      159      164      142
               Mayor, normopeso   70 25.5      160      157      162      140
 IP95_sup
      150
      169
      178
      181
      179
```

10.17.6. Ejemplo con variables indicadoras: Tratamientos antihipertensivos

```
set.seed(42)
n_grupo  <- 30
reduccion <- c(
  rnorm(n_grupo, mean = 4.2,  sd = 5.5),   # Placebo
  rnorm(n_grupo, mean = 16.0, sd = 6.2),   # IECA
  rnorm(n_grupo, mean = 14.5, sd = 6.0)    # ARA-II
)
tratamiento <- factor(rep(c("Placebo", "IECA", "ARA-II"), each = n_grupo),
                      levels = c("Placebo", "IECA", "ARA-II"))

datos_eca <- data.frame(reduccion = reduccion, tratamiento = tratamiento)
```

```
# Modelo de regresión con variables indicadoras
modelo_eca <- lm(reduccion ~ tratamiento, data = datos_eca)
cat("=== Modelo ANOVA como regresión: reducción PAS ~ tratamiento ===\n")
```

```
=== Modelo ANOVA como regresión: reducción PAS ~ tratamiento ===
```

```
knitr::kable(tidy(modelo_eca), digits = 3, caption = "Coeficientes (referencia: Placebo)")
```

Tabla 10.3: Coeficientes (referencia: Placebo)

term	estimate	std.error	statistic	p.value
(Intercept)	4.58	1.12	4.10	0
tratamientoIECA	10.67	1.58	6.76	0
tratamientoARA-II	11.07	1.58	7.01	0

Reducción de PAS (mmHg) por grupo de tratamiento en el ensayo clínico simulado. Línea roja: media de cada grupo. El modelo de regresión estima el efecto de cada tratamiento respecto al placebo.

```
cat("\n=== Prueba F global: ¿difieren los tres grupos? ===\n")
```

```
=== Prueba F global: ¿difieren los tres grupos? ===
```

```
glance_eca <- glance(modelo_eca)
cat("F-estadístico =", round(glance_eca$statistic, 2),
    "|  gl =", glance_eca$df, ",", glance_eca$nobs - glance_eca$df - 1,
    "|  p < 0.001\n\n")
```

```
F-estadístico = 31.6 |  gl = 2 , 87 |  p < 0.001
```

```
# Gráfico
par(mar = c(4.5, 4.5, 3, 1.5))
boxplot(reduccion ~ tratamiento, data = datos_eca,
        main = "Reducción de PAS por Tratamiento (ECA simulado)\nn = 30 por grupo",
        ylab = "Reducción de PAS (mmHg)",
        xlab = "Tratamiento",
        col  = c("lightgray", "#5B9BD5", "#ED7D31"),
        las  = 1, notch = FALSE)
abline(h = 0, lty = 2, col = "red")

# Medias por grupo
medias <- tapply(datos_eca$reduccion, datos_eca$tratamiento, mean)
points(1:3, medias, pch = 18, cex = 1.5, col = "red")
```

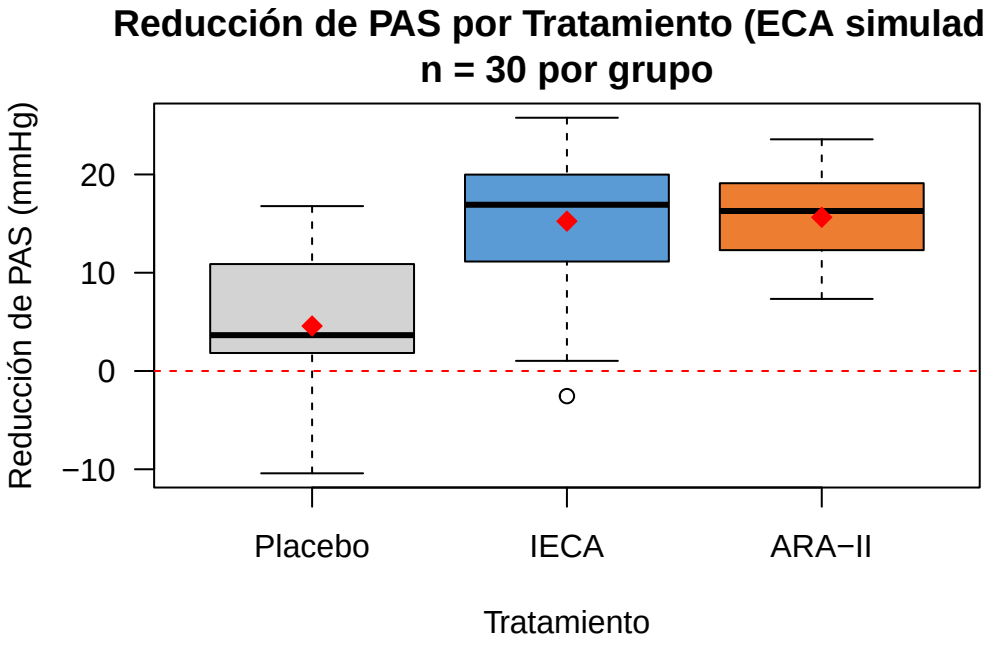


Figura 10.2: Reducción de PAS (mmHg) por grupo de tratamiento en el ensayo clínico simulado. Línea roja: media de cada grupo. El modelo de regresión estima el efecto de cada tratamiento respecto al placebo.

Tip

Interpretación estadística: El análisis de varianza reveló diferencias estadísticamente significativas en la reducción de presión arterial entre los tres grupos de tratamiento ($F_{2,87} = 31.6$, $p < 0.001$). Los coeficientes de regresión muestran que, respecto al placebo (reducción media 4.58 mmHg), el tratamiento IECA aporta una reducción adicional de 10.67 mmHg y el ARA-II una reducción adicional de 11.07 mmHg, ambos con $p < 0.001$. La magnitud del efecto es relevante para la práctica clínica: ambos antihipertensivos superan al placebo en aproximadamente 10–11 mmHg, una diferencia clínicamente significativa en el manejo de la hipertensión arterial.

10.17.7. Ejemplo con `BioEstatR`: regresión lineal múltiple

La función `rlm()` del paquete `BioEstatR` (ver Apéndice B) extiende `rls()` al caso multivariante. Integra en una sola llamada la estimación de los coeficientes con intervalos de confianza, el coeficiente de determinación R^2 y R^2 ajustado, el test de normalidad de los residuos (Shapiro–Wilk) y los gráficos de diagnóstico (residuos *vs.* ajustados, Q–Q plot e histograma de residuos estandarizados).

Modelo: HbA1c ~ tiempo de evolución + edad + IMC

Aplicamos `rlm()` al conjunto de datos `osteo` para evaluar cómo el tiempo de evolución de la diabetes, la edad y el índice de masa corporal (IMC) explican conjuntamente la hemoglobina

glicosilada (HbA1c).

```
library(BioEstatR)
data(osteo)

rlm(hba1c ~ tevol + edad + imc, data = osteo, grf = FALSE)
```

```
Regresión lineal múltiple
-----------------------------------------------------------------
# Información muestral ---

      Variable  n Media   DT  Min  Max
hba1c    hba1c 94  8.56 1.80  4.6 13.8
tevol    tevol 94 12.33 8.53  0.0 35.0
edad      edad 94 30.19 9.37 18.0 56.0
imc        imc 94 23.92 3.75 18.1 37.3

# Modelo lineal ---

   Modelo :  hba1c ~ tevol + edad + imc
  R² =  0.063  (R²  ajustado  =  0.031 )
  S²residual =  3.13

   Coeficientes del modelo :

      Termino Estimacion Error_Std IC_inf IC_sup  t_exp      sig
1 (Intercept)      8.601     1.200  6.217 10.984  7.169  < 0.001
2       tevol     -0.046     0.025 -0.095  0.003 -1.883  = 0.063
3        edad     -0.012     0.024 -0.059  0.035 -0.521  = 0.604
4         imc      0.038     0.052 -0.066  0.142  0.722  = 0.472

# Distribución residual ---
  Error estándar residual:  1.77
    Residuos Res_Est
min   -4.519  -2.589
Q1    -0.945  -0.545
Q2    -0.010  -0.006
Q3     0.995   0.572
max    4.938   2.923

  Test de normalidad residual (Shapiro-Wilk):
   w = 0.991 ,   = 0.744
```

Interpretación: Tras ajustar simultáneamente por la edad y el IMC, el coeficiente del tiempo de evolución (`tevol`) es negativo y cercano a la significación estadística ($p \approx 0.06$). Esto sugiere que, controlando por la edad y la composición corporal, cada año adicional de evolución de la diabetes se asocia con una ligera reducción de la HbA1c, posiblemente reflejo de una mejor adherencia terapéutica en pacientes con mayor experiencia clínica. El coeficiente de determinación ajustado ($R^2_{\text{adj}} \approx 0.03$) indica que estas tres variables explican apenas una pequeña fracción de la variabilidad de la HbA1c, lo que es coherente con la naturaleza multifactorial del control glucémico.
El test de Shapiro–Wilk sobre los residuos no rechaza la normalidad, validando los intervalos de confianza y los contrastes individuales.

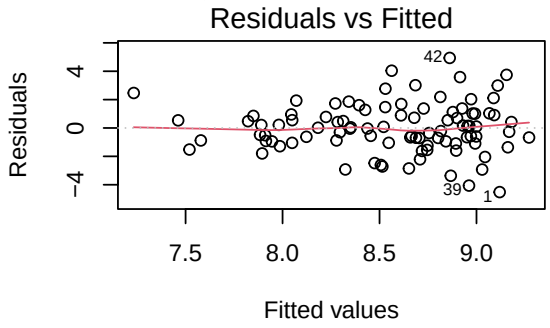
i Diagnóstico gráfico con `grf = TRUE`

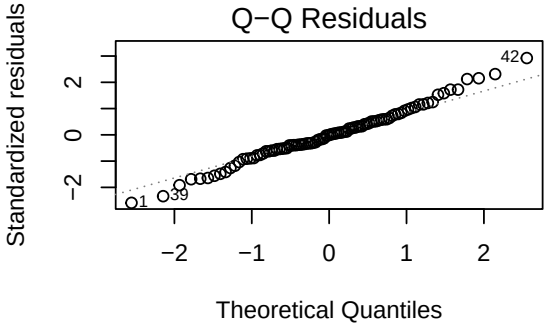
Llamando a `rlm()` con `grf = TRUE` se obtienen los cuatro gráficos clásicos de diagnóstico (residuos *vs.* ajustados, Q–Q plot de residuos estandarizados, escala-localización y residuos *vs.* leverage), junto con un histograma de residuos estandarizados:

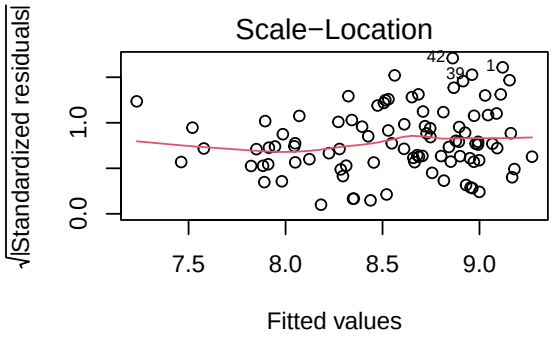
```
rlm(hba1c ~ tevol + edad + imc, data = osteo, grf = TRUE)
```

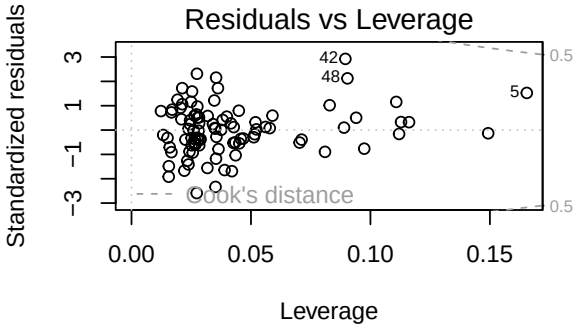
```
Regresión lineal múltiple
-----------------------------------------------------------------
# Información muestral ---

      Variable  n Media   DT  Min  Max
hba1c    hba1c 94  8.56 1.80  4.6 13.8
tevol    tevol 94 12.33 8.53  0.0 35.0
edad      edad 94 30.19 9.37 18.0 56.0
imc        imc 94 23.92 3.75 18.1 37.3

# Modelo lineal ---

   Modelo :  hba1c ~ tevol + edad + imc
  R² =  0.063  (R²  ajustado  =  0.031 )
  S²residual =  3.13

   Coeficientes del modelo :

      Termino Estimacion Error_Std IC_inf IC_sup  t_exp      sig
1 (Intercept)      8.601     1.200  6.217 10.984  7.169  < 0.001
2       tevol     -0.046     0.025 -0.095  0.003 -1.883  = 0.063
3        edad     -0.012     0.024 -0.059  0.035 -0.521  = 0.604
4         imc      0.038     0.052 -0.066  0.142  0.722  = 0.472
```

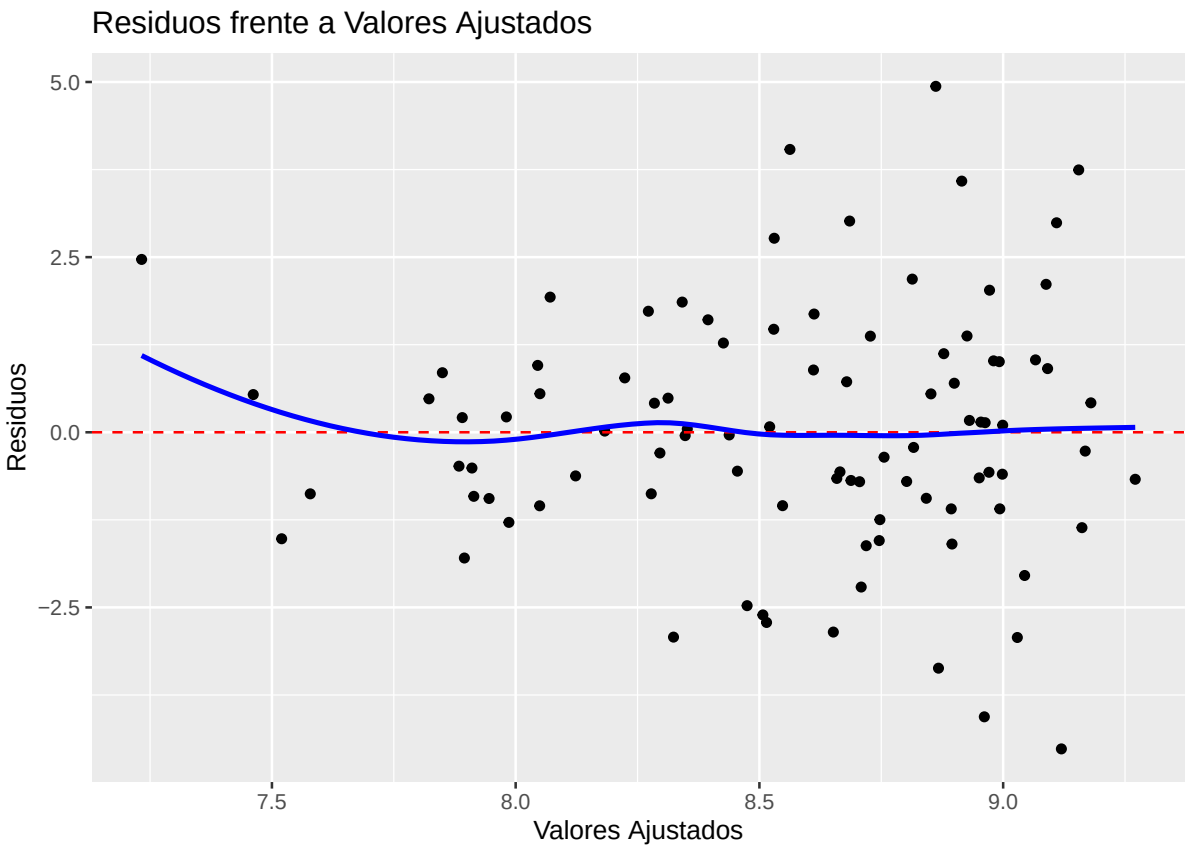
```
# Distribución residual ---
  Error estándar residual:  1.77
    Residuos Res_Est
min   -4.519  -2.589
Q1    -0.945  -0.545
Q2    -0.010  -0.006
Q3     0.995   0.572
max    4.938   2.923

  Test de normalidad residual (Shapiro-Wilk):
   w = 0.991 ,    = 0.744
```

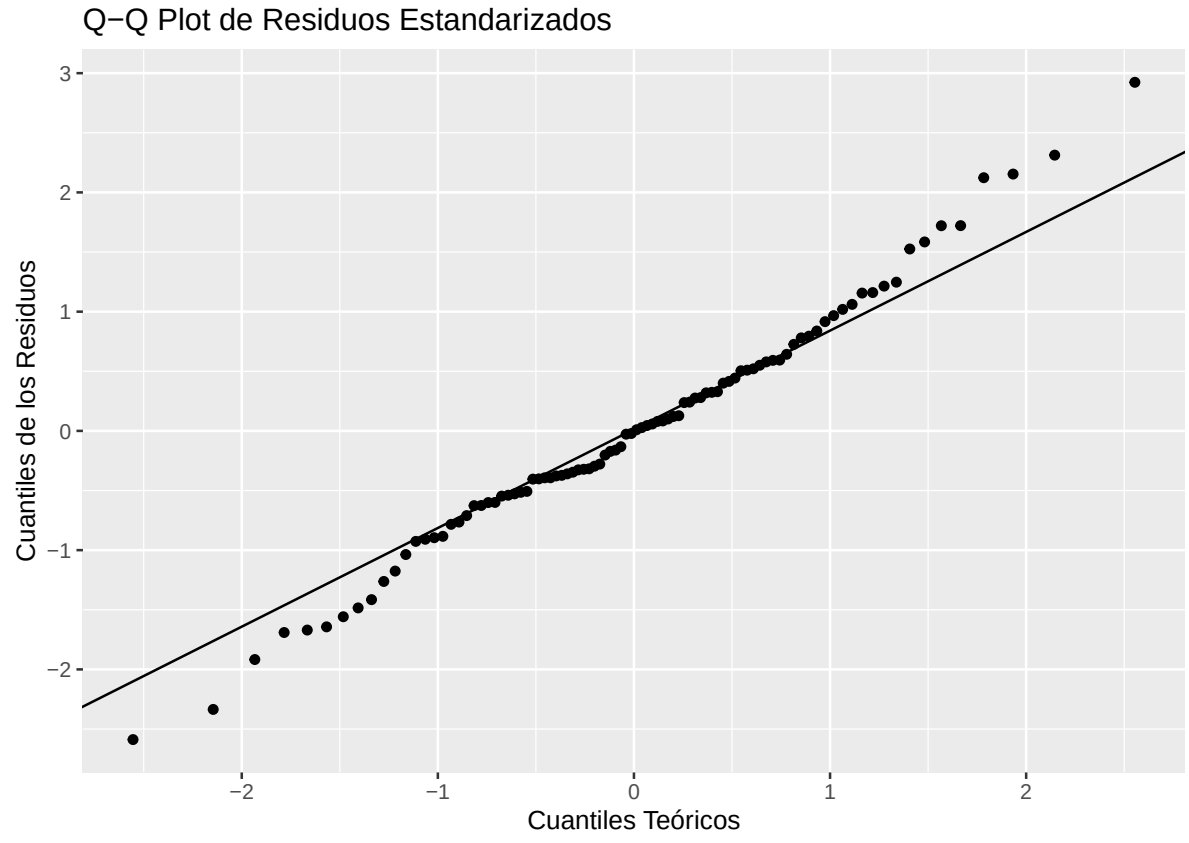


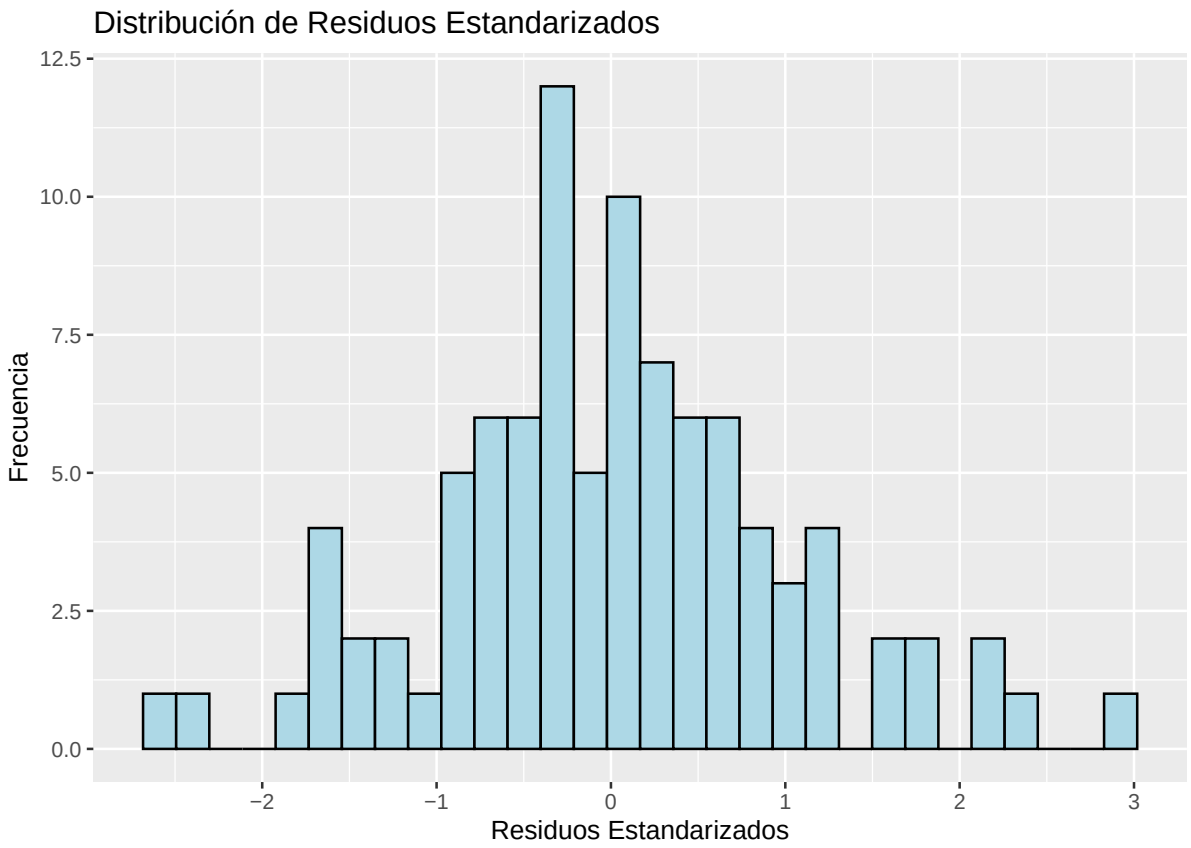







Residuos frente a Valores Ajustados
Residuos
5.0
2.5
0.0
−2.5
7.5
8.0
8.5
9.0
Valores Ajustados


Q−Q Plot de Residuos Estandarizados
Cuantiles de los Residuos
3
2
1
0
−1
−2
−2
−1
0
1
2
Cuantiles Teóricos



10.18. Resumen

Fórmulas Clave

Modelo y Estimación:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + U_i$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Varianza de Coeficientes:

$$\text{Var}(\hat{\beta}) = \sigma_U^2(\mathbf{X}^T\mathbf{X})^{-1}$$

$$s^2 = \frac{\text{RSS}}{n-p-1}$$

Bondad de Ajuste:

$$R^2 = 1 - \frac{\text{RSS}}{\text{SST}}$$

$$R^2_{\text{adj}} = 1 - (1-R^2)\frac{n-1}{n-p-1}$$

Tests e Intervalos:

$$t_j = \frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)} \sim t_{n-p-1}$$

$$F = \frac{(\text{RSS}_1 - \text{RSS}_2)/(p_2 - p_1)}{\text{RSS}_2/(n - p_2 - 1)} \sim F_{p_2-p_1, n-p_2-1}$$

Criterios de Información:

$$\text{AIC} = n \log(\text{RSS}/n) + 2(p+1)$$

$$\text{BIC} = n \log(\text{RSS}/n) + (p+1)\log(n)$$

Seleccionar el modelo con **menor** AIC o BIC.

10.19. Ejercicios

1. **Interpretación de Coeficientes Parciales:** En un modelo que predice la **hemoglobina glicosilada** (HbA1c, %) usando años de evolución de la diabetes, edad del paciente e IMC, el coeficiente de los años de evolución es 0.05. ¿Qué significa este número? ¿Cómo interpretaría el coeficiente del IMC?

2. **Matriz de Diseño:** Un estudio epidemiológico analiza la **presión arterial diastólica** con 3 variables predictoras (edad, IMC y tabaquismo en paquetes-año) y 80 observaciones. ¿Cuál es la dimensión de la matriz $\mathbf{X}$? ¿Qué contiene la primera columna? ¿Cuántos grados de libertad tienen los residuos?

3. **Problema de Multicolinealidad:** En un modelo de riesgo cardiovascular se incluyen simultáneamente el colesterol LDL (mg/dL) y el colesterol no-HDL (mg/dL). ¿Qué problema clínico y estadístico espera encontrar? ¿Cómo lo detectaría con el VIF y cómo lo resolvería?

4. R^2 **vs.** R^2_{adj}**:** Un investigador ajusta un modelo de predicción de supervivencia en cáncer de mama e incluye secuencialmente 15 variables (estadio, ganglios, Ki67, HER2, RE, RP, tamaño tumoral, edad, tratamiento, comorbilidades, etc.). ¿Por qué R^2 siempre aumenta al añadir cada variable? ¿Cuál criterio (AIC, BIC o R^2_{adj}) es más apropiado para seleccionar el modelo final?

5. **Prueba F para Modelos Anidados:** En un estudio de HbA1c se ajustan dos modelos:

   - Modelo A: $\text{HbA1c} = \beta_0 + \beta_1 \cdot \text{tevol} + U$
   - Modelo B: $\text{HbA1c} = \beta_0 + \beta_1 \cdot \text{tevol} + \beta_2 \cdot \text{Edad} + \beta_3 \cdot \text{IMC} + U$

   Se obtiene $\text{RSS}_A = 188.4$, $\text{RSS}_B = 172.6$, $n = 94$.

   a) ¿Qué hipótesis nula se está probando con la prueba F?
   b) Calcula el estadístico F
   c) Con $F_{0.05,2,90} \approx 3.10$, ¿es el modelo B significativamente mejor? Interpreta clínicamente

6. **Variables Indicadoras en Ensayos Clínicos:** En un ensayo clínico de 4 brazos se comparan tres tratamientos antidiabéticos (metformina, sitagliptina, empagliflozina) frente a placebo. ¿Cuántas variables indicadoras necesitas? Si el coeficiente estimado de empagliflozina es $-0.72\,\%$ de HbA1c con $p = 0.003$, interpreta este resultado en el contexto del ensayo. ¿Cuál es la categoría de referencia?

10.20. Respuestas a los Ejercicios

Ejercicio 1: El coeficiente de años de evolución (0.05) significa que, manteniendo la edad del paciente y el IMC constantes, cada año adicional de evolución de la diabetes se asocia con un aumento promedio de 0.05 puntos porcentuales en la HbA1c. El coeficiente del IMC representaría el cambio esperado en HbA1c por cada kg/m² adicional de IMC, ajustando por los años de evolución y la edad: si es positivo, mayor obesidad se asocia con peor control glucémico.

Ejercicio 2: Dimensión de $\mathbf{X}$: 80×4 (80 observaciones, 1 columna de unos para el intercepto + 3 columnas de predictores). La primera columna contiene todos los 1s (para estimar β_0). Los grados de libertad de los residuos son $n - p - 1 = 80 - 3 - 1 = 76$.

Ejercicio 3: El colesterol no-HDL = Colesterol Total − HDL, por lo que está determinísticamente relacionado con el LDL más el VLDL. Si ambos entran en el modelo, hay multicolinealidad severa (potencialmente casi perfecta). Consecuencias: errores estándar muy grandes, coeficientes inestables, posibles signos contraintuitivos. Detección: VIF tendería a infinito o sería muy alto (>10). Solución: utilizar solo uno de los dos indicadores lipídicos, preferiblemente el LDL por su mayor validación clínica, o usar el colesterol no-HDL como alternativa cuando el LDL no es medible.

Ejercicio 4: R^2 siempre aumenta porque es el ratio SCE/SCT, y añadir variables (aunque sean ruido puro) solo puede disminuir o mantener el RSS, nunca aumentarlo. El modelo siempre "aprovecha" cualquier correlación espuria con los datos de muestra. Para 15 variables y n no muy grande, el R^2 puede estar muy inflado. El BIC es el criterio más apropiado en investigación confirmatoria: penaliza más fuertemente la complejidad ($\log n$ vs. 2 en AIC), favoreciendo modelos parsimoniosos. El R^2_{adj} también es útil como referencia, pero no tiene la misma base probabilística.

Ejercicio 5: - a) $H_0 : \beta_2 = \beta_3 = 0$ (la edad y el IMC no aportan información sobre la HbA1c más allá del tiempo de evolución) - b) $F = \frac{(\text{RSS}_A - \text{RSS}_B)/(p_2 - p_1)}{\text{RSS}_B/(n - p_2 - 1)} = \frac{(188.4 - 172.6)/2}{172.6/90} = \frac{15.8/2}{1.918} = \frac{7.9}{1.918} = 4.12$ - c) $F = 4.12 > F_{0.05,2,90} = 3.10$, por lo que rechazamos H_0 ($p < 0.05$). El modelo B es significativamente mejor. Clínicamente: la edad y el IMC aportan información independiente sobre el control glucémico, más allá de los años de evolución de la diabetes. Esto sugiere que el tratamiento de la hiperglucemia debe considerar también la edad del paciente y su estado ponderal.

Ejercicio 6: Se necesitan **3 variables indicadoras** ($m - 1 = 4 - 1 = 3$): una para metformina, otra para sitagliptina y otra para empagliflozina. El placebo es la categoría de referencia. El coeficiente de empagliflozina ($-0.72\,\%$, $p = 0.003$) significa que, en promedio, los pacientes tratados con empagliflozina tienen una HbA1c 0.72 puntos porcentuales menor que los pacientes con placebo, y esta diferencia es estadísticamente significativa. En términos clínicos, una reducción de ~0.7 % en

HbA1c es clínicamente relevante (el umbral suele ser $0.5\,\%$).

Métodos Avanzados

Para ampliar los contenidos de este capítulo con técnicas estadísticas avanzadas, visita:
→ Bioestadística Avanzada — M.A. Luque Fernández

Parte V

Parte V: Análisis de Datos Categóricos

Capítulo 11

Semana 11 — Análisis de Datos Categóricos

En estadîstica médica y epidemiológica, muchas variables de interés son categoricas: enfermedad (sî/no), exposición a un factor de riesgo (presente/ausente), grupo de tratamiento, estadio clînico. Este capîtulo introduce las herramientas para analizar este tipo de datos: las pruebas Chi-cuadrado de independencia y homogeneidad, la prueba de McNemar para datos apareados, las medidas de asociación en tablas de contingencia según el diseño del estudio (cohorte, transversal, caso-control), y los métodos exactos cuando los supuestos paramétricos no se cumplen.

11.1. Variables Categóricas y Tablas de Contingencia

Una **tabla de contingencia** es una representación matricial de las frecuencias conjuntas de dos o más variables categóricas. Es la base del análisis de datos categóricos. La tabla 2x2 es un caso particular de tablas de contingencia, que pueden tener más de dos categorîas en filas y columnas. Por ejemplo, una tabla 3x4 podrîa analizar la asociación entre un factor de riesgo con tres niveles (bajo, medio, alto) y un resultado con cuatro categorîas (ausente, leve, moderado, severo). En estos casos, el análisis se realiza utilizando pruebas de Chi-cuadrado generalizadas para tablas r×c, que evalúan la independencia entre las variables categóricas sin necesidad de reducirlas a dicotómicas.

11.1.1. Tabla 2×2: Estructura y Notación Epidemiológica

La tabla 2×2 es el caso más frecuente en epidemiologîa y ensayos clînicos:

	Resultado: Sî	**Resultado: No**	**Total**
Exposición: Sî	a	b	$a+b$
Exposición: No	c	d	$c+d$
Total	$a+c$	$b+d$	n

	Resultado: Sí	Resultado: No	Total

Donde:

- a = expuestos con resultado positivo (casos expuestos)
- b = expuestos sin resultado
- c = no expuestos con resultado positivo
- d = no expuestos sin resultado
- $n = a + b + c + d$ = total de observaciones

Advertencia

Sin embargo, hay circunstancias donde la tabla es presentada transpuesta, es decir, con la enfermedad en las filas y la exposición en las columnas. En este caso, la interpretación de los elementos de la tabla se mantiene pero se debe tener cuidado al calcular las medidas de asociación (RR, OR) para no invertir los grupos. En general, lo más común es encontrar la enfermedad en las columnas y la exposición en las filas, pero siempre es importante verificar la disposición de los datos antes de realizar los cálculos.
Tabla 2x2 (transpuesta):

	Exposición: Sí	Exposición: No	Total
Resultado: Sí	a	c	$a + c$
Resultado: No	b	d	$b + d$
Total	$a + b$	$c + d$	n

Nota: En este formato, a representa los casos expuestos, c los casos no expuestos, b los controles expuestos y d los controles no expuestos y es el que se utiliza con el paquete **BioEstatR**

11.1.2. Tablas r×c Generalizadas

Para r categorías de fila y c categorías de columna, la tabla es de dimensión $r \times c$ con frecuencias observadas O_{ij}, marginales de fila $O_{i\cdot} = \sum_j O_{ij}$ y de columna $O_{\cdot j} = \sum_i O_{ij}$.

11.2. Prueba ² de Independencia

La prueba Chi-cuadrado de independencia contrasta si dos variables categóricas son estadísticamente independientes en la población.

11.2.1. Hipótesis

H_0 : las variables son independientes

$$\Leftrightarrow \quad P(X = i, Y = j) = P(X = i)\, P(Y = j)$$

H_1 : las variables no son independientes (están asociadas)

11.2.2. Frecuencias Esperadas Bajo H_0

Si H_0 es verdadera, las frecuencias esperadas son:

$$E_{ij} = \frac{O_{i\cdot} \cdot O_{\cdot j}}{n}$$

⚠ Estadístico: Pearson [2]

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Bajo H_0: $\chi^2 \sim \chi^2_{(r-1)(c-1)}$ (asintóticamente).
Para tabla 2×2: $(r-1)(c-1) = 1$ grado de libertad.
Condición de validez: Todas las frecuencias esperadas deben ser $E_{ij} \geq 5$. Si alguna es menor, usar la corrección de Yates o la **prueba exacta de Fisher**.

11.2.3. Cálculo Manual para Tabla 2×2

Para tabla 2×2 con notación (a, b, c, d):

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

11.2.4. Ejemplo Médico: Tabaquismo y Osteoporosis

💡 Ejemplo: tabla2x2() — datos osteo (Fac. Medicina UGR)

Analizamos la asociación entre tabaquismo (`tabaco`) y osteoporosis del cuello femoral (`osteo_cue`) en 94 pacientes diabéticos en un estudio ficticion de casos y controles usando el software **BioEstatR**. Notese que la exposición debe ser introducida por columnas y la enfermedad en las filas. Este es el consenso para este programa pero en general lo más frecuente es encontrar la enfermedad en las columnas y la exposicion en las filas de la tabla de 2x2 como se ha comentado anteriormente.

```
# Análisis de tablas 2x2
# ----------------------
```

```
# Frecuencias observadas
                     Fumador   No fumador Total
  Con osteoporosis        44            9    53
  Sin osteoporosis        26           15    41
  Total                   70           24    94


# Test Chi-cuadrado para un estudio retrospectivo

   ² = 4.651,   gl = 1,  p = 0.031, (cpc = 1)
  Validez: Frecuencia mínima esperada = 10.47  > 7.7

  Test exacto de Fisher (bilateral): p = 0.035

  --- Otros criterios  ²:
   ² = 4.673,   gl = 1,  p = 0.054, (sin cpc)
   ² = 3.699,   gl = 1,  p = 0.054, (cpc de Yates = 47.00)

# Medidas de asociación para un estudio retrospectivo

  Riesgo atribuible*:
  Ra=0.406; 95%-IC(Ra)= (0.082, 0.615)
  * La estimación de Ra para estudios retrospectivos es una aproximación válida si la prevale

  Razón del producto cruzado (odds ratio):
  OR=2.821; 95%-IC(OR)= (1.070, 7.013)
  * La estimación para OR sirve de aproximación al riesgo relativo siempre que la prevalencia
```

Comprobación manual de E_{ij}:

$$E_{11} = \frac{53 \times 70}{94} = 39.47, \quad E_{12} = \frac{53 \times 24}{94} = 13.53$$

$$E_{21} = \frac{41 \times 70}{94} = 30.53, \quad E_{22} = \frac{41 \times 24}{94} = 10.47$$

Minima frecuencia esperada: $E_{22} = 10.47 > 5$ (condición de validez cumplida).

i Interpretación

El estadîstico $^2 = 4.65$ con 1 grado de libertad produce p $= 0.031 < 0.05$, rechazando la hipótesis nula de independencia. Existe asociación estadîsticamente significativa entre tabaquismo y osteoporosis en esta muestra de 94 pacientes diabéticos. Los casos con osteoporosis de la cabez del cuello del femur presentan una odds ratio de 2.82 (IC 95 %: 1.07–7.01), indicando

aproximadamente 2.8 veces mayor probabilidad de exposición al hábito tabaquico comparado con los controles sin osteoporosis. En el contexto clínico, esta asociación es relevante para estratificar riesgo en pacientes diabéticos, sugiriendo que el cese del tabaquismo podría ser una intervención preventiva importante en esta población de alto riesgo metabólico.

11.3. Prueba ² de Homogeneidad

La prueba de homogeneidad contrasta si la distribución de una variable categórica es **la misma en k poblaciones o grupos** predefinidos. Matemáticamente usa el mismo estadístico χ^2, pero el diseño del estudio es diferente.

11.3.1. Independencia vs. Homogeneidad

Aspecto	Independencia	Homogeneidad
Muestreo	Muestra única	k muestras independientes
H_0	Independencia entre variables	Igual distribución en k grupos
Ejemplo	¿Tabaquismo y cáncer relacionados?	¿Mismo % de curación en 3 tratamientos?

11.3.2. Procedimiento

Se aplica el mismo estadístico $\chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ con $(r-1)(c-1)$ grados de libertad.

Ejemplo: tablarxc() — eficacia de tratamiento en 3 grupos

Ensayo clínico con 3 grupos de tratamiento (Control, Dosis baja, Dosis alta) y resultado binario (Curado/No curado):

```
# Test Chi-cuadrado para tablas RxC
# ---------------------------------

# Frecuencias observadas
               Curado   No_curado Total
  Control          20          30    50
  Dosis_baja       35          25    60
  Dosis_alta       45          15    60
  Total           100          70   170

# Test chi-cuadrado
  Validez: Frecuencia mínima esperada = 20.59
           0 frecuencias esperadas son menores a 1
           0 son menores a 5 (el 0% de la tabla)
```

```
²(2 gl) = 13.802, p = 0.001
```

i Interpretación

El estadístico $^2 = 13.802$ con 2 grados de libertad produce $p < 0.001$, rechazando la hipótesis nula de homogeneidad entre grupos. La distribución de curación no es homogénea entre los 3 grupos de tratamiento (Control: 40 %, Dosis baja: 58 %, Dosis alta: 75 %). Los resultados demuestran que la dosis alta de tratamiento se asocia significativamente con tasas de curación más altas comparado con control y dosis baja. Este patrón dosis-respuesta es clinicamente relevante y sugiere que la escalada de dosis podría optimizar eficacia, aunque el balance entre beneficio y tolerabilidad debe evaluarse conjuntamente en la toma de decisiones terapéuticas.

11.4. Condiciones de Aplicación y Alternativas

11.4.1. Condición de Frecuencias Esperadas

La regla clásica de Cochran (1954) exige, para que la aproximación $\chi^2 \sim \chi^2_{(r-1)(c-1)}$ sea válida, que **todas** las frecuencias esperadas E_{ij} sean ≥ 5 en una tabla 2×2 y que, en tablas $r \times c$, al menos el 80 % de las E_{ij} sean ≥ 5 y ninguna < 1.

Sin embargo, este umbral único es excesivamente conservador. La escuela granadina de bioestadística (Martín Andrés and Luna del Castillo [2004]; Martín Andrés and Silva Mato [1994]) ha demostrado, mediante estudios de simulación extensivos, que el valor mínimo de E_{ij} que mantiene un error de tipo I cercano al nominal **depende del tamaño muestral y del diseño del estudio**:

! Valores mínimos de E_{ij} recomendados — Tablas 2×2

Tabla 11.1: Umbrales para la validez del χ^2 en tablas 2×2 según diseño y tamaño muestral. Adaptado de Martín Andrés and Luna del Castillo [2004] y Martín Andrés and Silva Mato [1994].

Tamaño muestral n	Diseño con un margen fijo (cohortes, caso-control)	Diseño con ambos márgenes libres (transversal, prevalencia)	Prueba recomendada
$n \leq 20$	—	—	**Fisher exacta** (siempre)
$20 < n \leq 40$	$E_{\text{mín}} \geq 5$	$E_{\text{mín}} \geq 5$	χ^2 con corrección de Yates
$40 < n \leq 100$	$E_{\text{mín}} \geq 3$	$E_{\text{mín}} \geq 4$	χ^2 de Pearson (sin corrección)

$100 < n \leq 200$	$E_{\text{mín}} \geq 2$	$E_{\text{mín}} \geq 3$	χ^2 de Pearson
$n > 200$	$E_{\text{mín}} \geq 1$	$E_{\text{mín}} \geq 2$	χ^2 de Pearson
Ambos márgenes fijos (raro)	—	—	**Fisher exacta** (test condicional óptimo)

Cuando el diseño es de "margen fijo" vs. "márgenes libres"

- **Un margen fijo** (lo más habitual en epidemiología): el investigador fija a priori el tamaño de los grupos de comparación.
  - *Cohortes*: se fija el número de expuestos y no expuestos; se observa la incidencia.
  - *Caso-control*: se fija el número de casos y controles; se observa la exposición.
- **Ambos márgenes libres**: el muestreo es transversal/de prevalencia y solo se fija n total; los dos márgenes (exposición y enfermedad) son aleatorios.
- **Ambos márgenes fijos**: situación poco frecuente en investigación clínica (p.ej., experimento controlad como es el caso del té de Fisher). Es la única en la que la **prueba exacta de Fisher** es el test condicional óptimo desde un punto de vista frecuentista.

Para tablas $r \times c$ con r o $c > 2$, la regla práctica recomendada por estos autores es: $\geq 80\,\%$ de las E_{ij} por encima del umbral correspondiente al n total (columna apropiada), y ninguna $E_{ij} < 1$. Para detalles teóricos y simulaciones que sustentan estos umbrales, véase Martín Andrés and Luna del Castillo [2004].

11.4.2. Corrección de Continuidad de Yates (Tabla 2×2)

El estadístico χ^2 de Pearson es una suma de términos calculados sobre frecuencias enteras que se aproxima por la distribución continua $\chi^2_{(r-1)(c-1)}$. La **corrección de continuidad de Yates** reduce el estadístico para mejorar la aproximación:

Estadístico: χ^2 de Yates (con corrección de continuidad)

Para una tabla 2×2 con frecuencias a, b, c, d:

$$\chi^2_{\text{Yates}} = \frac{n\left(|ad - bc| - \frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

- Se aplica cuando alguna $E_{ij} \in [5, 10)$
- Siempre da $\chi^2_{\text{Yates}} \leq \chi^2_{\text{Pearson}} \rightarrow$ p-valor más conservador
- En R: `chisq.test(tabla, correct = TRUE)` (opción por defecto para tablas 2×2)

Ejemplo: Pearson vs. Yates en tabla con frecuencias moderadas

```
            Caso Control
Expuesto       7       4
No expuesto    3      11
    Pearson's Chi-squared test

data:  m
X-squared = 5, df = 1, p-value = 0.03
    Pearson's Chi-squared test with Yates' continuity correction

data:  m
X-squared = 3, df = 1, p-value = 0.08
```

Interpretación

La prueba de Pearson (sin corrección) rechaza H ($^2 = 5$, p = 0.03), mientras que la corrección de Yates no rechaza ($^2 = 3$, p = 0.08). La discordancia ocurre porque las frecuencias esperadas están en el rango [5,10), donde la aproximación normal es moderada. La corrección de Yates reduce el estadístico penalizando por el uso de una distribución continua con datos discretos, produciendo un resultado más conservador. En práctica clínica, con muestras pequeñas donde E [5,10), la versión conservadora (Yates) es preferible para evitar falsos positivos, aunque para grandes muestras ambas convergen.

La siguiente figura compara las distribuciones empíricas de los estadísticos con y sin corrección frente a la distribución $\chi^2(1)$ teórica, usando simulaciones bajo H_0 con $n = 25$:

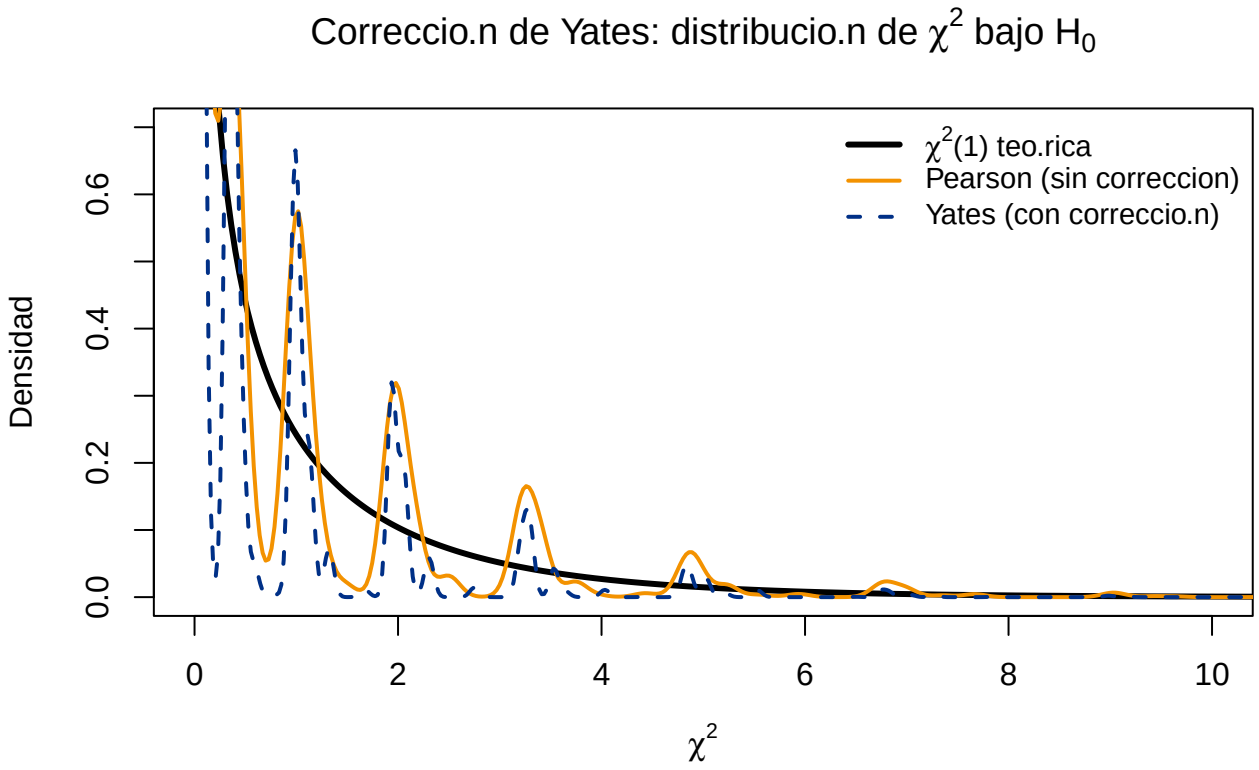


Figura 11.1: Comparación de la distribución empírica de ^{2} de Pearson (sin corrección, naranja) y ^{2} de Yates (con corrección, azul) con la curva $^2(1)$ teórica (negro), bajo H con n = 25. La corrección de Yates produce una aproximación más conservadora pero mejor calibrada para muestras pequeñas.

11.4.3. Prueba Exacta de Fisher

Cuando alguna $E_{ij} < 5$, la prueba de Fisher calcula la probabilidad exacta de observar la tabla obtenida (o más extrema) bajo H_0, utilizando la distribución hipergeométrica:

$$P = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}$$

Ejemplo: Prueba exacta de Fisher en muestra pequeña

```
# Tabla con frecuencias pequeñas (Eij < 5)
m <- matrix(c(4, 1, 2, 8), nrow = 2,
                   dimnames = list(c("Expuesto", "No expuesto"),
                                   c("Caso", "Control")));m
```

```
            Caso Control
Expuesto       4       2
No expuesto    1       8
```

```
fisher.test(m)
```

```
    Fisher's Exact Test for Count Data

data:  m
p-value = 0.09
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.747 875.880
sample estimates:
odds ratio
      12.5
```

Nota

Notese que en este caso también sería apropiado utilizar un test de permutación ya introducido en temas anteriores. Para el ejemplo anterior procederîamos como sigue:

```
# Tabla con frecuencias pequeñas
m <- matrix(c(4, 1, 2, 8), nrow = 2,
                    dimnames = list(c("Expuesto", "No expuesto"),
                                    c("Caso", "Control")));m
```

```
            Caso Control
Expuesto       4       2
No expuesto    1       8
```

```
# Permutation test nativo en R
set.seed(134)
test_simulado <- chisq.test(m, correct = FALSE, simulate.p.value = TRUE, B = 1000)
print(test_simulado)
```

```
    Pearson's Chi-squared test with simulated p-value (based on 1000
    replicates)

data:  m
X-squared = 5, df = NA, p-value = 0.08
```

Interpretación

Con n = 15 y varias frecuencias esperadas < 5, la aproximación ^{2} es inapropiada, haciendo que la prueba exacta de Fisher sea el método correcto. Fisher calcula mediante la distribución hipergeométrica la probabilidad exacta de observar la tabla observada (u otra más extrema) bajo H de independencia. Siendo un método no paramético, el resultado proporciona un

p-valor exacto y una odds ratio (OR) con intervalo de confianza exacto, basados en el espacio muestral discreto de tablas 2×2. Este enfoque es especialmente valioso en estudios de casos y controles de pequeño tamaño o cuando la exposición es rara, situaciones comunes en medicina clínica donde los datos categóricos a menudo tienen frecuencias limitadas.
El test de permutación también es una alternativa válida, aunque la prueba exacta de Fisher es el estándar de referencia para tablas 2×2 con frecuencias pequeñas debido a su eficiencia computacional y su fundamento teórico sólido. El test se basa en generar una distribución empírica del estadístico de prueba bajo la hipótesis nula mediante permutaciones aleatorias de los datos, proporcionando un p-valor no paramétrico que también es exacto en el sentido de que no depende de aproximaciones asintóticas. En este caso, ambos métodos (Fisher y permutación) deberían converger a resultados similares, aunque Fisher es mas directo para tablas 2×2.

11.5. Prueba de McNemar para Datos Apareados

11.5.1. ¿Cuándo usar McNemar?

La prueba χ^2 de independencia asume que las observaciones son **independientes**. Sin embargo, en muchos diseños médicos los datos son **apareados**: se mide el mismo paciente en dos condiciones o momentos distintos.

Situaciones típicas en medicina:

- **Antes/después**: ¿Mejora la proporción de pacientes con PA controlada tras 6 meses de intervención?
- **Dos pruebas diagnósticas**: ¿Concuerdan el test rápido de PCR y el cultivo bacteriano aplicados a los mismos pacientes?
- **Pares emparejados**: ¿Difiere la detección de lesiones en ojo derecho vs. ojo izquierdo del mismo paciente?
- **Diseños cruzados**: ¿Es diferente la proporción de respuestas con tratamiento A vs. B en el mismo individuo?

En estos diseños, la **unidad de análisis es el par**, no el individuo, y solo los pares **discordantes** aportan información sobre la diferencia.

11.5.2. Estructura de la Tabla Apareada

La tabla de datos apareados tiene una estructura especial:

	Condición B: Positivo	**Condición B: Negativo**	**Total**
Condicion A: Positivo	a (concordantes $+/+$)	b (discordantes $+/-$)	$a+b$
Condición A: Negativo	c (discordantes $-/+$)	d (concordantes $-/-$)	$c+d$
Total	$a+c$	$b+d$	n

Donde:

- a: pares concordantes positivos (ambos positivos)
- d: pares concordantes negativos (ambos negativos)
- b: discordantes — positivo en A, negativo en B
- c: discordantes — negativo en A, positivo en B

La hipótesis nula es que la proporción marginal de positivos es la misma en ambas condiciones, equivalente a:

$$H_0 : P(\mathrm{A}^+\mathrm{B}^-) = P(\mathrm{A}^-\mathrm{B}^+) \quad \Longleftrightarrow \quad \pi_b = \pi_c = 0.5$$

11.5.3. Estadístico de McNemar

⚠ Estadístico: Prueba de McNemar

Solo los pares **discordantes** (b y c) aportan información sobre la diferencia entre condiciones.
Sin corrección de continuidad (preferida cuando $b + c \geq 25$):

$$\chi^2_{\text{McNemar}} = \frac{(b-c)^2}{b+c} \sim \chi^2_1 \quad \text{bajo } H_0$$

Con corrección de continuidad de Yates (recomendada cuando $b + c < 25$):

$$\chi^2_{\text{McNemar,Yates}} = \frac{(|b-c|-1)^2}{b+c}$$

Para muestras muy pequeñas ($b + c < 10$): usar el test binomial exacto sobre los pares discordantes bajo $H_0 : p = 0.5$.
Regla de decisión: Rechazar H_0 si $\chi^2 > \chi^2_{0.05,1} = 3.84$ (o p-valor < 0.05).

11.5.4. Ejemplo: Intervención para el Control de la Hipertensión

Ejemplo: Estudio antes-después control de la presión arterial

En una consulta de medicina interna del Hospital Universitario de Granada, 100 pacientes hipertensos participan en un programa de intervención multidisciplinar (dieta, ejercicio y educación sanitaria) durante 6 meses. Se evalúa si el paciente tiene la PA controlada (< 140/90 mmHg) antes y después del programa.
Tabla de pares apareados:

	Post: PA Controlada	**Post: PA No controlada**	**Total**
Pre: PA Controlada	42 (a)	8 (b)	50
Pre: PA No controlada	28 (c)	22 (d)	50
Total	70	30	100

Interpretación de la tabla: - 42 pacientes ya tenían la PA controlada y mantuvieron el control tras la intervención - 22 pacientes no la controlaron ni antes ni después - $b = 8$: 8 pacientes que tenían la PA controlada la "perdieron" tras el programa - $c = 28$: 28 pacientes que no la tenían controlada **la consiguieron** con la intervención
Pares discordantes: $b + c = 8 + 28 = 36 \geq 25 \rightarrow$ no es necesaria la corrección de Yates.
Cálculo manual:

$$\chi^2_{\text{McNemar}} = \frac{(b-c)^2}{b+c} = \frac{(8-28)^2}{36} = \frac{400}{36} = 11.11$$

Como $11.11 > 3.84$ ($p \approx 0.001$), rechazamos H_0. La intervención mejora significativamente el control de la PA: 28 pacientes mejoraron frente a solo 8 que empeoraron.
Nota sobre las proporciones marginales: - Antes: $50/100 = 50\,\%$ con PA controlada
- Después: $70/100 = 70\,\%$ con PA controlada
- Diferencia: +20 puntos porcentuales ($p < 0.001$ por McNemar)
Si se hubiera aplicado incorrectamente un χ^2 de independencia (ignorando el apareamiento), la prueba sería errónea porque las observaciones no son independientes.

11.5.5. Implementación: BioEstatR vs. Base R

BioEstatR — testmcnemar(): Datos apareados

Para datos apareados (donde las observaciones son dependientes), utilizamos la función `testmcnemar()` del paquete `BioEstatR`.

```
library(BioEstatR)

# BioEstatR Tabla 2×2 apareada: testmcnemar()
testmcnemar(n11 = 42, n12 = 8, n21 = 28, n22 = 22,
            fcat = c("Pre: Controlada", "Pre: No controlada"),
            ccat = c("Post: Controlada", "Post: No controlada"))
```

```
# Inferencia con dos proporciones (muestras apareadas)
# ----------------------------------------------------

# Frecuencias observadas pretest x posttest
                         Post: Controlada    Post: No controlada Total
   Pre: Controlada                     42                      8    50
   Pre: No controlada                  28                     22    50
   Total                               70                     30   100


# Proporciones observadas pretest x posttest
                         Post: Controlada    Post: No controlada Total
   Pre: Controlada                   0.42                   0.08  0.50
   Pre: No controlada                0.28                   0.22  0.50
   Total                             0.70                   0.30  1.00

# Test de McNemar: H :  =
   Validez: n +n  = 36 > 10 el test es válido
   Zexp =  3.2500
               valor.p Alternativa
    Bilateral   0.0012  H :
    Unilateral  0.0006  H :  <


# Test exacto de Fisher:
  H :  =0.5 para n  ~ B(n +n ,   )
               Valor.p Alternativa
    Bilateral   0.0141  H :  0.5
```

```
    Unilateral  0.0006  H :  <0.5

  ----
  * Aquí se alude a la probabilidad total de la discordancia, es decir que   +  =1

# Estimación de las proporciones individuales de discordancias    y    (método de Wald ajusta
   [1] p  = 0.0962, 95%-IC(  ) = (0.0395, 0.1528)
   [2] p  = 0.2885, 95%-IC(  ) = (0.2014, 0.3755)



# Intervalo de confianza para la diferencia de 2 proporciones apareadas
   [1] Método de Wald (clásico con cpc):
      Estimación puntual de   -   = -0.2000
      Validez: n +n  = 36 > 5, el IC es válido
      95%-IC(  -  ) = (-0.3117, -0.0883)

   [2] Método de Agresti-Min:
      Estimación puntual de   -   = -0.1961
      Validez: siempre es válido
      95%-IC(  -  ) = (-0.3066, -0.0856)
```

```r
# Base R: mcnemar.test() - control de PA antes y después del programa
m_paired <- matrix(c(42, 28, 8, 22), nrow = 2,
                   dimnames = list(
                     c("Pre: Controlada", "Pre: No controlada"),
                     c("Post: Controlada", "Post: No controlada")))

cat("=== mcnemar.test() - Base R ===\n")
```

```
=== mcnemar.test() - Base R ===
```

```r
mcnemar.test(m_paired, correct = FALSE)  # Sin corrección (b+c=36 >= 25)
```

```
    McNemar's Chi-squared test

data:  m_paired
McNemar's chi-squared = 11, df = 1, p-value = 0.0009
```

La prueba de McNemar se centra exclusivamente en los pares discordantes: aquellos pacientes cuyo estado clínico cambió entre las dos mediciones. En nuestro ejemplo, el resultado de `testmcnemar()` nos ofrece tres claves para la interpretación:

1. **Identificación de Cambios (Pares Discordantes):**
   - $b = 8$: Pacientes que estaban controlados "antes" y pasaron a estar "no controlados" "después" (empeoraron).

- $c = 28$: Pacientes que no estaban controlados "antes" y lograron el control "después" (mejoraron).

2. **Lógica Estadística:** La prueba ignora los pares concordantes (los que siempre estuvieron controlados o nunca lo estuvieron) y evalúa si el número de pacientes que mejoró es significativamente diferente del número que empeoro. Bajo la hipótesis nula (H_0), esperaríamos que el número de cambios en ambas direcciones fuera igual ($b = c$).

3. **Conclusión Clínica:** Dado que nuestro p-valor es `< 0.05` y observamos que 28 pacientes mejoraron frente a solo 8 que empeoraron ($c > b$), rechazamos H_0. Concluimos que la intervención tiene un efecto beneficioso estadísticamente significativo sobre el control de la presión arterial.

En resumen, la prueba de McNemar nos permite afirmar que el cambio observado en las proporciones marginales de control de PA (del 50 % al 70 %) no es debido al azar, sino que refleja un impacto real del programa de intervención sobre los pacientes.

11.6. Diseños de Estudio y Medidas de Asociación

La eleccion de la medida de asociación depende del **diseño del estudio**. Todas se calculan a partir de la tabla 2×2 pero tienen interpretaciones distintas.

11.6.1. Estructura Epidemiológica de la Tabla 2×2

	Enfermedad (D+)	**Sin enfermedad (D−)**	**Total**
Exposición (E+)	a	b	$n_1 = a + b$
Sin exposición (E−)	c	d	$n_2 = c + d$
Total	$M_1 = a + c$	$M_0 = b + d$	n

Caso transpuesto:

	Exposición: Sí	**Exposición: No**	**Total**
Enfermedad: Sí	a	c	$a + c$
Sin enfermedad: No	b	d	$b + d$
Total	$a + b$	$c + d$	n

Riesgos (proporciones) en tabla caso típico:

$$p_1 = \frac{a}{a+b} \quad \text{(riesgo en expuestos)}, \qquad p_2 = \frac{c}{c+d} \quad \text{(riesgo en no expuestos)}$$

Riesgos (proporciones) en tabla transpuesta:

$$p_1 = \frac{a}{a+c} \quad \text{(riesgo en casos)}, \qquad p_2 = \frac{b}{b+d} \quad \text{(riesgo en controles)}$$

Tipos de estudios

Diseño	Unidad de análisis	Medida de asociación
Estudio de Cohortes	Individuo	Riesgo Relativo (RR)
Estudio de Casos y Controles	Individuo	Odds Ratio (OR)
Estudio Transversal	Individuo	Razón de Prevalencias (RP)
Estudio Apareado (McNemar)	Par de individuos	Razón de discordancias (b/c)

El estudio de cohortes es el único diseño que permite estimar riesgos de incidencia directamente, por lo que la medida de asociación más adecuada es el **Riesgo Relativo (RR)**. En estudios de casos y controles, donde se fija el número de casos y controles, no se pueden calcular riesgos, por lo que se utiliza la **Odds Ratio (OR)** como medida de asociación. En estudios transversales, donde se mide prevalencias, la medida adecuada es la **Razón de Prevalencias (RP)**.

11.6.2. Medidas de Asociación

Las medidas de asociación se calculan a partir de las frecuencias de la tabla 2×2: La diferencia de riesgos (DR) se calcula como la resta de las proporciones de riesgo entre expuestos y no expuestos:

$$DR = p_1 - p_2 = \frac{a}{a+b} - \frac{c}{c+d}$$

El riesgo relativo (RR) se calcula como el cociente de las proporciones de riesgo entre expuestos y no expuestos:

$$RR = \frac{p_1}{p_2} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Y la razón del producto cruzado u odds ratio (OR) se calcula como el cociente de los productos cruzados de la tabla:

$$OR = \frac{a \cdot d}{b \cdot c}$$

El intervalo de confianza de las medidas de asociación se calcula sobre la escala logarítmica para garantizar que los límites sean positivos, y luego se transforma exponencialmente para obtener el

intervalo en la escala original del RR. Su base teórica se fundamenta en la distribución asintótica normal del logaritmo del RR, lo que permite construir un intervalo de confianza que refleja la incertidumbre de la estimación. La fórmula del error estándar del log(RR) se deriva de la varianza de las proporciones en cada grupo, considerando la independencia de las muestras y el tamaño muestral. La distribución asintótica permite aplicar el teorema del límite central (TLC) y por ende el **método Delta**.

11.6.2.1. El Método Delta para Intervalos de Confianza (95 %) de las medidas de asociación en Epidemiología

1. El Problema: La asimetría de las medidas epidemiológicas

En epidemiología, usamos medidas de asociación como el **Odds Ratio (OR)** o el **Riesgo Relativo (RR)**. Estas medidas son cocientes. Su interpretación básica es:

- Si una exposición protege, el resultado está entre **0** y **1**.
- Si no hay efecto, el resultado es exactamente **1**.
- Si la exposición es de riesgo, el resultado va desde **1** hasta el ∞

Esta distribución es terriblemente asimétrica. Si intentáramos calcular un Intervalo de Confianza del **95 %** sumando y restando un margen de error directamente al OR o al RR, podríamos obtener límites inferiores matemáticamente imposibles (como un OR o RR negativo).

La solución matemáica radica en aplicar el logaritmo natural (ln) a la medida. El logaritmo "estira" la parte entre **0** y **1**, y "comprime" la parte del **1** al infinito, creando una distribución simétrica (una curva normal) perfecta para calcular intervalos. Sin embargo la transformación logarítmica requiere de el **Método Delta**.

2. ¿Qué es el Metodo Delta? (Fundamentos Teóricos)

El Método Delta es una técnica estadística que nos permite estimar la varianza de una función de una variable, cuando solo conocemos la varianza de la variable original. Se basa en una aproximación matemática llamada **Serie de Taylor** de primer orden. En términos sencillos: si tienes una curva complicada (como un logaritmo), el Método Delta "hace un zoom" tan de cerca en un punto específico que la curva se aproxima como si fuera una línea recta. Matemáticamente, si tienes un estadístico T con una varianza conocida $Var(T)$, y le aplicas una función $f(T)$, la varianza de esa nueva función se aproxima usando la primera derivada de la función multiplicada al cuadrado por la varianza original:

$$Var(f(T)) \approx [f'(T)]^2 \cdot Var(T)$$

3. La Aplicación en Epidemiología

Gracias a esa fórmula teórica, se aplica el **Método Delta** a las fórmulas del OR y el RR de la clásica tabla de 2x2 (con las celdas a, b, c, d), logrando unas fórmulas de varianza increíbles por su

simplicidad (se omite su derivación pero el siguiente enlace proporciona información muy detalla con respecto a su derivación y cálculo: The Delta-Method and Influence Function in Medical Statistics: a Reproducible Tutorial

Para el Odds Ratio (OR)

La varianza del logaritmo natural del OR se calcula simplemente sumando las inversas de las cuatro celdas de tu tabla:

$$Var(\ln(OR)) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

Para el Riesgo Relativo (RR)

La varianza del logaritmo natural del RR (donde los totales de las filas son $a+b$ y $c+d$) es:

$$Var(\ln(RR)) \approx \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}$$

Aquí tienes las secciones ampliadas con sus respectivas fórmulas en LaTeX y la explicación para que mantengan el mismo estilo didáctico del documento y puedas copiarlas y pegarlas directamente:

Para la diferencia de riesgos (DR)

La varianza de la diferencia de riesgos ($RD = p_1 - p_2$) tiene una ventaja: **no requiere transformar la medida a logaritmos**, porque la simple resta de dos proporciones ya genera una distribución bastante simétrica. Se calcula directamente sumando las varianzas de cada proporción individual.

Si definimos n_1 como el total de expuestos ($a+b$) y n_2 como el total de no expuestos ($c+d$), la fórmula clásica es:

$$Var(RD) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

> **Nota**
>
> *Nota: Como no usamos logaritmos aquí, el Error Estándar ($SE = \sqrt{Var(RD)}$) se suma y se resta directamente a la medida original para sacar el IC 95 %: $RD \pm 1.96 \cdot SE$.*

Para la razón de prevalencias (RP)

La varianza de la razón de prevalencias ($RP = \frac{p_1}{p_2}$) se calcula usando la fórmula del Método Delta aplicada al logaritmo natural de la función de cociente de proporciones. En los estudios transversales, la RP se calcula matemáticamente de la **misma manera** que el Riesgo Relativo (RR) en los estudios

de cohortes. Por lo tanto, el Método Delta nos lleva exactamente a la misma varianza elegante que vimos antes:

$$Var(\ln(RP)) \approx \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}$$

> **Nota**
>
> *Nota: Al igual que con el OR y el RR, para obtener el IC 95 % final se debe calcular el intervalo en la escala logarítmica y luego "deshacerlo" aplicando la función exponencial.*

4. Los 4 pasos para calcular el IC 95 % (Demostracion en R)

Cuando en un software calculas el intervalo de confianza, lo que ocurre en segundo plano gracias al Método Delta es lo siguiente:

1. **Transformar:** Se calcula la medida original y se aplica el logaritmo natural $\ln(OR)$.
2. **Calcular el Error Estandar (SE):** Se usa la fórmula del Método Delta. El SE es la raíz cuadrada de la varianza.
3. **Construir el intervalo logarítmico:** Aplicamos la regla de la curva normal para el **95 %** usando el valor **1.96**.
4. **Deshacer la transformación:** Se aplica la función exponencial (anti-logaritmo) a los límites calculados para volver a la escala original.

A continuación, ejecutamos esto en R paso a paso con una tabla ficticia:

```
# 0. Definir las celdas de la tabla 2x2
a <- 4   # Expuestos - Caso
b <- 1   # Expuestos - Control
c <- 2   # No expuestos - Caso
d <- 8   # No expuestos - Control

# Calcular el OR
OR <- (a / b) / (c / d)

# PASO 1: Transformación logarítmica
ln_OR <- log(OR)

# PASO 2: Varianza (Método Delta) y Error Estándar (SE)
var_ln_OR <- (1/a) + (1/b) + (1/c) + (1/d)
SE_ln_OR <- sqrt(var_ln_OR)

# PASO 3: Intervalos en escala logarítmica (Z = 1.96 para 95%)
lim_inf_ln <- ln_OR - (1.96 * SE_ln_OR)
```

```r
lim_sup_ln <- ln_OR + (1.96 * SE_ln_OR)

# PASO 4: Deshacer transformación (exponencial) para obtener los límites reales
IC_inferior <- exp(lim_inf_ln)
IC_superior <- exp(lim_sup_ln)

# Resultados
cat("Odds Ratio:", OR, "\n")
```

```
Odds Ratio: 16
```

```r
cat("IC 95% Inferior:", round(IC_inferior, 2), "\n")
```

```
IC 95% Inferior: 1.09
```

```r
cat("IC 95% Superior:", round(IC_superior, 2), "\n")
```

```
IC 95% Superior: 234
```

Ejemplo: Estudio de Cohorte — tabaquismo y cardiopatia isquémica

Seguimiento de 500 fumadores y 500 no fumadores durante 10 años:

	Cardiopatía	Sin cardiopatía	Total
Fumadores	80	420	500
No fumadores	30	470	500

```r
# Cálculo de RR e IC
a <- 80; b <- 420; c_ <- 30; d <- 470
p1 <- a / (a + b); p2 <- c_ / (c_ + d)
RR <- p1 / p2
se_logRR <- sqrt(b / (a*(a+b)) + d / (c_*(c_+d)))
IC_RR <- exp(log(RR) + c(-1, 1) * 1.96 * se_logRR)
cat("p1 (riesgo fumadores)     =", round(p1, 4), "\n")
```

```
p1 (riesgo fumadores)     = 0.16
```

```r
cat("p2 (riesgo no fumadores)  =", round(p2, 4), "\n")
```

```
p2 (riesgo no fumadores)  = 0.06
```

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

```
RR = 2.67
```

```r
cat("IC 95%: [", round(IC_RR[1], 3), ",", round(IC_RR[2], 3), "]\n")
```

```
IC 95%: [ 1.79 , 3.98 ]
```

```r
# Cálculo con BioEstatR
tabla2x2(o11 = 80, o12 =30 , o21 = 420, o22 = 470, # Nota: transposicion de columnas por fila
         ccat = c("Fumadores", "No fumadores"),
         fcat = c("Cardiopatía", "Sin cardiopatía"),
         estudio = "P")  #P: Prospectivo (Cohorte))
```

```
# Análisis de tablas 2x2
# ----------------------

# Frecuencias observadas
                    Fumadores   No fumadores Total
  Cardiopatía              80             30   110
  Sin cardiopatía         420            470   890
  Total                   500            500  1000


# Test Chi-cuadrado para un estudio prospectivo

   ² = 25.532,   gl = 1,   p < 0.001, (cpc = 2)
  Validez: Frecuencia mínima esperada = 55.00  > 14.9

  Test exacto de Fisher (bilateral): p < 0.001

  --- Otros criterios  ²:
```

```
   ² = 25.536,   gl = 1,  p < 0.001, (sin cpc)
   ² = 24.525,   gl = 1,  p < 0.001, (cpc de Yates = 500.00)

# Medidas de asociación para un estudio prospectivo
  [!] Las medidas de riesgo se calculan como riesgo de la categoría
      en la 1a columna (frente a la 2a) para la categoría en la 1a
      fila (frente a la 2a)

  Riesgo absoluto (diferencia de Berkson; método de Agresti-Caffo):
  d=0.100; 95%-IC(d)=(0.061, 0.138)

  Riesgo relativo:
  Rr=2.667; 95%-IC(Rr)=(1.773, 3.929)

  Razón del producto cruzado (odds ratio):
  OR=2.984; 95%-IC(OR)= (1.908, 4.572)
```

Interpretación: Los fumadores tienen **2.67 veces más riesgo** de cardiopatía isquémica que los no fumadores ($RR = 2.67$; $IC_{95\,\%}$: 1.79–3.98). El intervalo no contiene el 1, confirmando asociación estadísticamente significativa.

11.6.3. Diferencia de Riesgos (RD) y NNT

Definición: Diferencia de Riesgos (RD)

$$\widehat{RD} = p_1 - p_2 = \frac{a}{a+b} - \frac{c}{c+d}$$

Intervalo de confianza al 95 %:

$$IC_{95\,\%}(RD) = \widehat{RD} \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Número Necesario a Tratar/Perjudicar (NNT/NNH):

$$NNT = \frac{1}{|\widehat{RD}|}$$

El NNT indica cuántas personas hay que tratar (o exponer) para observar un caso adicional de resultado.

Para el ejemplo anterior: $RD = 0.16 - 0.06 = 0.10 \rightarrow NNH = 1/0.10 = 10$ (por cada 10 fumadores seguidos durante 10 años, 1 caso adicional de cardiopatía).

11.6.4. Estudio Transversal: Razón de Prevalencias (RP)

En un estudio transversal, las proporciones son **prevalencias** (no riesgos de incidencia). La medida de elección es la **Razón de Prevalencias (RP)**, cuya fórmula es idéntica al RR pero con prevalencias:

$$\widehat{RP} = \frac{\text{prevalencia en expuestos}}{\text{prevalencia en no expuestos}} = \frac{a/(a+b)}{c/(c+d)}$$

El intervalo de confianza se calcula con la misma formula que el RR. **Nota importante:** el OR no es una buena aproximación de la RP cuando la prevalencia es $> 10\,\%$.

11.6.5. Estudio Caso-Control: Odds Ratio (OR)

En un estudio caso-control el muestreo es por resultado (se seleccionan casos y controles), por lo que **no se pueden estimar riesgos ni RR directamente**. La medida correcta es el **Odds Ratio**.

Definición: Odds Ratio (OR)

$$\widehat{OR} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

Intervalo de confianza al 95 % (método de Woolf/Wald):

$$IC_{95\,\%}(OR) = \exp\left[\ln \widehat{OR} \pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right]$$

Propiedades: - $OR = 1$: sin asociación - $OR > 1$: exposición asociada con mayor odds de resultado - $OR \approx RR$ cuando la enfermedad es **rara** (prevalencia $< 10\,\%$) - `tabla2x2()` de BioEstatR calcula OR e IC directamente

Ejemplo: Caso-Control — alcohol y cirrosis hepática

```
library(BioEstatR)

# 200 casos de cirrosis y 200 controles
# Consumo excesivo de alcohol: 120 casos, 60 controles
tabla2x2(o11 = 120, o12 = 80, o21 = 60, o22 = 140,
         fcat = c("Alcohol excesivo", "Alcohol no excesivo"),
         ccat = c("Caso (cirrosis)", "Control"),
         estudio = "R")  #R: Retrospectivo: Caso-Control") # Nota: T: transversal, P: prospec

# Análisis de tablas 2x2
# ----------------------

# Frecuencias observadas
```

```
                       Caso (cirrosis)   Control Total
  Alcohol excesivo                 120        80   200
  Alcohol no excesivo               60       140   200
  Total                            180       220   400


# Test Chi-cuadrado para un estudio retrospectivo

   ² = 36.352,   gl = 1,  p < 0.001, (cpc = 2)
  Validez: Frecuencia mínima esperada = 90.00  > 7.7

  Test exacto de Fisher (bilateral): p < 0.001

  --- Otros criterios ²:
   ² = 36.364,   gl = 1,  p < 0.001, (sin cpc)
   ² = 35.162,   gl = 1,  p < 0.001, (cpc de Yates = 200.00)

# Medidas de asociación para un estudio retrospectivo

  Riesgo atribuible*:
  Ra=0.476; 95%-IC(Ra)= (0.341, 0.584)
  * La estimación de Ra para estudios retrospectivos es una aproximación válida si la prevale

  Razón del producto cruzado (odds ratio):
  OR=3.500; 95%-IC(OR)= (2.300, 5.253)
  * La estimación para OR sirve de aproximación al riesgo relativo siempre que la prevalencia
```

Interpretación: $OR = 3.50$ ($IC_{95\,\%}$: 2.30–5.25). Las personas con consumo excesivo de alcohol tienen **3.5 veces más odds** de cirrosis que las de consumo no excesivo. El resultado es altamente significativo ($p < 0.001$).

11.6.6. Resumen de Intervalos de Confianza

Medida	Fórmula del IC 95 %
RR	$\exp\left[\ln \widehat{RR} \pm 1.96\sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}\right]$
OR	$\exp\left[\ln \widehat{OR} \pm 1.96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right]$
RD	$\widehat{RD} \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Cuándo OR RR

El OR sobreestima al RR cuando la prevalencia/incidencia del resultado es alta ($> 10\,\%$). En enfermedades raras, $OR \approx RR$. En estudios caso-control, siempre calcular OR (el RR no es estimable sin incidencia).

11.7. Cálculo en R base de las medidas de asociación y sus intervalos de confianza

```
# Funcion para RR, RD, OR con IC (diseños no apareados)
calcular_medidas <- function(a, b, c, d, conf = 0.95) {
  z  <- qnorm(1 - (1 - conf)/2)
  n1 <- a + b; n2 <- c + d
  p1 <- a/n1;  p2 <- c/n2

  RR    <- p1/p2
  IC_RR <- exp(log(RR) + c(-1,1) * z * sqrt(b/(a*n1) + d/(c*n2)))

  OR    <- (a*d)/(b*c)
  IC_OR <- exp(log(OR) + c(-1,1) * z * sqrt(1/a + 1/b + 1/c + 1/d))

  RD    <- p1 - p2
  IC_RD <- RD + c(-1,1) * z * sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)

  cat(sprintf("RR  =%.3f  IC%d%%: [%.3f,%.3f]\n",
              RR, round(conf*100), IC_RR[1], IC_RR[2]))
  cat(sprintf("OR  =%.3f  IC%d%%: [%.3f,%.3f]\n",
              OR, round(conf*100), IC_OR[1], IC_OR[2]))
  cat(sprintf("RD  =%.3f  IC%d%%: [%.3f,%.3f]\n",
              RD, round(conf*100), IC_RD[1], IC_RD[2]))
  cat(sprintf("NNT/NNH =%.1f\n", 1/abs(RD)))
}

# Cohorte: tabaquismo y cardiopatía isquémica
cat("=== Cohorte: tabaquismo y cardiopatía ===\n")
```

```
=== Cohorte: tabaquismo y cardiopatía ===
```

```
calcular_medidas(a=80, b=420, c=30, d=470)
```

```
RR  = 2.667  IC95%: [1.786, 3.982]
OR  = 2.984  IC95%: [1.922, 4.632]
```

```
RD  = 0.100  IC95%: [0.062, 0.138]
NNT/NNH = 10.0
```

11.8. Ajuste por Confusión: El Estimador de Mantel-Haenszel

En estudios epidemiológicos, a menudo observamos una asociación entre una exposición y una enfermedad que parece variar según los niveles de una tercera variable, llamada **variable de confusión** (o *confounder*). Si no controlamos esta variable, podemos llegar a conclusiones erróneas —incluso observar una **Paradoja de Simpson**, donde una asociación aparente en el total desaparece o se invierte al estratificar por la variable de confusión.

11.8.1. El Estimador de Mantel-Haenszel (MH)

El estimador de Mantel-Haenszel es una técnica clásica para combinar los odds ratios (OR) obtenidos de varios estratos (k) de una variable de confusion, proporcionando un OR ajustado global. Se calcula como:

$$R_{MH} = \frac{\sum_{i=1}^{k} \frac{a_i d_i}{n_i}}{\sum_{i=1}^{k} \frac{b_i c_i}{n_i}}$$

Donde para el estrato i:

- a_i, b_i, c_i, d_i: celdas de la tabla 2×2.
- n_i: total de individuos en el estrato.

Criterio del 10 %: Para evaluar si una variable es un confounder clínicamente relevante, comparamos el OR crudo (total) con el OR ajustado (MH). Si la diferencia entre ambos es $\geq 10\,\%$, se concluye que existe una confusión sustancial, y el OR ajustado por Mantel-Haenszel debe utilizarse como la medida de asociacion más precisa.

11.8.2. Ejemplo Numérico en R: Evaluación de Confounding

Supongamos un estudio sobre el efecto de un nuevo fármaco (Fármaco A) en la recuperación de una enfermedad respiratoria, estratificado por la gravedad del paciente (Leve vs. Grave).

```r
library(epitools)

# Estrato 1: Leve
# Fármaco A (expuesto), recuperación (caso)
a1 <- 40; b1 <- 10; c1 <- 10; d1 <- 40
table1 <- matrix(c(a1, b1, c1, d1), nrow = 2, byrow = TRUE)
cat(sprintf("Tabla estrato 1 (Leve):\n")); print(table1)
```

```
Tabla estrato 1 (Leve):
```

```
     [,1] [,2]
[1,]   40   10
[2,]   10   40
```

```
or1 <- (a1 * d1) / (b1 * c1)  # OR estrato
print(sprintf("OR estrato 1 (Leve) =%.3f", or1))
```

```
[1] "OR estrato 1 (Leve) = 16.000"
```

```
# Estrato 2: Grave
a2 <- 10; b2 <- 40; c2 <- 5; d2 <- 45
table2 <- matrix(c(a2, b2, c2, d2), nrow = 2, byrow = TRUE)
or2 <- (a2 * d2) / (b2 * c2)  # OR estrato
cat(sprintf("\nTabla estrato 2 (Grave):\n")); print(table2)
```

```
Tabla estrato 2 (Grave):
```

```
     [,1] [,2]
[1,]   10   40
[2,]    5   45
```

```
print(sprintf("OR estrato 2 (Grave) =%.3f", or2))
```

```
[1] "OR estrato 2 (Grave) = 2.250"
```

```
# Array de tablas (estratificación)
tables <- array(c(a1, b1, c1, d1, a2, b2, c2, d2), dim = c(2, 2, 2))
tables <- aperm(tables, c(2, 1, 3))  # Reordenar dimensiones para epitools

# 1. OR Crudo (sumamos los estratos)
total_table <- table1 + table2
or_crudo <- (total_table[1,1] * total_table[2,2]) / (total_table[1,2] * total_table[2,1])
cat(sprintf("Tabla total:\n")); print(total_table)
```

```
Tabla total:
```

```
     [,1] [,2]
[1,]   50   50
[2,]   15   85
```

```
# 2. OR Ajustado por Mantel-Haenszel
mh <- mantelhaen.test(tables, correct = FALSE)
or_mh <- mh$estimate

cat(sprintf("OR Crudo: %.3f\n", or_crudo))
```

```
OR Crudo: 5.667
```

```
cat(sprintf("OR Mantel-Haenszel: %.3f\n", or_mh))
```

```
OR Mantel-Haenszel: 6.833
```

```
# 3. Aplicación de la regla del 10%
diff_pct <- abs(or_crudo - or_mh) / or_crudo * 100
cat(sprintf("Diferencia relativa: %.1f%%\n", diff_pct))
```

```
Diferencia relativa: 20.6%
```

Interpretación Epidemiológica: En este estudio, el OR crudo (total) es 5.67, mientras que el OR ajustado (MH) es 6.83. La diferencia relativa es de 20.6 %. Al ser mayor del 10 %, existe un efecto de confusión sustancial provocado por la "gravedad".

La gravedad actúa como un factor de confusión ya que está asociada tanto a la probabilidad de recibir el Fármaco A (sesgo de indicación) como a la probabilidad de recuperación (pronóstico). En este caso concreto, el OR crudo **subestima** el efecto real ($5.67 < 6.83$) al mezclar el impacto del fármaco con la mayor o menor gravedad basal de los pacientes; en otros contextos la confusión puede ir en direccion contraria y sobreestimar el efecto. Al ajustar por MH, eliminamos esta confusión.

Sin embargo, es crucial notar que existe una **interacción** (modificación del efecto): el fármaco parece ser más eficaz en pacientes leves que en pacientes graves. En presencia de una interacción tan marcada, la mejor recomendación epidemiológica es presentar los **resultados estratificados**, ya que un único OR ajustado (MH) ocultaría esta heterogeneidad clínicamente relevante.

Por ultimo, para superar las limitaciones de los métodos tradicionales, la generalización de la estandarización mediante la **g-formula** (Robins, 1987) representa hoy en día una alternativa avanzada y robusta para ajustar el efecto del fármaco. Para aquellos interesados en introducirse en la inferencia causal desde una perspectiva tanto histórica como computacional, se recomienda como punto de partida el tutorial *Computational Causal Inference* (Smith & Luque-Fernandez, 2020, publicado en *Statistics in Medicine*), el cual ofrece una excelente guía con ejemplos prácticos implementados en **R, Python y Stata**.

11.8.3. Control multivariable de la confusión: regresión logística múltiple con `rlogitm()`

El estimador de MantelHaenszel es elegante y didácticamente potente, pero presenta dos limitaciones operativas importantes:

1. **Sólo permite controlar por una variable de confusión categórica** (la que define los estratos). Si necesitamos ajustar simultáneamente por edad, IMC, tabaquismo y tiempo de evolución, la estratificación produce demasiados estratos con frecuencias muy pequeñas en cada uno.
2. **Asume homogeneidad del OR a través de los estratos**. Cuando esa hipótesis se rompe (interacción, como en el ejemplo previo), el OR de M–H promedia efectos heterogéneos y oculta la modificación del efecto.

La **regresión logística múltiple** generaliza el ajuste por confusión: permite incluir un número arbitrario de covariables (continuas o categoricas) simultáneamente, estima un OR ajustado para cada predictor manteniendo constantes las demás, y proporciona una medida global de capacidad discriminante mediante el **área bajo la curva ROC** (AUC). El paquete `BioEstatR` integra todo este flujo en la función `rlogitm()` (ver Apéndice B), que devuelve:

- Estimación de los coeficientes con sus odds ratios e intervalos de confianza.
- Bondad de ajuste mediante el **test de Hosmer–Lemeshow**.
- Pseudo-R^2 de **Nagelkerke**.
- **AUC** con su intervalo de confianza al 95 %.
- Curva ROC (opcional, con `grf = TRUE`).

Ejemplo: osteoporosis del cuello femoral ajustada por múltiples factores

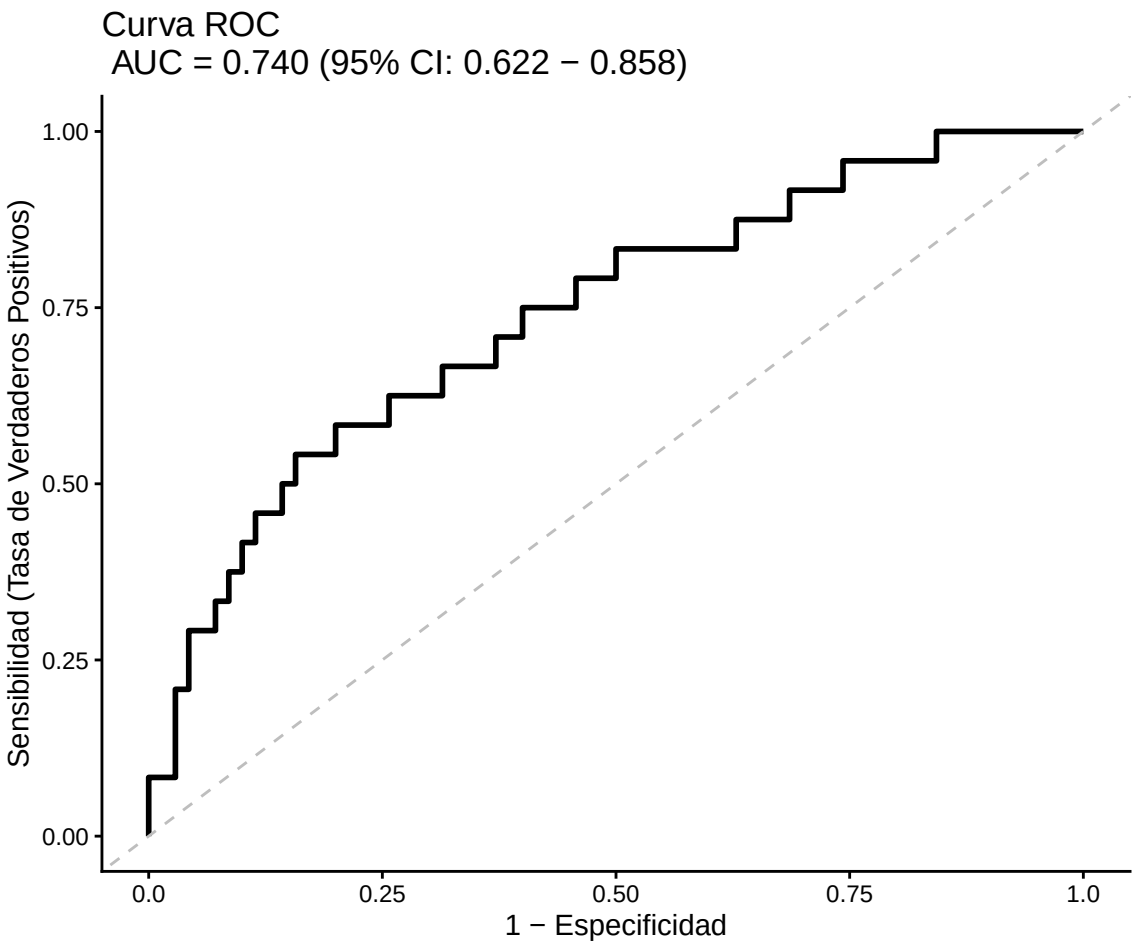
Volvamos al estudio observacional de la sección de tablas de contingencia, donde habíamos detectado una asociación entre tabaquismo y osteoporosis del cuello femoral (`osteo_cue`) en 94 pacientes diabéticos. La pregunta clínica relevante es: **¿persiste la asociación tras controlar simultáneamente por edad, indice de masa corporal y tiempo de evolución de la diabetes?**

```
library(BioEstatR)
data(osteo)

# Modelo: osteoporosis del cuello femoral ajustada por tabaco + edad + IMC + tevol
rlogitm(osteo_cue ~ tabaco + edad + imc + tevol,
        data = osteo,
        grf  = TRUE)
```

```
Regresión logística  multiple
----------------------------------------------------------------
# Información muestral ---

   Tamaño muestral (N inicial) :  94
   Tamaño muestral tras eliminar valores perdidos (Casos completos) :  94
   Mínima frecuencia de eventos (n efectivo) :  24

# Distribución de la variable respuesta (osteo_cue) ---

  Categoria  n Porcentaje
1        No 70       74.5
2        Sí 24       25.5

# Modelo logístico ---  ---
```

```
 Modelo :  osteo_cue ~ tabaco + edad + imc + tevol
Devianza residual:  92  (Nula:  107 )
AIC:  102
R² de Nagelkerke:  0.214

 Test de bondad de ajuste de Hosmer-Lemeshow :
X² =  2.07 , gl =  8 ,   = 0.979

 Capacidad discriminante :
 AUC (Area bajo la curva ROC)  =  0.74

 Coeficientes del modelo :
```

	Termino	Estimacion	Error_Std	z_exp		sig	OR	OR_inf	OR_sup
1	(Intercept)	2.893	2.269	1.275	=	0.202	18.045	0.275	2237.565
2	tabacoSí	0.865	0.549	1.575	=	0.115	2.376	0.821	7.218
3	edad	-0.004	0.037	-0.105	=	0.917	0.996	0.925	1.072
4	imc	-0.224	0.105	-2.140	=	0.032	0.799	0.638	0.965
5	tevol	0.071	0.035	2.020	=	0.043	1.074	1.004	1.154



Interpretación epidemiológica:

- **Tabaquismo (variable de exposición principal):** tras ajustar por edad, IMC y tiempo de evolución, los fumadores presentan un OR ajustado ≈ 2.4 (IC 95 %: 0.82 – 7.22) frente a los no fumadores. La asociación cruda obtenida en la tabla 2×2 (OR ≈ 2.8, $p = 0.031$) se atenúa ligeramente al ajustar por las demás covariables y pierde significación estadística al 5 %, lo que sugiere que parte del efecto observado podría estar mediado o confundido por la edad y el IMC. Este es un ejemplo clásico de cómo el **OR ajustado por regresión multivariable** difiere del OR crudo y, también, del OR de Mantel–Haenszel restringido a un único estratificador.
- **IMC:** muestra una asociación inversa estadísticamente significativa (OR ≈ 0.80, $p \approx 0.03$). Cada unidad adicional de IMC reduce las odds de osteoporosis del cuello femoral aproximadamente un 20 %, en consonancia con la literatura sobre el efecto protector de la masa corporal sobre la densidad ósea.
- **Tiempo de evolución de la diabetes (`tevol`):** asociación positiva significativa (OR ≈ 1.07, $p \approx 0.04$). Cada año adicional de evolución incrementa las odds de osteoporosis del cuello femoral en torno a un 7 %.
- **Edad:** no aporta información independiente del resto de predictores ($p \approx 0.92$), proba-

blemente porque su efecto está capturado por `tevol` (variable altamente correlacionada con la edad en pacientes diabéticos).

Bondad de ajuste y discriminación:

- El test de Hosmer–Lemeshow no rechaza el ajuste del modelo ($\chi^2 \approx 2.07$, $p \approx 0.98$): las probabilidades predichas son coherentes con las frecuencias observadas en deciles de riesgo.
- El pseudo-R^2 de Nagelkerke alcanza ≈ 0.21, un valor moderado para datos observacionales en bioestadística clínica.
- El AUC de la curva ROC es ≈ 0.74, indicando una **capacidad discriminante aceptable**: en términos prácticos, si tomamos al azar a un paciente con osteoporosis y a otro sin osteoporosis, la probabilidad predicha de osteoporosis es mayor en el primero el 74 % de las veces.

Mantel–Haenszel vs. regresión logística múltiple

Criterio	Mantel–Haenszel	Regresión logística múltiple
Nº de covariables de ajuste	1 (categórica, define estratos)	Arbitrario (continuas y categóricas)
Supuesto sobre OR	Homogeneidad entre estratos	Linealidad en el log-odds; no interacción salvo que se modelice
Output principal	OR ajustado único	OR ajustado por covariable + bondad de ajuste + AUC
Detección de interacción	Test de Breslow–Day	Términos producto en la fórmula
Cuando elegirlo	Análisis exploratorio rápido, una sola covariable de confusión	Análisis confirmatorio, múltiples confusores, predicción

En la práctica epidemiológica moderna ambas técnicas se utilizan de forma complementaria: M–H ofrece una **lectura rápida y transparente** del efecto de una variable de exposición estratificada, mientras que la **regresión logística múltiple** proporciona el modelo de ajuste definitivo con todas las covariables relevantes y una medida global de discriminación (AUC).

11.9. Ejercicios del tema

Ejercicio 11.1: En un estudio sobre el consumo de alcohol y diabetes tipo 2 se obtiene la siguiente tabla 2×2:

	Diabetes	Sin diabetes	Total
Bebedor habitual	45	105	150
No bebedor	20	130	150

a) Calcule las frecuencias esperadas bajo H_0. b) Calcule el estadístico χ^2 manualmente. c) ¿Existe asociación con $\alpha = 0.05$?

Ejercicio 11.2: Un ensayo clínico compara tres tratamientos para la hipertensión: A (control), B (dosis baja), C (dosis alta). Los resultados son:

Tratamiento	Controlada	No controlada
A	25	35
B	40	20
C	45	15

Usando la prueba χ^2 de homogeneidad con $\alpha = 0.05$, ¿difieren los tratamientos en su eficacia? ($gl = 2$, $\chi^2_{0.05} = 5.99$).

Ejercicio 11.3: Una tabla 2×2 tiene celdas $(3, 12, 8, 7)$. a) Compruebe si se cumple la condición de aplicación del χ^2. b) ¿Qué alternativa utilizaría? c) Aplique la alternativa e interprete.

Ejercicio 11.4 (McNemar): En un estudio de concordancia diagnóstica, 80 pacientes con sospecha de infección por *Helicobacter pylori* son evaluados con dos pruebas: test de aliento con urea (TAU) y serología (SERO). Los resultados son:

	SERO: +	**SERO: −**	**Total**
TAU: +	38	12	50
TAU: −	5	25	30
Total	43	37	80

a) ¿Por qué es incorrecto aplicar el test χ^2 de independencia aquí?
b) Identifique los pares discordantes (b y c) e indique cuántos son.
c) Calcule el estadístico de McNemar (con y sin corrección de Yates).
d) Difieren significativamente las dos pruebas diagnósticas en su tasa de positividad? ¿Cuál detecta más casos?

Ejercicio 11.5: Un estudio de cohortes sigue a 1000 trabajadores expuestos a un contaminante y 1000 no expuestos durante 5 años. Se observan 60 casos de enfermedad pulmonar en expuestos y 20 en no expuestos. a) Calcule el RR e interprete. b) Calcule el IC 95 %. c) Calcule el RD y el NNH.

Ejercicio 11.6: En un estudio caso-control sobre infarto de miocardio, se estudia la asociación con hipertensión. Se reclutan 200 casos (infartos) y 200 controles. De los casos, 120 son hipertensos; de

los controles, 80. a) Construya la tabla 2×2. b) Calcule el OR. c) Calcule el IC 95 % del OR. d) ¿Por qué no es adecuado calcular el RR en este diseño?

Ejercicio 11.7: En un estudio transversal sobre la prevalencia de diabetes y obesidad en 800 adultos, 200 son obesos (IMC ≥ 30) y 600 no obesos. Entre los obesos, 60 tienen diabetes; entre los no obesos, 40 tienen diabetes. a) Calcule la RP. b) Calcule el OR. c) ¿Cuál es la medida más apropiada? d) ¿Son similares los resultados? Explique por qué.

11.10. Respuestas a los Ejercicios

Ejercicio 11.1: a) $E_{11} = 150 \times 65/300 = 32.5$; $E_{12} = 150 \times 235/300 = 117.5$; $E_{21} = 32.5$; $E_{22} = 117.5$. b) $\chi^2 = (45-32.5)^2/32.5 + (105-117.5)^2/117.5 + (20-32.5)^2/32.5 + (130-117.5)^2/117.5 \approx 4.81 + 1.33 + 4.81 + 1.33 = 12.28$. c) $\chi^2_{0.05,1} = 3.84$. Como $12.28 > 3.84$ ($p < 0.001$), rechazamos H_0: existe asociación significativa entre consumo de alcohol y diabetes tipo 2.

Ejercicio 11.2: Frecuencias esperadas: $n = 180$, marginales columna: 110 controladas, 70 no controladas. $E_{A,ctrl} = 60 \times 110/180 = 36.67$; calculando todos: $\chi^2 \approx 15.2 > 5.99$, $gl = 2$, $p < 0.001$. Los tres tratamientos tienen eficacias significativamente diferentes. La dosis alta (75 % controlados) es la más eficaz.

Ejercicio 11.3: a) $E_{11} = 11 \times 11/30 = 4.03 < 5$; condición NO cumplida. b) Usar **prueba exacta de Fisher**. c) Con $n = 30$ y celdas $(3, 12, 8, 7)$: $OR = 3 \times 7/(12 \times 8) = 0.22$. La prueba exacta de Fisher daría $p \approx 0.07$, no significativo al 5 %.

Ejercicio 11.4 (McNemar): a) Las dos pruebas se aplican al mismo paciente, por lo que las observaciones son dependientes entre sí. El test χ^2 de independencia requiere que todas las observaciones sean independientes; violarlo inflaría artificialmente el tamaño de muestra efectivo. b) Pares discordantes: $b = 12$ (TAU+/SERO−) y $c = 5$ (TAU−/SERO+). Total: $b + c = 17$. c) Como $b + c = 17 < 25$, se recomienda la corrección de Yates:

- Sin corrección: $\chi^2 = (12-5)^2/17 = 49/17 = 2.88$
- Con corrección (Yates): $\chi^2 = (|12-5|-1)^2/17 = 36/17 = 2.12$

d) Sin corrección: $\chi^2 = 2.88$, $p \approx 0.090$. Con corrección de Yates: $\chi^2 = 2.12$, $p \approx 0.145$. Con $\alpha = 0.05$, **no rechazamos** H_0 en ningún caso: no hay evidencia suficiente de que las dos pruebas difieran significativamente en su tasa de positividad. No obstante, el TAU detecta más positivos que la serología ($b = 12 > c = 5$), lo que sugiere una tendencia a mayor sensibilidad del TAU, aunque no alcanza significancia estadistica con esta muestra.

Ejercicio 11.5: a) $RR = (60/1000)/(20/1000) = 3.0$. Los expuestos tienen **3 veces más riesgo** de enfermedad pulmonar. b) $SE_{\ln RR} = \sqrt{940/(60 \times 1000) + 980/(20 \times 1000)} = \sqrt{0.01567 + 0.049} = 0.256$; $IC_{95\,\%} = \exp(\ln 3 \pm 0.501) = [1.50, 5.99]$. c) $RD = 0.060 - 0.020 = 0.040$; $NNH = 1/0.040 = 25$ (por cada 25 trabajadores expuestos durante 5 años, 1 caso adicional de enfermedad pulmonar).

Ejercicio 11.6: a) Tabla: hipertensos $(120, 80, 200)$; normotensos $(80, 120, 200)$; totales $(200, 200, 400)$. b) $OR = (120 \times 120)/(80 \times 80) = 14400/6400 = 2.25$. c) $IC_{95\,\%}$: $\exp(\ln 2.25 \pm 1.96\sqrt{1/120 + 1/80 + 1/80 + 1/120}) = \exp(0.811 \pm 0.369) = [1.44, 3.51]$. d) En un estudio caso-control, el muestreo es por resultado (se seleccionan casos y controles), por lo que los totales de fila son fijados por el investigador y no reflejan la incidencia real → no se puede estimar el riesgo ni el RR directamente.

Ejercicio 11.7: a) $RP = (60/200)/(40/600) = 0.30/0.067 = 4.50$. Los obesos tienen 4.5 veces la prevalencia de diabetes. b) $OR = (60 \times 560)/(140 \times 40) = 33600/5600 = 6.0$. c) La medida apropiada en un estudio transversal es la **RP**, no el OR. d) Los resultados difieren notablemente ($RP = 4.5$ vs $OR = 6.0$) porque la prevalencia de diabetes no es rara (50 % en obesos): el OR sobreestima la RP cuando la prevalencia es alta ($> 10\,\%$). Usar siempre RP en estudios transversales.

11.11. Recomendaciones de Métodos Avanzados

> **Tip**
>
> Para profundizar en estos temas, se recomienda consultar textos especializados como *Categorical Data Analysis* de Alan Agresti o literatura específica en bioestadística avanzada enfocada en inferencia causal.
> Para ampliar los contenidos de este capítulo con tecnicas estadísticas avanzadas, visita:
> **→ Bioestadística Avanzada — M.A. Luque Fernández**

11.12. Lecturas Recomendadas Adicionales

Tema	Referencia	Relevancia para este capítulo
Datos categóricos	Agresti [2013]	Tablas de contingencia, ^{2}, modelos log-lineales
Datos apareados	McNemar [1947]	Artículo original de la prueba de McNemar
Tasas y proporciones	Fleiss et al. [2003]	IC exactos, OR, RR y prueba de McNemar
Estadística médica	Altman [1991]	Tablas 2×2 y medidas de asociación
Epidemiología	Rothman et al. [2008]	RR, OR, diseños cohorte/caso-control
Inferencia Causal	Robins [1987]	G-formula para estandarización

Tema	Referencia	Relevancia para este capítulo
Inferencia Causal	Smith and et al. [2020]	Tutorial de inferencia causal computacional (R, Python, Stata)
Epidemiología analítica	Szklo and Nieto [2007]	Medidas de efecto y diseños epidemiológicos
Introducción estadística	Diez et al. [2019]	Inferencia sobre proporciones

Referencias

Sobre esta sección

Esta página reúne las referencias bibliográficas utilizadas en el curso **Matemáticas para la Estadística**. Los materiales están organizados por temas para facilitar la búsqueda según los tópicos de interés.

Libros de Referencia General

Los siguientes textos proporcionan cobertura completa de los temas tratados en el curso:

- Casella and Berger [2002] — Texto clásico y riguroso sobre inferencia estadística, fundamental para cursos avanzados.
- Rice [2007] — Excelente balance entre teoría y práctica con muchos ejemplos.
- Wackerly et al. [2008] — Orientado a aplicaciones en ingeniería y ciencias, con numerosos ejercicios.
- Wasserman [2006] — Curso conciso pero riguroso en inferencia estadística.
- Wood [2015] — Introducción accesible a estadística desde una perspectiva moderna.

Análisis Exploratorio de Datos

Para los tópicos de la Semana 1 y 2 (EDA):

- Tukey [1961] — Artículo seminal que define el Análisis Exploratorio de Datos (EDA).
- Wickham and Grolemund [2016] — Guía moderna para manipulación y visualización de datos en R.

Probabilidad y Variables Aleatorias

Referencias para profundizar en los tópicos de las Semanas 3 y 4:

- Rice [2007] — Excelentes capítulos sobre probabilidad y variables aleatorias.
- Ross [2014] — Enfoque práctico para ingenieros y científicos.
- Devore [2020] — Texto ampliamente usado en cursos de ingeniería.

Teoría del Muestreo e Inferencia Estadística

Para las Semanas 5, 7, 8 y 9:

- Casella and Berger [2002] — Riguroso tratamiento de estimación y contrastes de hipótesis.
- Agresti [2018] — Métodos estadísticos aplicados a ciencias sociales.
- Benjamini and Hochberg [1995] — Control de tasa de falsos positivos en pruebas múltiples.

Regresión Lineal

Para las Semanas 10 y 11:

- James et al. [2013] — Introducción moderna a aprendizaje estadístico con énfasis en regresión.
- Freedman [2007] — Enfoque crítico y reflexivo sobre la práctica de la regresión.
- Wooldridge [2020] — Aplicaciones econométricas de modelos de regresión.
- Greene [2018] — Referencia avanzada en econometría.
- Angrist and Pischke [2008] — Métodos empíricos para inferencia causal.

Análisis de Datos Categóricos y Epidemiología

Para la Semana 12 (tablas de contingencia, Chi-cuadrado, McNemar, medidas de asociación):

- Agresti [2013] — Referencia estándar para el análisis de datos categóricos: tablas de contingencia, pruebas Chi-cuadrado y modelos log-lineales.
- Fleiss et al. [2003] — Tratamiento riguroso de proporciones, tasas y pruebas para tablas 2×2, incluida la prueba de McNemar para datos apareados.
- McNemar [1947] — Artículo original de Quinn McNemar donde se propone la prueba para muestras relacionadas.
- Altman [1991] — Estadística médica aplicada: medidas de asociación (RR, OR, RD), tablas 2×2 y pruebas para datos apareados.
- Rothman et al. [2008] — Epidemiología moderna: diseños de cohorte, caso-control y transversal; estimación e interpretación de RR y OR.
- Szklo and Nieto [2007] — Presentación accesible de epidemiología analítica con énfasis en medidas de efecto.
- Femia Marzo et al. [2024] — Paquete R del grupo de Bioestadística de la UGR (Pedro Femia et al.): funciones `tabla2x2()`, `tablarxc()` y `testnormal()` con el dataset `osteo`.
- Diez et al. [2019] — Introducción accesible a la inferencia sobre proporciones y tablas de contingencia.

Recursos en R

El curso utiliza ampliamente **R** para la implementación de conceptos estadísticos:

- R Core Team [2024] — Documentación oficial del proyecto R.

- Wickham and Grolemund [2016] — Guía práctica para análisis de datos con R.
- Kabacoff [2015] — Referencia completa de gráficos y manipulación en R.

Perspectivas Bayesianas

Para estudiantes interesados en métodos Bayesianos:

- Koop [2011] — Introducción a econometría Bayesiana.
- Gelman et al. [2013] — Texto completo sobre análisis Bayesiano de datos.
- McElreath [2020] — Curso moderno en inferencia Bayesiana con Stan.
- Peng and Dominici [2015] — Pensamiento estadístico desde perspectiva Bayesiana.

Tópicos Especializados

Lecturas complementarias para temas avanzados:

- James et al. [2013] — Aprendizaje estadístico (machine learning clásico).
- Efron and Hastie [2016] — Perspectiva histórica sobre estadística en la era computacional.
- Hyndman and Athanasopoulos [2018] — Métodos de pronóstico y series de tiempo.
- Pearl [2009] — Fundamentos matemáticos de causalidad.
- Varian [2014] — Aplicaciones de estadística a economía y "big data".
- Montgomery et al. [2020] — Estadística aplicada a ingeniería.
- Diez et al. [2019] — Recurso abierto gratuito para aprendizaje de estadística.

Diccionarios y Referencias Rápidas

- Dodge [2006] — Diccionario definitivo de términos estadísticos en inglés.

Cómo citar este curso

Si utiliza los materiales de este curso en su investigación o trabajo, le recomendamos citar:

> Luque Fernández, Miguel Ángel. (2025–2026). Matemáticas para la Estadística Médica. Universidad de Granada. Disponible en: https://migariane.github.io/CursoMatematicaEstadistica/

Acceso Abierto

Alan Agresti. *Categorical Data Analysis*. Wiley, Hoboken, NJ, 3rd edition, 2013.

Alan Agresti. *Statistical Methods for the Social Sciences*. Pearson, 5th edition, 2018.

Douglas G. Altman. *Practical Statistics for Medical Research*. Chapman and Hall/CRC, London, 1991.

Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, Pacific Grove, CA, 2nd edition, 2002.

Laura M. Chihara and Tim C. Hesterberg. *Mathematical Statistics with Resampling and R*. Wiley, Hoboken, NJ, 2nd edition, 2019.

Jay L. Devore. *Probability and Statistics for Engineering and the Sciences*. Cengage Learning, 9th edition, 2020.

David M. Diez, Christopher D. Barr, and Mine Cetinkaya-Rundel. *OpenIntro Statistics*. OpenIntro Inc., 4th edition, 2019. URL https://www.openintro.org.

Yadolah Dodge. *The Oxford Dictionary of Statistical Terms*. Oxford University Press, 2nd edition, 2006.

Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, 2016.

Pedro Femia Marzo, Pedro Carmona Montes, and Miguel Angel Luque-Fernandez. BioEstatR: Paquete de funciones de bioestadística para R, 2024. URL https://github.com/migariane/BioEstatR. Paquete R. Grupo de Bioestadística, Departamento de Estadística e Investigación Operativa, Universidad de Granada.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical Methods for Rates and Proportions*. Wiley, Hoboken, NJ, 3rd edition, 2003.

David A. Freedman. *Statistical Models: Theory and Practice*. Cambridge University Press, 2nd edition, 2007.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition, 2013.

William H. Greene. *Econometric Analysis*. Pearson, 8th edition, 2018.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2nd edition, 2018. URL https://otexts.com/fpp2/.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, New York, 2013.

Robert I. Kabacoff. *R in Action: Data Analysis and Graphics with R*. Manning Publications, 2nd edition, 2015.

Gary Koop. *Bayesian Econometrics*. Wiley, 2nd edition, 2011.

Antonio Martín Andrés and Juan de Dios Luna del Castillo. *Bioestadística para las Ciencias de la Salud*. Norma-Capitel, Madrid, 5th edition, 2004.

Antonio Martín Andrés and Agustín Silva Mato. Choosing the optimal unconditional test for comparing two independent proportions. *Computational Statistics & Data Analysis*, 17(5): 555–574, 1994. doi: 10.1016/0167-9473(94)90148-1.

Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press, 2nd edition, 2020.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Douglas C. Montgomery, George C. Runger, and Norma F. Hubele. *Engineering Statistics*. Wiley, 6th edition, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Roger D. Peng and Francesca Dominici. *Statistical Rethinking*. Chapman and Hall/CRC, 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org/.

John A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, 3rd edition, 2007.

James M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7 (5-8):1393–1512, 1987. doi: 10.1016/0270-0255(86)90088-6.

Sheldon M. Ross. *Introduction to Probability and Statistics for Engineers and Scientists*. Academic Press, 5th edition, 2014.

Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. *Modern Epidemiology*. Lippincott Williams and Wilkins, Philadelphia, PA, 3rd edition, 2008.

Feras A. Saad, Cameron E. Freer, Martin C. Rinard, and Vikash K. Mansinghka. Optimal approximate sampling from discrete probability distributions. *Proceedings of the ACM on Programming Languages*, 4(POPL), 2020. doi: 10.1145/3371104.

Samuel Sanford Shapiro and Martin Bradley Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611, 1965.

Matthew Smith and et al. Computational causal inference. *Statistics in Medicine*, 39(1):1–35, 2020. doi: 10.1002/sim.9234.

Moyses Szklo and F. Javier Nieto. *Epidemiology: Beyond the Basics*. Jones and Bartlett, Sudbury, MA, 2nd edition, 2007.

John W. Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1961.

Hal R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2): 3–28, 2014.

Dennis D. Wackerly, William Mendenhall, and Richard L. Scheaffer. *Mathematical Statistics with Applications*. Duxbury Press, Belmont, CA, 7th edition, 2008.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, 2006.

Hadley Wickham and Garrett Grolemund. *R for Data Science*. O'Reilly Media, 2016. URL https://r4ds.had.co.nz/.

Simon N. Wood. *Core Statistics*. Cambridge University Press, 2015.

Jeffrey M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, 7th edition, 2020.

Última actualización: Semestre 2025–2026

Apéndice A

Apéndice A: Repaso Matemático Riguroso

Este apéndice proporciona los fundamentos matemáticos necesarios para un análisis riguroso de la probabilidad y la estadística. Este contenido está estructurado como una guía de referencia basada en principios de bioestadística.

A.1. A.1 Funciones y Gráficas

Las funciones mapean elementos de un dominio a un codominio ($f : X \to Y$).

- **Notación:** $y = f(x)$, donde x es la variable independiente y y es la variable dependiente.
- **Dominio:** Conjunto de todos los posibles valores de entrada (x).
- **Codominio/Rango:** Conjunto de todos los posibles valores de salida (y).
- **Funciones Inversas:** Si una función f es biyectiva (inyectiva y sobreyectiva), existe una función inversa f^{-1} tal que $f(f^{-1}(x)) = x$ y $f^{-1}(f(x)) = x$.
- **Composición:** La composición de funciones $(f \circ g)(x)$ significa aplicar g primero, y luego f al resultado: $(f \circ g)(x) = f(g(x))$.

Ejemplos de funciones comunes: * **Lineal:** $f(x) = mx + b$ * **Cuadrática:** $f(x) = ax^2 + bx + c$ * **Exponencial:** $f(x) = a^x$ * **Logarítmica:** $f(x) = \log_b(x)$

Ejemplo en R: Definición y uso de funciones

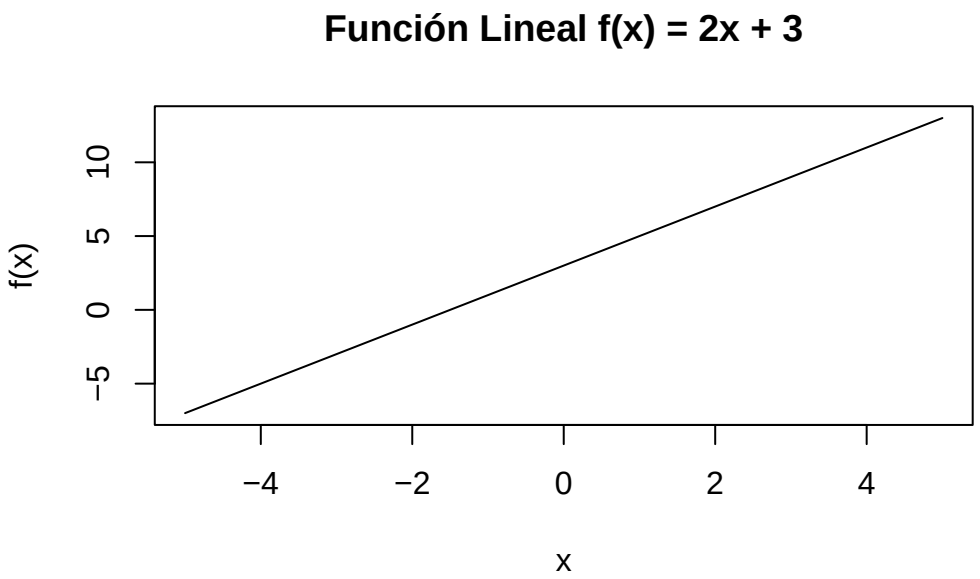
Cuadro A.1 Código R

```
# Definir una función lineal
f_lineal <- function(x) {
  2*x + 3
}
f_lineal(5) # Resultado: 13
```

```
[1] 13
```

Cuadro A.2 Código R

```
# Graficar una función
curve(f_lineal, from = -5, to = 5,
      main = "Función Lineal f(x) = 2x + 3",
      xlab = "x", ylab = "f(x)")
```



A.2. A.2 Exponenciales y Logaritmos

Fundamentales para modelizar relaciones de crecimiento/decrecimiento y simplificar funciones de verosimilitud en estadística.

- **Propiedades Exponenciales:**
  - $a^x a^y = a^{x+y}$
  - $(a^x)^y = a^{xy}$
  - $a^0 = 1$
  - $a^{-x} = 1/a^x$
  - $(ab)^x = a^x b^x$
  - En R, `exp(x)` calcula e^x.
- **Propiedades Logarítmicas (base natural** $\ln$ **o** $\log_e$**):**
  - $\ln(xy) = \ln(x) + \ln(y)$
  - $\ln(x/y) = \ln(x) - \ln(y)$
  - $\ln(x^k) = k \ln(x)$
  - $\ln(1) = 0$
  - $\ln(e) = 1$
  - $\log_b x = \frac{\ln x}{\ln b}$ (cambio de base)
  - La relación inversa: $e^{\ln x} = x$ y $\ln(e^x) = x$.

- En R, `log(x)` calcula $\ln(x)$ por defecto; `log10(x)` para base 10, `log(x, base=b)` para otras bases.

Ejemplo en R: Cálculos con exponenciales y logaritmos

Cuadro A.3 Código R

```
# Cálculos exponenciales
exp(1)      # e^1
```

```
[1] 2.718282
```

Cuadro A.4 Código R

```
exp(2)      # e^2
```

```
[1] 7.389056
```

Cuadro A.5 Código R

```
# Cálculos logarítmicos
log(exp(1)) # ln(e) = 1
```

```
[1] 1
```

```
[1] 4.60517
```

```
[1] 2
```

```
[1] FALSE
```

A.3. A.3 Sumatorias y Productos

La notación $\sum$ representa sumas y $\prod$ representa productos. Son esenciales para definir medias, varianzas y funciones de verosimilitud.

- **Notación de Sumatoria:** $\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$
- **Propiedades de la Suma:**
  - Linealidad: $\sum_{i=1}^{n}(aX_i + bY_i) = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} Y_i$
  - Suma de una constante: $\sum_{i=1}^{n} c = nc$
- **Sumas dobles:** $\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}$ para sumar elementos en una tabla o matriz.
- **Notación de Producto:** $\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot \ldots \cdot x_n$

Ejemplo en R: Sumas y Productos

```
[1] 30
```

```
Media muestral: 6
```

Cuadro A.6 Código R

```r
log(100)    # ln(100)
```

Cuadro A.7 Código R

```r
log(100, base = 10) # log10(100) = 2
```

```
[1] 3840
```

```
     [,1] [,2]
[1,]    1    3
[2,]    2    4
```

```
[1] 10
```

A.4. A.4 Teoría de Conjuntos

La teoría de conjuntos es el lenguaje fundamental para describir eventos en probabilidad.

- **Conjunto:** Colección de objetos distintos.
  - $A = \{1, 2, 3\}$
  - Ω = Espacio muestral (conjunto de todos los resultados posibles)
- **Elemento:** Un objeto dentro de un conjunto.
  - $x \in A$ (x es un elemento de A)
  - $y \notin A$ (y no es un elemento de A)
- **Subconjunto:** $A \subset B$ (todos los elementos de A también están en B).
- **Conjunto Vacío:** $\emptyset$ o `{}` (contiene cero elementos).
- **Operaciones:**
  - **Unión ($\cup$):** $A \cup B = \{x | x \in A \text{ o } x \in B\}$. Elementos en A O B (o ambos).
  - **Intersección ($\cap$):** $A \cap B = \{x | x \in A \text{ y } x \in B\}$. Elementos en A Y B.
  - **Complemento (A^c o A'):** Elementos en el espacio muestral Ω que no están en A. Formalmente, $A^c = \{x | x \in \Omega \text{ y } x \notin A\}$.
  - **Diferencia ($A \setminus B$):** Elementos en A pero no en B. Equivalente a $A \cap B^c$.
- **Conjuntos Disjuntos:** A y B son disjuntos si $A \cap B = \emptyset$. No tienen elementos en común.
- **Partición:** Una colección de conjuntos $A_1, A_2, \ldots, A_k$ forma una partición de Ω si son mutuamente disjuntos y su unión es Ω.

Ejemplo en R: Operaciones con conjuntos (usando vectores lógicos)

```
[1] 1 2 3 4 5 6 7
```

```
[1] 4 5
```

```
[1] 1 2 3
```

```
[1]  6  7  8  9 10
```

Cuadro A.8 Código R

```
# Propiedad: log(x^k) = k * log(x)
x_val <- 5
k_val <- 3
log(x_val^k_val) == k_val * log(x_val)
```

Cuadro A.9 Código R

```
# Vector de datos
datos <- c(2, 4, 6, 8, 10)

# Suma de los datos
sum(datos)
```

```
[1] FALSE
```

A.5. A.5 Cálculo: Derivadas e Integrales

Herramientas fundamentales para la optimización (minimización de funciones de pérdida, maximización de verosimilitud) usando derivadas y para el cálculo de probabilidades e inferencia usando integrales.

A.5.1. A.5.1 Derivadas

Miden la tasa de cambio instantánea de una función. En estadística, se usan para encontrar máximos y mínimos de funciones (como funciones de verosimilitud o de costo).

- **Notación:** $f'(x)$, $\frac{dy}{dx}$, $\frac{d}{dx}f(x)$.
- **Reglas Básicas:**
  - Constante: $\frac{d}{dx}c = 0$
  - Potencia: $\frac{d}{dx}x^n = nx^{n-1}$
  - Suma/Resta: $\frac{d}{dx}(f(x) \pm g(x)) = f'(x) \pm g'(x)$
  - Producto: $\frac{d}{dx}(f(x)g(x)) = f'(x)g(x) + f(x)g'(x)$
  - Cociente: $\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}$
  - Regla de la cadena: $\frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$
  - Exponencial: $\frac{d}{dx}e^x = e^x$
  - Logaritmo: $\frac{d}{dx}\ln(x) = \frac{1}{x}$

Ejemplo: Minimización de una función (encontrar el vértice de una parábola) Sea $f(x) = x^2 - 4x + 5$. Para encontrar el mínimo, derivamos e igualamos a cero: $f'(x) = 2x - 4$ $2x - 4 = 0 \implies x = 2$. Sustituyendo en $f(x)$: $f(2) = 2^2 - 4(2) + 5 = 4 - 8 + 5 = 1$. El mínimo es en $(2, 1)$.

Ejemplo en R: Derivación simbólica (para funciones simples)

```
2 * x - 4
```

Cuadro A.10 Código R

```r
# Media muestral usando sumatoria
n_datos <- length(datos)
media <- sum(datos) / n_datos
cat("Media muestral:", media, "\n")
```

Cuadro A.11 Código R

```r
# Producto de los datos
prod(datos)
```

```
[1] 0
```

A.5.2. A.5.2 Integrales

Representan el área bajo una curva y se utilizan para calcular probabilidades a partir de funciones de densidad, esperanzas y varianzas.

- **Integral Indefinida (Antiderivada):** $\int f(x)dx = F(x) + C$, donde $F'(x) = f(x)$.
- **Integral Definida:** $\int_a^b f(x)dx = F(b) - F(a)$. Esto calcula el área neta entre $f(x)$ y el eje x desde a hasta b.
- **Integrales Múltiples:** Para funciones de densidad conjuntas.

Ejemplo: Cálculo de una probabilidad a partir de una PDF Si $f(x)$ es la PDF, $P(a \leq X \leq b) = \int_a^b f(x)dx$.

Ejemplo en R: Integración numérica

```
[1] 0.6826895
```

```
[1] 1
```

A.6. A.6 Álgebra Matricial

El álgebra matricial es indispensable en estadística multivariante, regresión lineal (donde los coeficientes se estiman usando inversas matriciales), análisis de componentes principales, y más.

- **Matriz:** Un arreglo rectangular de números.
  - Notación: $\mathbf{A}_{m \times n}$ (m filas, n columnas).
- **Vector:** Una matriz con una sola fila (vector fila) o una sola columna (vector columna).
- **Tipos de Matrices:**
  - **Identidad (I):** Matriz cuadrada con 1s en la diagonal y 0s en el resto. $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$.
  - **Transpuesta ($\mathbf{A}^T$):** Se obtienen intercambiando filas por columnas. $(A^T)_{ij} = A_{ji}$.
  - **Simétrica:** $\mathbf{A} = \mathbf{A}^T$.
- **Operaciones Básicas:**
  - **Suma/Resta:** Elemento a elemento (las matrices deben tener las mismas dimensiones).

Cuadro A.12 Código R

```
# Suma doble (ej. suma de una matriz)
matriz_ej <- matrix(1:4, nrow=2)
print(matriz_ej)
```

Cuadro A.13 Código R

```
sum(matriz_ej) # Suma todos los elementos
```

- **Multiplicación por un escalar:** Multiplicar cada elemento por el escalar.
- **Producto Matricial (AB):** El número de columnas de $\mathbf{A}$ debe ser igual al número de filas de $\mathbf{B}$.
  - Si $\mathbf{A}_{m\times p}$ y $\mathbf{B}_{p\times n}$, entonces $(\mathbf{AB})_{m\times n}$.
  - $(\mathbf{AB})_{ij} = \sum_{k=1}^{p} A_{ik}B_{kj}$.
  - NO es conmutativo: $\mathbf{AB} \neq \mathbf{BA}$ en general.
- **Inversa de una Matriz ($\mathbf{A}^{-1}$):** Para una matriz cuadrada $\mathbf{A}$, si existe $\mathbf{A}^{-1}$, entonces $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. Solo existe si el determinante es distinto de cero.
- **Determinante ($\det(\mathbf{A})$ o $|\mathbf{A}|$):** Un escalar asociado a una matriz cuadrada que indica, entre otras cosas, si la matriz es invertible (si $\det(\mathbf{A}) \neq 0$).

Ejemplo en R: Operaciones con matrices

```
[1] "Matriz A:"

     [,1] [,2]
[1,]    1    2
[2,]    3    4

[1] "Matriz B:"

     [,1] [,2]
[1,]    5    6
[2,]    7    8

[1] "A + B:"

     [,1] [,2]
[1,]    6    8
[2,]   10   12

[1] "A %*% B:"

     [,1] [,2]
[1,]   19   22
[2,]   43   50

[1] "Transpuesta de A (A^T):"
```

Cuadro A.14 Código R

```r
# Representación de conjuntos como vectores lógicos o numéricos
universo <- 1:10
conjuntoA <- c(1, 2, 3, 4, 5)
conjuntoB <- c(4, 5, 6, 7)

# Unión
union(conjuntoA, conjuntoB)
```

Cuadro A.15 Código R

```r
# Intersección
intersect(conjuntoA, conjuntoB)
```

```
     [,1] [,2]
[1,]    1    3
[2,]    2    4

[1] "Inversa de A:"

     [,1] [,2]
[1,] -2.0  1.0
[2,]  1.5 -0.5

[1] "A %*% solve(A):"

     [,1]         [,2]
[1,]    1 1.110223e-16
[2,]    0 1.000000e+00

[1] "Determinante de A:"

[1] -2
```

A.7. A.7 Referencias

1. **Iowa Biostatistics Math Review:** https://iowabiostat.github.io/math-review/
2. **Casella, G., & Berger, R. L. (2002).** *Statistical Inference*. (Excelente para fundamentos matemáticos de la inferencia estadística).
3. **Spivak, M. (2006).** *Calculus*. (Un texto clásico para un tratamiento riguroso del cálculo).
4. **Harville, D. A. (1997).** *Matrix Algebra From a Statistician's Perspective*. (Referencia avanzada para álgebra matricial en estadística).

Cuadro A.16 Código R

```
# Diferencia (elementos en A que no están en B)
setdiff(conjuntoA, conjuntoB)
```

Cuadro A.17 Código R

```
# Complemento (respecto a 'universo')
setdiff(universo, conjuntoA)
```

Cuadro A.18 Código R

```
# Comprobación de disjunción (no disjuntos en este caso)
length(intersect(conjuntoA, conjuntoB)) == 0
```

Cuadro A.19 Código R

```
# Definir una expresión simbólica
expr <- expression(x^2 - 4*x + 5)
derivada <- D(expr, "x")
print(derivada)
```

Cuadro A.20 Código R

```
# Evaluar la derivada en un punto
eval(derivada, list(x = 2)) # Debería ser 0
```

Cuadro A.21 Código R

```
# Calcular la probabilidad P(-1 <= X <= 1) para una Normal Estándar
# Usamos la función dnorm (densidad de la normal)
resultado_integral <- integrate(dnorm, lower = -1, upper = 1)
print(resultado_integral$value) # Aproximadamente 0.6827 (regla del 68%)
```

Cuadro A.22 Código R

```
# Para la esperanza de una variable continua con PDF f(x): E(X) = integral(x * f(x) dx)
# Ejemplo con una función simple x * exp(-x) de 0 a Inf
f_ejemplo_esperanza <- function(x) { x * dexp(x, rate = 1) } # E(X) para Exp(1) es 1/rate = 1
esperanza_numerica <- integrate(f_ejemplo_esperanza, lower = 0, upper = Inf)
print(esperanza_numerica$value) # Debería ser cercano a 1
```

Cuadro A.23 Código R

```
# Definir matrices
A <- matrix(c(1, 2, 3, 4), nrow=2, byrow=TRUE) # Matriz 2x2
B <- matrix(c(5, 6, 7, 8), nrow=2, byrow=TRUE) # Matriz 2x2
C <- matrix(c(1, 0, 0, 1), nrow=2, byrow=TRUE) # Matriz identidad I

print("Matriz A:")
```

Cuadro A.24 Código R

```
print(A)
```

Cuadro A.25 Código R

```
print("Matriz B:")
```

Cuadro A.26 Código R

```
print(B)
```

Cuadro A.27 Código R

```
# Suma de matrices
print("A + B:")
```

Cuadro A.28 Código R

```
print(A + B)
```

Cuadro A.29 Código R

```
# Producto matricial
print("A %*% B:")
```

Cuadro A.30 Código R

```
print(A %*% B)
```

Cuadro A.31 Código R

```
# Transpuesta de A
print("Transpuesta de A (A^T):")
```

Cuadro A.32 Código R

```
print(t(A))
```

Cuadro A.33 Código R

```
# Inversa de A
# Solo si la matriz es cuadrada y no singular (determinante != 0)
print("Inversa de A:")
```

Cuadro A.34 Código R

```
print(solve(A))
```

Cuadro A.35 Código R

```
# Verificar A * A_inv = I
print("A %*% solve(A):")
```

Cuadro A.36 Código R

```
print(A %*% solve(A)) # Debería ser la matriz identidad
```

Cuadro A.37 Código R

```
# Determinante de A
print("Determinante de A:")
```

Cuadro A.38 Código R

```
print(det(A)) # (1*4 - 2*3) = 4 - 6 = -2
```

Apéndice B

Apéndice B: El Paquete `BioEstatR` — Guía de Referencia

El paquete **`BioEstatR`** (versión 1.0.0, 2026) ha sido desarrollado en la Unidad Docente de Bioestadística del Departamento de Estadística e Investigación Operativa de la Universidad de Granada por **Pedro Femia Marzo** y **Miguel Ángel Luque Fernández**. Proporciona un conjunto de funciones de alto nivel que integran descripción, inferencia, modelización (lineal y logística) y visualización en una sola llamada, diseñadas específicamente para la docencia de bioestadística en ciencias de la salud.

- **Sitio web del paquete:** https://migariane.github.io/BioEstatR/
- **Repositorio GitHub:** https://github.com/migariane/BioEstatR
- **Viñeta introductoria:** https://migariane.github.io/BioEstatR/articles/BioEstatR-overview.html
- **Contacto:** `pfemia@ugr.es`

B.1. B.1 Instalación y el conjunto de datos `osteo`

Instalación

Cuadro B.1 Instalación de BioEstatR desde GitHub

```
# 1. Instalar el paquete remotes si no lo tienes
install.packages("remotes")

# 2. Instalar BioEstatR directamente desde GitHub (compila para tu sistema)
remotes::install_github("migariane/BioEstatR")

# 3. Cargar el paquete y el conjunto de datos osteo
library(BioEstatR)
data(osteo)
```

Para obtener la referencia bibliográfica del paquete:

```
citation("BioEstatR")
```

El paquete incluye el dataset `osteo`, con datos reales de **94 pacientes diabéticos** evaluados en la Unidad de Diabetes de la Facultad de Medicina de la Universidad de Granada. Es el conjunto de referencia de todos los ejemplos de este apéndice y del libro.

El dataset `osteo`

Cuadro B.2 Estructura del dataset osteo

```
library(BioEstatR)
data(osteo)
str(osteo)
```

```
'data.frame':   94 obs. of  27 variables:
 $ num       : num  1 2 3 4 5 6 7 8 9 10 ...
 $ edad      : num  23 35 51 44 54 36 21 20 27 27 ...
 $ grupo_edad: Factor w/ 3 levels "< 25","> 33",..: 1 2 2 2 2 2 1 1 3 3 ...
 $ sexo      : Factor w/ 2 levels "Hombre","Mujer": 1 1 1 1 2 2 1 2 2 2 ...
 $ peso      : num  72.3 82.5 61 67 51 61.5 73.8 62 46 49 ...
 $ talla     : num  175 173 170 167 168 158 180 170 148 144 ...
 $ imc       : num  23.6 27.6 21.1 24 18.1 ...
 $ tevol     : num  2 4 17 17 30 2 3 12 16 19 ...
 $ tabaco    : Factor w/ 2 levels "No","Sí": 2 1 2 2 2 2 2 2 1 1 ...
 $ alcohol   : Factor w/ 3 levels "Excesivo","Moderado",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ ingca     : Factor w/ 2 levels "Insuficiente",..: 2 2 2 1 2 2 2 2 2 1 ...
```

```
 $ acfis     : Factor w/ 2 levels "No","Sí": 2 1 2 2 2 2 1 2 2 1 ...
 $ retin     : Factor w/ 3 levels "Grave","Leve",..: 3 3 3 2 3 3 3 3 2 2 ...
 $ nefro     : Factor w/ 3 levels "Grave","Leve",..: 3 3 3 3 3 3 3 2 2 3 ...
 $ neuro     : Factor w/ 3 levels "Grave","Leve",..: 3 3 3 1 3 3 3 3 3 3 ...
 $ hba1c     : num  4.6 6.1 6.7 8.2 9.7 8.4 10.1 10.3 5.4 8.7 ...
 $ ca        : num  9.6 9.5 9.9 8.3 9.5 9.9 9.8 9.2 9.6 8.9 ...
 $ p         : num  4.2 3.4 4.7 3.2 3.7 4.7 4.7 4.2 2.6 3.9 ...
 $ cr        : num  1.02 1.07 1.35 0.99 0.86 0.8 1 0.7 0.92 1.08 ...
 $ pthm      : num  2.9 3.1 NA NA 3.1 1.1 3.4 NA 0.7 2.6 ...
 $ pthi      : num  49 NA 54.3 NA 49.8 59.9 20.9 NA NA 10 ...
 $ bmdcue    : num  22 NA 23.7 NA 35.7 38.3 6.2 NA NA 10.3 ...
 $ szl24     : num  1.07 0.83 0.83 0.72 0.56 0.92 1 0.83 0.72 0.69 ...
 $ sztri     : num  0.55 -1.62 -1.28 -1.11 -1.68 2.05 -0.75 -1.11 -0.98 -
0.91 ...
 $ szcue     : num  1.44 -0.93 -0.07 -2.07 -3.1 0.35 0.08 -1.62 -2.25 -
1.41 ...
 $ osteo_cue : Factor w/ 2 levels "No","Sí": 1 1 1 2 2 1 1 1 2 1 ...
 $ osteo_tri : Factor w/ 2 levels "No","Sí": 1 1 1 1 1 1 1 1 1 1 ...
```

Las 27 variables cubren factores de riesgo (tabaco, alcohol, actividad física), biomarcadores (HbA1c, calcio, creatinina, PTH), complicaciones (retinopatía, nefropatía, neuropatía) y medidas óseas (z-scores DXA, osteoporosis del cuello femoral y trocánter).

B.2. B.2 Análisis descriptivo: `freq()` y `grps()`

B.2.1. B.2.1 Tablas de frecuencias: `freq()`

Cuadro B.3 Signatura de freq()

```
freq(x, acum = TRUE, cuts = 0, agrup = TRUE, decs = 3, grf = TRUE)
```

Genera una tabla de frecuencias absolutas, relativas y acumuladas. Si `cuts > 0`, agrupa en intervalos de igual amplitud. El argumento `grf = FALSE` suprime el gráfico.

Ejemplo B.1: Distribución de la edad

Cuadro B.4 freq() sobre la edad del dataset osteo

```
freq(osteo$edad, grf = FALSE)
```

```
Distribución de frecuencias
```

```
--------------------------------
Variable:  osteo$edad
n = 94

        x Freq  Prop Prop.Acum
1 (18,23]   27 0.287     0.287
2 (23,28]   22 0.234     0.521
3 (28,33]   14 0.149     0.670
4 (33,38]   15 0.160     0.830
5 (38,43]    4 0.043     0.873
6 (43,48]    5 0.053     0.926
7 (48,53]    3 0.032     0.958
8 (53,58]    3 0.032     0.990
```

La distribución es **asimétrica positiva**: más de la mitad de los pacientes son jóvenes en estadios iniciales de la diabetes.

Cuadro B.5 freq() con 5 clases de igual amplitud para el IMC

```
freq(osteo$imc, cuts = 5, grf = FALSE)
```

```
Distribución de frecuencias
--------------------------------
Variable:  osteo$imc
n = 94

        x Freq  Prop Prop.Acum
1 (18,22]   37 0.394     0.394
2 (22,26]   36 0.383     0.777
3 (26,30]   13 0.138     0.915
4 (30,34]    5 0.053     0.968
5 (34,38]    1 0.011     0.979
```

B.2.2. B.2.2 Estadísticas por grupos: grps()

Cuadro B.6 Signatura de grps()

```
grps(x, f, ic = FALSE, grf = TRUE, alfa = 0.05, conf = 0.95, decs = 3)
```

Proporciona media, desviación típica y tamaño de cada grupo. Con `ic = TRUE` añade intervalos de confianza por grupo.

Ejemplo B.2: IMC por sexo

Cuadro B.7 IMC según sexo en el dataset osteo

```
grps(osteo$imc, osteo$sexo, grf = FALSE)
```

```
        n  media    dt
Hombre 45 23.514 2.890
Mujer  49 24.294 4.389
```

Cuadro B.8 IMC según sexo con intervalos de confianza

```
grps(osteo$imc, osteo$sexo, ic = TRUE, grf = FALSE)
```

```
        n  media    dt ic_inf ic_sup
Hombre 45 23.514 2.890 22.646 24.382
Mujer  49 24.294 4.389 23.033 25.555
```

B.3. B.3 Intervalos de confianza: `icm()`, `icp()`, `icl()`

B.3.1. B.3.1 IC para la media: `icm()`

Cuadro B.9 Signatura de icm()

```
icm(x = 0, n = 0, m = 0, s = 0, conf = 0.95, alfa = 0.05, decs = 3)
```

Intervalo de confianza bilateral para la media de una distribución normal. Acepta los estadísticos suficientes (n, $\bar{x}$, s) o directamente los datos (`x`).

Ejemplo B.3: IC para la media de HbA1c

Cuadro B.10 IC 95 % para la media poblacional de HbA1c

```
icm(m = mean(osteo$hba1c), s = sd(osteo$hba1c), n = nrow(osteo))
```

```
Intervalo de confianza bilateral para la media de una VA normal
----------------------------------------------------------------
Información muestral:
  Tamaño muestral: n =  94
  Media: m =  8.565
  Desviación típica: s =  1.799
  Error estándar de la media: sem =  0.186
```

```
Estimación:
 95%-IC(μ):  (8.196, 8.933)
  Precisión obtenida: 0.368
```

Interpretación: Con 95 % de confianza, la media poblacional de HbA1c en diabéticos de la UGR está aproximadamente entre 8.2 % y 8.9 %, lo que indica un control glucémico deficiente (meta clínica: HbA1c < 7 %).

B.3.2. B.3.2 IC para proporción: `icp()`

Cuadro B.11 Signatura de icp()

```
icp(x = 0, n = 0, conf = 0.95, alfa = 0.05, decs = 4)
```

Calcula el IC para una proporción con varios métodos: Clopper–Pearson (exacto), Wilson, Wald clásico y Agresti–Coull.

Ejemplo B.4: IC para la prevalencia de tabaquismo

Cuadro B.12 IC 95 % para la prevalencia de tabaquismo

```
icp(x = sum(osteo$tabaco == "Sí"), n = nrow(osteo))
```

```
Intervalo de confianza para una proporción binomial
---------------------------------------------------

Información muestral:
  Tamaño de muestra: n = 94
  Estimación puntual clásica: p=x/n = 0.4362, q=(1-p)=0.5638
  Casos observados: x = 41

# Método exacto (Clooper-Pearson):
  Pseudo-estimación puntual: p' = 0.4405, q'=(1-p')=0.5595
  95%-IC( ): (0.3385, 0.5424)
  Semiamplitud: 0.1019

# Método de Wilson (con cpc):
  Pseudo-estimación puntual: p' = 0.4388, q'=(1-p')=0.5612
  95%-IC( ): (0.3354, 0.5422)
  Semiamplitud: 0.1034

# Método de Wald (con cpc):
```

```
  Estimación puntual (clásica): p=x/n = 0.4362, q=(1-p)=0.5638
  95%-IC( ): (0.3306, 0.5417)
  Precisión: 0.1056

# Método de Wald ajustado (Agresti-Coull):
  Estimación puntual: p=(x+2)/(n+4) = 0.4388, q=(1-p)=0.5612
  95%-IC( ): (0.3405, 0.537)
  Precisión: 0.0982
```

B.3.3. B.3.3 IC para tasa Poisson: `icl()`

Cuadro B.13 Signatura de icl()

```
icl(x = 0, n = 0, conf = 0.95, decs = 4)
```

IC para la tasa λ de una distribución de Poisson basado en el método exacto (gamma–Poisson). Útil en vigilancia epidemiológica para tasas de incidencia por persona-año.

Ejemplo B.5: Tasa de incidencia de complicaciones diabéticas

En un registro hospitalario se observan 12 nuevos casos de retinopatía diabética en 1000 personas-año de seguimiento. Estimamos la tasa λ con un IC al 95 % pasando el **recuento de eventos** como un único valor observado (`n` se deja en su valor por defecto, $n = 1$, que indica que los 12 eventos provienen de una única unidad de seguimiento):

Cuadro B.14 icl() — IC exacto para una tasa de incidencia

```
icl(x = 12)
```

```
Intervalo de confianza bilateral para el parámetro    de una VA con distribución de Poisson
-------------------------------------------------------------------------------------------
Información muestral:
  Muestra de una sola observación
  Tamaño muestral: n =  1
  Media observada: m =  12

Estimación:
  [1] Método exacto:
      95 %-IC( ):  ( 6.2006 ,  20.9616 )
      Semiamplitud del intervalo: 7.3805

  [2] Aproximación a la normal (transformación de la raiz):
      Validez de la aproximación: Σx =  12  < 15 --- NO es válida ---
```

```
95 %-IC():  ( 6.1709 ,  21.0271 )
Precisión obtenida: 7.4281
```

Interpretación: la tasa estimada es de 12 casos por 1000 personas-año (IC 95 %: aproximadamente 6.2 – 21.0 por 1000 p-año, método exacto gamma–Poisson). El método exacto es preferible al de la aproximación normal cuando el número de eventos es pequeño.

> **Nota sobre los argumentos:** si se llama `icl(x = 12, n = 1000)` la función interpreta `x` como la **media muestral observada en `n` unidades** ($\sum x_i = n \cdot x = 12\,000$ eventos en 1000 unidades), no como "12 eventos en 1000 personas-año". La salida en ese caso sería un IC muy estrecho alrededor de 12 ($\approx$ [11.79, 12.22]).

B.4. B.4 Tamaño muestral: `nm()`, `np()` y `nl()`

B.4.1. B.4.1 Para la media: `nm()`

Cuadro B.15 Signatura de nm()

```
nm(d, n, m, s, alfa = 0.05)
```

Calcula el tamaño muestral necesario para estimar una media con precisión d (semiamplitud del IC) dada una desviación típica esperada s.

Ejemplo B.6: ¿Cuántos pacientes para estimar HbA1c con mayor precisión?

Cuadro B.16 Tamaño muestral para estimar HbA1c con d = 0.3

```
nm(d = 0.3, n = 94, m = mean(osteo$hba1c), s = sd(osteo$hba1c), alfa = 0.05)
```

```
# Tamaño de muestra para la estimación de la media de una VA normal o su aproximación
# ------------------------------------------------------------------------------

# Muestra piloto:
  Tamaño muestral:  n = 94
  Media: m = 8.5649
  Desviación típica: s = 1.7987
  Error estandar de la media: sem = 0.1855
  Precisión observada: d = 0.3684

# Estimación del tamaño muestral:
  Precisión deseada:   = 0.3000
  Tamaño muestral necesario: n  142
```

B.4.2. B.4.2 Para proporciones: np()

Cuadro B.17 Signatura de np()

```
np(x, n, d, conf = 0.95, decs = 5)
```

Tamaño muestral para estimar una proporción con precisión d, usando la estimación piloto $\hat{p} = x/n$.

B.4.3. B.4.3 Para tasa Poisson: nl()

Cuadro B.18 Signatura de nl()

```
nl(x, n = 0, d, lmax = 0, conf = 0.95, alfa = 0.05)
```

Tamaño muestral para estimar el parámetro λ de una distribución de Poisson con la precisión deseada. Se puede informar a partir de observaciones piloto o del valor máximo esperado del parámetro (`lmax`).

Ejemplo B.7: Diseño de un estudio de vigilancia epidemiológica

Se planea un estudio para estimar la tasa de incidencia de una infección hospitalaria con una **precisión de ± 1 caso por persona-año**. Un estudio piloto previo observó 3 casos en una sola unidad de seguimiento ($n = 1$, media observada $\bar{x} = 3$).

Cuadro B.19 nl() — tamaño muestral a partir de observación piloto

```
nl(x = 3, d = 1)
```

```
Tamaño de muestra necesario para estimar el parámetro
de una VA con distribución de Poisson con precisión
------------------------------------------------------------------------
Muestra piloto:
  Muestra de una sola observación
  Tamaño muestral: n =  1
  Media observada: m =  3

  Estimación considerando la información muestral:
     95 %-max( ) =  7.754  (método exacto)
     Precisión deseada:   =  1
     Tamano muestral sin cpc: n    30
     Tamano muestral con cpc: n    31
```

Interpretación: se requieren al menos $n \approx 31$ unidades de seguimiento (con corrección por continuidad) para conseguir la precisión deseada al 95 % de confianza.

Si en lugar de datos piloto se dispone de una **cota superior esperada** del parámetro (p. ej. $\lambda \leq 4.5$), se utiliza el argumento `lmax`:

Cuadro B.20 nl() — tamaño muestral a partir de un valor máximo esperado

```
nl(lmax = 4.5, d = 1)
```

```
Tamaño de muestra necesario para estimar el parámetro
de una VA con distribución de Poisson con precisión
------------------------------------------------------------------------
Estimación con el valor máximo propuesto para el parámetro:
  Valor máximo propuesto:   =  4.5
  Precisión deseada:   =  1
  Tamano muestral sin cpc: n    18
  Tamano muestral con cpc: n    19
```

B.5. B.5 Contrastes de hipótesis

B.5.1. B.5.1 Test de normalidad: `testnormal()`

Cuadro B.21 Signatura de testnormal()

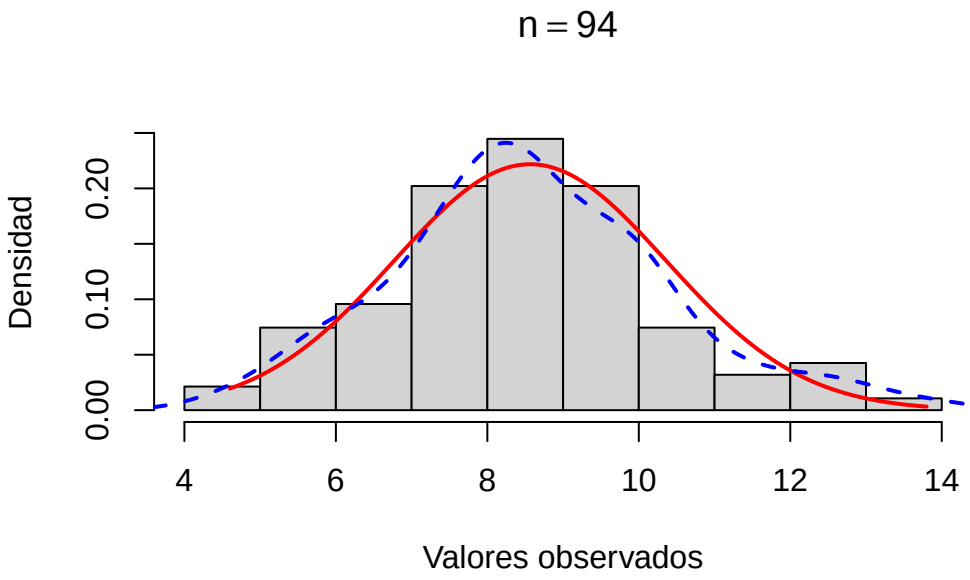
```
testnormal(x, obs = TRUE, mod = TRUE, dens = TRUE, sw = TRUE, decs = 3, grf = TRUE)
```

Contraste de Shapiro–Wilk con visualización integrada (histograma + densidad teórica + Q–Q plot).

Ejemplo B.8: ¿Es normal la HbA1c?

Cuadro B.22 Test de normalidad para HbA1c

```
testnormal(osteo$hba1c, grf = FALSE)
```




```
# Test de normalidad de Shapiro-Wilk
 -----------------------------------
   n = 94,  W = 0.983,  p = 0.276
```

En esta muestra la HbA1c es compatible con una distribución aproximadamente normal — contrasta con la edad, que es claramente asimétrica.

B.5.2. B.5.2 Test t: `testt()`

Cuadro B.23 Signatura de testt()

```
testt(m, s, n, m0, delta, potencia)
```

Integra en una sola función: verificación de normalidad (Shapiro–Wilk), t-test de una o dos muestras (pareadas o independientes), test de homogeneidad de varianzas (Welch automático) e IC.

Ejemplo B.9a: Contraste de una muestra — ¿HbA1c > 7.5 %?

Cuadro B.24 testt() de una muestra: HbA1c contra el umbral clínico

```
# H0: mu = 7.5  vs  H1: mu > 7.5
testt(m = osteo$hba1c, m0 = 7.5, grf = FALSE)
```

```
# t-Test con una muestra
# ----------------------

# Resumen de 'osteo$hba1c'
    n = 94.000
    media = 8.565
```

```
    d.t. = 1.799
    sem = 0.186

# Estimación de la media  :
  95%-IC( ) = (8.196, 8.933)

# Test de normalidad de Shapiro-Wilk:
  W = 0.983, gl = 94, p = 0.276

# Test de Student para contrastar H : =  con  =7.500
  texp = 5.740, gl = 93
    p < 0.001 para la alternativa bilateral H :
    p < 0.001 para la alternativa unilateral H : >

  Estimación del efecto bruto
  95%-IC( - ) = (0.696, 1.433)
```

Rechazamos H_0: la media de HbA1c supera significativamente el umbral clínico de 7.5 %.

Ejemplo B.9b: Contraste de dos muestras — HbA1c entre hombres y mujeres

Cuadro B.25 testt() de dos muestras independientes

```
testt(grupos = osteo$sexo, m = osteo$hba1c, grf = FALSE)
```

```
# t-test para 2 Muestras Independientes
# -------------------------------------

# Información muestral y estimación de las medias
  Niveles de agrupación: Hombre, Mujer

                        n media    dt   sem             IC
  osteo$hba1c [Hombre] 45 8.553 1.745 0.260 (8.029, 9.078)
  osteo$hba1c [Mujer]  49 8.576 1.864 0.266  (8.04, 9.111)
  ----
  * IC elaborados al 95% de confianza para estimar   y   respectivamente


# Pruebas de normalidad (test de Shapiro-Wilk)
[1] Para grupo = Hombre, W = 0.958, gl = 45, p = 0.102
```

```
[2] Para grupo = Mujer, W = 0.980, gl = 49, p = 0.566

# Test de homogeneidad de varianzas. Fexp = (var /var )
  Fexp = 1.141, gl  = 48, gl  = 44, p = 0.660

# Diferencia de medias (osteo$hba1c [Mujer] - osteo$hba1c [Hombre])
  Hipótesis a contrastar: H : =  ( - =0)

  a) Test de Student (varianzas homogéneas)
  texp = 0.059, gl = 92
    p = 0.953 para la alternativa bilateral H :
    p = 0.476 para la alternativa unilateral H : <
  95%-IC( - ) = (-0.719, 0.764)

  b) Test de Welch (varianzas no homogéneas)
  texp = 0.060, gl = 91.96
    p = 0.953 para la alternativa bilateral H :
    p = 0.476 para la alternativa unilateral H : <
  95%-IC( - ) = (-0.717, 0.762)
```

B.5.3. B.5.3 Test F de varianzas: `testf()`

Cuadro B.26 Signatura de testf()

```
testf(s1, n1, s2, n2)
```

Contraste $H_0 : \sigma_1^2 = \sigma_2^2$ usando el estadístico $F = s_1^2/s_2^2 \sim F_{n_1-1,\, n_2-1}$.

Ejemplo B.10: ¿Misma variabilidad del IMC en hombres y mujeres?

Cuadro B.27 Test F de varianzas del IMC por sexo

```
testf(s1 = sd(osteo$imc[osteo$sexo == "Hombre"]),
      n1 = sum(osteo$sexo == "Hombre"),
      s2 = sd(osteo$imc[osteo$sexo == "Mujer"]),
      n2 = sum(osteo$sexo == "Mujer"))
```

```
[[1]]
[1] 2.306145

[[2]]
[1] 48
```

```
[[3]]
[1] 44

[[4]]
[1] 0.00584978
```

B.5.4. B.5.4 Test de proporciones: testp()

Cuadro B.28 Signatura de testp()

```
testp(x, n, p0, delta = 0.05, potencia = 0.9)
```

Contraste para una proporción (contra un valor hipotético π_0) o para comparar dos proporciones independientes o pareadas.

Ejemplo B.11: ¿La prevalencia de tabaquismo supera el 35 %?

Cuadro B.29 testp() de una proporción

```
testp(x = sum(osteo$tabaco == "Sí"), n = nrow(osteo), p0 = 0.35)
```

```
# Test para contrastar una proporción binomial
# ---------------------------------------------

# Información muestral
  n = 94
  x = 41   n-x=53
  p = 0.436; q = (1-p) = 0.564

# Test Ho: =0.350
  [1] Método exacto
                      H1  Fexp Valor.p
    Cola derecha  >0.350 1.410   0.052
    Bilateral     0.350     -   0.103

    95%-IC( ) = (0.339, 0.542) (método de Clooper-Pearson)

  [2] Método aproximado a la distribución normal
    Validez: min(n , n(1- )) = 32.9 (>5, el método es válido)
    zexp = 1.643,  p  = 0.100
```

```
    95%-IC( ) = (0.335, 0.542) (método de Wilson)
```

B.5.5. B.5.5 Test de Wilcoxon: testwx()

Cuadro B.30 Signatura de testwx()

```
testwx(m1, m2, par = FALSE)
```

Test de Wilcoxon–Mann–Whitney para dos muestras independientes (par = FALSE) o test de los rangos con signo de Wilcoxon para datos pareados (par = TRUE). Devuelve el tamaño del efecto r y la probabilidad de superioridad.

Ejemplo B.12: IMC en diabéticos jóvenes vs. mayores

Cuadro B.31 Wilcoxon–Mann–Whitney sobre IMC por grupo de edad

```
testwx(m1 = osteo$imc[osteo$grupo_edad == "< 25"],
       m2 = osteo$imc[osteo$grupo_edad == "> 33"],
       par = FALSE,
       grf = FALSE)
```

```
Test de Wilcoxon/Mann-Whithney para dos muestras independientes
-----------------------------------------------------------------
# Información muestral ---

                                Muestra  n    min     Q1     Q2     Q3    max
1 osteo$imc[osteo$grupo_edad == "< 25"] 32 19.628 21.421 22.845 24.149 32.475
2 osteo$imc[osteo$grupo_edad == "> 33"] 30 18.070 23.382 25.889 28.945 37.333
    RIQ
1 2.728
2 5.563

# Rangos ---

                                Muestra  n Suma_rangos Rango_medio       U
1 osteo$imc[osteo$grupo_edad == "< 25"] 32         807      25.219 681.000
2 osteo$imc[osteo$grupo_edad == "> 33"] 30        1146      38.200 279.000

# Test ---

  U = 279.000; Z = 2.831; W = 279.000; p = 0.004
```

```
# Tamaño del efecto ---

  Diferencia de localización: -2.668    95%-IC = (-4.518, -0.778)
  r = 0.360  (criterio: 0.1 pequeño; 0.3 mediano; >0.5 grande)
  Probabilidad de superioridad PS = 0.709
  (probabilidad de que un valor al azar de M1 sea < a un valor al azar de M2)
```

B.5.6. B.5.6 Test de McNemar: `testmcnemar()`

Cuadro B.32 Signatura de testmcnemar()

```
testmcnemar(n11, n12, n21, n22, delta = 0.05, alfa = 0.05, beta = 0.15)
```

Para comparar dos proporciones en muestras apareadas (antes/después, o dos pruebas diagnósticas en los mismos sujetos). El contraste se basa en las celdas discordantes n_{12} y n_{21}.

Ejemplo B.13: Cambio en actividad física tras una intervención

Cuadro B.33 testmcnemar() sobre datos apareados antes/después

```
# n11 = activo->activo, n12 = activo->inactivo
# n21 = inactivo->activo, n22 = inactivo->inactivo
testmcnemar(n11 = 35, n12 = 10, n21 = 5, n22 = 44,
            fcat = c("Activo", "Inactivo"),
            ccat = c("Antes", "Después"))
```

```
# Inferencia con dos proporciones (muestras apareadas)
# ----------------------------------------------------

# Frecuencias observadas pretest x posttest
               Antes    Después Total
   Activo         35         10    45
   Inactivo        5         44    49
   Total          40         54    94


# Proporciones observadas pretest x posttest
               Antes    Después  Total
   Activo     0.3723     0.1064 0.4787
   Inactivo   0.0532     0.4681 0.5213
   Total      0.4255     0.5745 1.0000
```

```
# Test de McNemar: H :  =
   Validez: n +n  = 15 > 10 el test es válido
   Zexp =  1.1619
               valor.p Alternativa
    Bilateral   0.2453  H :
    Unilateral  0.1226  H :  >


# Test exacto de Fisher:
  H :  =0.5 para n  ~ B(n +n ,  )
               Valor.p Alternativa
    Bilateral   0.3018  H :  0.5
    Unilateral  0.1509  H :  >0.5
  ----
  * Aquí se alude a la probabilidad total de la discordancia, es decir que  +  =1

# Estimación de las proporciones individuales de discordancias   y   (método de Wald ajusta
   [1] p  = 0.1224, 95%-IC( ) = (0.0575, 0.1873)
   [2] p  = 0.0714, 95%-IC( ) = (0.0204, 0.1224)


# Intervalo de confianza para la diferencia de 2 proporciones apareadas
   [1] Método de Wald (clásico con cpc):
      Estimación puntual de  -  = 0.0532
      Validez: n +n  = 15 > 5, el IC es válido
      95%-IC( - ) = (-0.0282, 0.1346)

   [2] Método de Agresti-Min:
      Estimación puntual de  -  = 0.0521
      Validez: siempre es válido
      95%-IC( - ) = (-0.0289, 0.1331)
```

B.6. B.6 Regresión lineal: `rls()` y `rlm()`

B.6.1. B.6.1 Regresión lineal simple: `rls()`

Cuadro B.34 Signatura de rls()

```
rls(f, data, pred = NULL, grf = TRUE, alfa = 0.05, decs = 3)
```

Ajusta el modelo $Y = \beta_0 + \beta_1 X + \varepsilon$ e informa de: correlación de Pearson con IC, coeficientes con IC, R^2, diagnóstico de residuos y (opcionalmente) predicciones puntuales con IC.

Ejemplo B.14: ¿La evolución de la diabetes predice la HbA1c?

Cuadro B.35 rls() — regresión lineal simple de HbA1c sobre tevol

```
rls(hba1c ~ tevol, data = osteo, grf = FALSE)
```

```
Regresión lineal simple
---------------------------------------------------------------
# Información muestral ---

  variable  n  media    dt Min  Max Rango
1    hba1c 94  8.565 1.799 4.6 13.8   9.2
2    tevol 94 12.330 8.534 0.0 35.0  35.0

  Cov(hba1c,tevol) = -3.64

# Correlación de Pearson ---

      r IC_inf IC_sup gl   texp     sig
  -0.237  -0.42 -0.036 92 -2.341 = 0.021

# Modelo lineal ---

  Modelo:  hba1c ~ tevol
  R² =  0.056
  S²residual =  3.087
  :

         Coef  estim    se ic_inf ic_sup   texp    sig
1 (Constante)  9.181 0.320  8.546  9.816 28.730 <0.001
2       tevol -0.050 0.021 -0.092 -0.008  2.341  0.021

# Distribución residual ---
  Error estándar residual: 1.757
      res   zres
min -4.481 -2.585
Q1  -0.981 -0.562
Q2  -0.056 -0.032
```

```
Q3   0.956  0.551
max  4.669  2.698

  Test de normalidad residual (Shapiro-Wilk):
  w =0.989, p= 0.625
```

B.6.2. B.6.2 Regresión lineal múltiple: `rlm()`

Cuadro B.36 Signatura de rlm()

```
rlm(f, data, pred = NULL, grf = TRUE, dfout = FALSE, alfa = 0.05, decs = 3)
```

Extiende `rls()` al caso multivariante: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$. Devuelve la tabla de coeficientes con IC, R^2 y R^2 ajustado, distribución de los residuos con su test de normalidad y, si `grf = TRUE`, los cuatro gráficos clásicos de diagnóstico más histograma de residuos estandarizados.

Ejemplo B.15: HbA1c ~ tevol + edad + IMC

Cuadro B.37 rlm() — regresión lineal múltiple sobre el dataset osteo

```
rlm(hba1c ~ tevol + edad + imc, data = osteo, grf = FALSE)
```

```
Regresión lineal múltiple
-----------------------------------------------------------------
# Información muestral ---

      Variable  n  Media    DT   Min    Max
hba1c    hba1c 94  8.565 1.799  4.60 13.800
tevol    tevol 94 12.330 8.534  0.00 35.000
edad      edad 94 30.191 9.366 18.00 56.000
imc        imc 94 23.921 3.748 18.07 37.333

# Modelo lineal ---

   Modelo :  hba1c ~ tevol + edad + imc
  R² =  0.063  (R²  ajustado  =  0.031 )
  S²residual =  3.134

   Coeficientes del modelo :

      Termino Estimacion Error_Std IC_inf IC_sup  t_exp      sig
1 (Intercept)      8.601     1.200  6.217 10.984  7.169  < 0.001
```

```
2       tevol     -0.046     0.025 -0.095  0.003 -1.883  = 0.063
3        edad     -0.012     0.024 -0.059  0.035 -0.521  = 0.604
4         imc      0.038     0.052 -0.066  0.142  0.722  = 0.472

# Distribución residual ---
  Error estándar residual:  1.77
    Residuos Res_Est
min   -4.519  -2.589
Q1    -0.945  -0.545
Q2    -0.010  -0.006
Q3     0.995   0.572
max    4.938   2.923

  Test de normalidad residual (Shapiro-Wilk):
   w = 0.991 ,    = 0.744
```

Tras ajustar simultáneamente por edad e IMC, el efecto del tiempo de evolución sobre la HbA1c se atenúa, sugiriendo confusión parcial. El R^2 ajustado es modesto: estas tres variables explican una fracción pequeña de la variabilidad del control glucémico (ver discusión completa en Capítulo 10).

Predicciones con el modelo

`rlm()` acepta un `data.frame` en el argumento `pred` para devolver predicciones puntuales con intervalos de confianza (de la media) y de predicción (de una nueva observación):

Cuadro B.38 rlm() con predicciones para pacientes hipotéticos

```
nuevos <- data.frame(tevol = c(5, 15), edad = c(30, 45), imc = c(25, 28))
rlm(hba1c ~ tevol + edad + imc, data = osteo, pred = nuevos, grf = FALSE)
```

```
Regresión lineal múltiple
-----------------------------------------------------------------
# Información muestral ---

      Variable  n  Media    DT   Min    Max
hba1c    hba1c 94  8.565 1.799  4.60 13.800
tevol    tevol 94 12.330 8.534  0.00 35.000
edad      edad 94 30.191 9.366 18.00 56.000
imc        imc 94 23.921 3.748 18.07 37.333

# Modelo lineal ---
```

```
   Modelo :  hba1c ~ tevol + edad + imc
  R² =  0.063  (R²  ajustado  =  0.031 )
  S²residual =  3.134

   Coeficientes del modelo :

      Termino Estimacion Error_Std IC_inf IC_sup  t_exp      sig
1 (Intercept)      8.601     1.200  6.217 10.984  7.169  < 0.001
2       tevol     -0.046     0.025 -0.095  0.003 -1.883  = 0.063
3        edad     -0.012     0.024 -0.059  0.035 -0.521  = 0.604
4         imc      0.038     0.052 -0.066  0.142  0.722  = 0.472

# Pronósticos con el modelo ---
  Pronosticos puntuales y bandas al 95 % de confianza para
  promedios IC(m), y para una nueva observación: IC(obs)

  tevol edad imc  Puntual IC_m_inf IC_m_sup IC_obs_inf IC_obs_sup
1     5   30  25 8.947237 8.428009 9.466464   5.392127   12.50235
2    15   45  28 8.413571 7.678597 9.148545   4.820606   12.00654

# Distribución residual ---
  Error estándar residual:  1.77
    Residuos Res_Est
min   -4.519  -2.589
Q1    -0.945  -0.545
Q2    -0.010  -0.006
Q3     0.995   0.572
max    4.938   2.923

  Test de normalidad residual (Shapiro-Wilk):
   w = 0.991 ,   = 0.744
```

B.7. B.7 Regresión logística: `rlogits()` y `rlogitm()`

B.7.1. B.7.1 Regresión logística simple: `rlogits()`

Cuadro B.39 Signatura de rlogits()

```
rlogits(f, data, grf = FALSE, alfa = 0.05, decs = 3)
```

Ajusta un modelo logístico binario simple $\text{logit}\,\pi = \beta_0 + \beta_1 X$ y devuelve coeficientes, odds ratio con

IC, devianzas, AIC, pseudo-R^2 de Nagelkerke, test de Hosmer–Lemeshow y AUC.

Ejemplo B.16: ¿El IMC predice la osteoporosis del cuello femoral?

Cuadro B.40 rlogits() — regresión logística simple sobre osteoporosis

```
rlogits(osteo_cue ~ imc, data = osteo)
```

```
Regresión logística  simple
----------------------------------------------------------------
# Información muestral ---

   Tamaño muestral (N inicial) :  94
   Tamaño muestral tras eliminar valores perdidos (Casos completos) :  94
   Mínima frecuencia de eventos (n efectivo) :  24

# Distribución de la variable respuesta (osteo_cue) ---

  Categoria  n Porcentaje
1        No 70     74.468
2        Sí 24     25.532

# Modelo logístico ---  ---

   Modelo :  osteo_cue ~ imc
  Devianza residual:  100.176  (Nula:  106.804 )
  AIC:  104.176
  R² de Nagelkerke:  0.1

   Test de bondad de ajuste de Hosmer-Lemeshow :
  X² =  7.041 , gl =  8 ,   = 0.532

   Capacidad discriminante :
   AUC (Area bajo la curva ROC)  =  0.649

   Coeficientes del modelo :

      Termino Estimacion Error_Std  z_exp      sig     OR OR_inf   OR_sup
1 (Intercept)      3.620     2.044  1.771  = 0.077 37.348  0.937 3055.247
2         imc     -0.202     0.089 -2.256  = 0.024  0.817  0.672    0.957
```

B.7.2. B.7.2 Regresión logística múltiple: `rlogitm()`

Cuadro B.41 Signatura de rlogitm()

```
rlogitm(f, data, pred = NULL, grf = FALSE, alfa = 0.05, decs = 3)
```

Extiende `rlogits()` al caso multivariante. Devuelve la tabla de coeficientes con OR ajustados e IC, devianzas, AIC, pseudo-R^2 de Nagelkerke, test de Hosmer–Lemeshow, AUC y, opcionalmente, la curva ROC (con `grf = TRUE`).

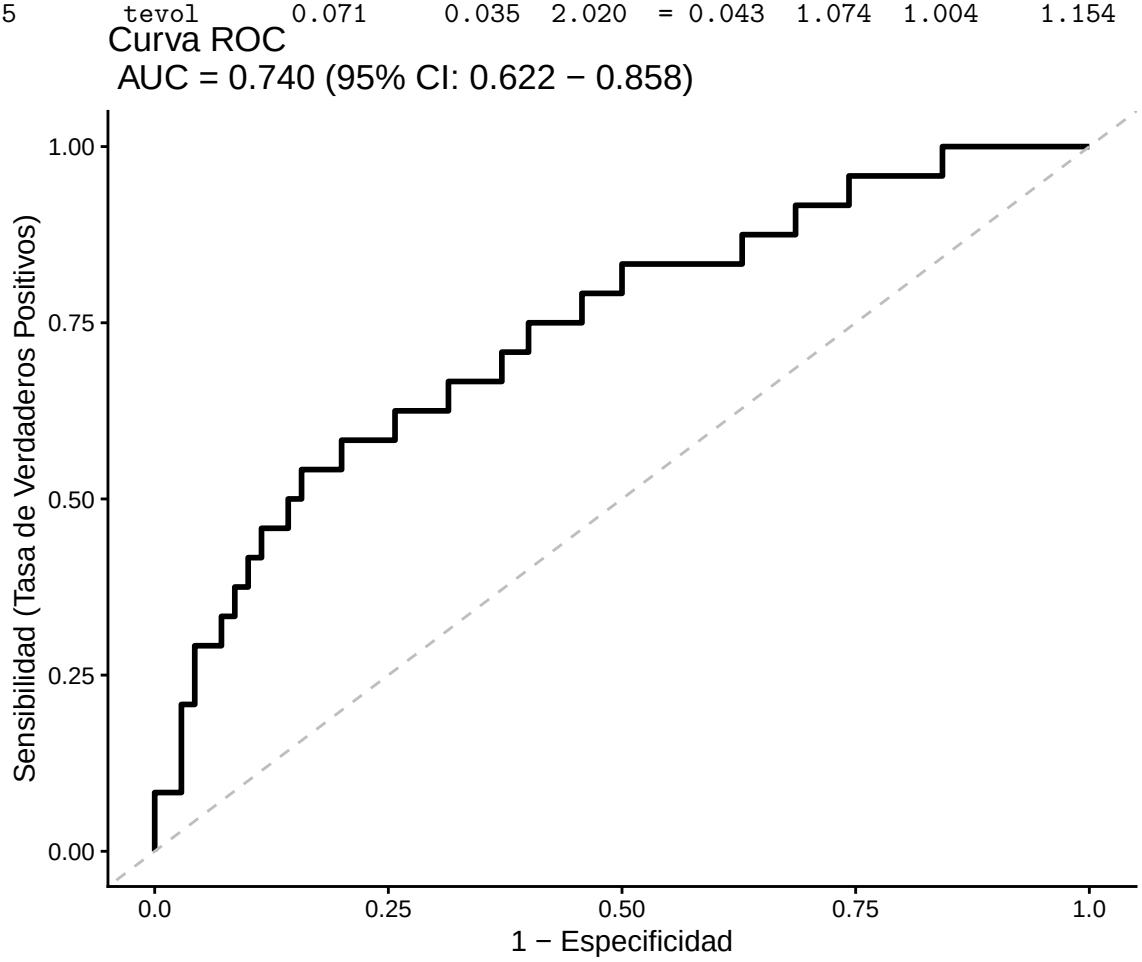
Ejemplo B.17: Osteoporosis ajustada por tabaco + edad + IMC + tevol

Cuadro B.42 rlogitm() — regresión logística múltiple con curva ROC

```
rlogitm(osteo_cue ~ tabaco + edad + imc + tevol,
        data = osteo,
        grf  = TRUE)
```

```
Regresión logística  multiple
----------------------------------------------------------------
# Información muestral ---

   Tamaño muestral (N inicial) :  94
   Tamaño muestral tras eliminar valores perdidos (Casos completos) :  94
   Mínima frecuencia de eventos (n efectivo) :  24

# Distribución de la variable respuesta (osteo_cue) ---

  Categoria  n Porcentaje
1        No 70     74.468
2        Sí 24     25.532

# Modelo logístico ---  ---

   Modelo :  osteo_cue ~ tabaco + edad + imc + tevol
  Devianza residual:  92.03  (Nula:  106.804 )
  AIC:  102.03
  R² de Nagelkerke:  0.214

   Test de bondad de ajuste de Hosmer-Lemeshow :
  X² =  2.067 , gl =  8 ,   = 0.979

   Capacidad discriminante :
```

```
   AUC (Area bajo la curva ROC)  =  0.74

   Coeficientes del modelo :

      Termino Estimacion Error_Std  z_exp      sig     OR OR_inf   OR_sup
1 (Intercept)      2.893     2.269  1.275  = 0.202 18.045  0.275 2237.565
2    tabacoSí      0.865     0.549  1.575  = 0.115  2.376  0.821    7.218
3        edad     -0.004     0.037 -0.105  = 0.917  0.996  0.925    1.072
4         imc     -0.224     0.105 -2.140  = 0.032  0.799  0.638    0.965
5       tevol      0.071     0.035  2.020  = 0.043  1.074  1.004    1.154
```



La interpretación epidemiológica completa de este modelo se desarrolla en Sección 11.8.3: los OR ajustados, el test de Hosmer–Lemeshow, el pseudo-R^2 de Nagelkerke y el AUC se combinan para evaluar simultáneamente la asociación independiente de cada predictor y la capacidad discriminante global del modelo.

B.8. B.8 Tablas de contingencia: tabla2x2() y tablarxc()

B.8.1. B.8.1 Tabla 2×2: tabla2x2()

Cuadro B.43 Signatura de tabla2x2()

```
tabla2x2(o, estudio = "T", fcat, ccat, tablas = c("F", "E", "S"))
```

Calcula: frecuencias observadas y esperadas, χ^2 de Pearson (con y sin corrección de Yates), test exacto de Fisher y medidas de asociación (OR, RR, RD) con sus IC al 95 %, adaptadas al tipo de estudio (`estudio = "T"` transversal, `"C"` cohorte, `"CC"` caso-control, `"AP"` apareado).

Ejemplo B.18: Tabaquismo y osteoporosis del cuello femoral

Cuadro B.44 tabla2x2() — tabaquismo y osteoporosis

```
tabla2x2(fvar = osteo$tabaco, cvar = osteo$osteo_cue, o = osteo)
```

```
# Análisis de tablas 2x2
# ----------------------

# Frecuencias observadas
          C1   C2 Total
  F1      44    9    53
  F2      26   15    41
  Total   70   24    94


# Test Chi-cuadrado para un estudio transversal

   ² = 4.662,   gl = 1,  p = 0.031, (cpc = 0.5)
  Validez: Frecuencia mínima esperada = 10.47  > 3.9

  Test exacto de Fisher (bilateral): p = 0.035

  --- Otros criterios  ²:
   ² = 4.673,   gl = 1,  p = 0.054, (sin cpc)
   ² = 3.699,   gl = 1,  p = 0.054, (cpc de Yates = 47.00)

# Estimación de la prevalencia  en un estudio transversal
  Método de Wald ajustado:
  p=0.561; 95%-IC( )=(0.463, 0.659)
```

```
# Medidas de asociación para un estudio transversal
  [!] Las medidas de riesgo se calculan como riesgo de la categoría
      en la 1a columna (frente a la 2a) para la categoría en la 1a
      fila (frente a la 2a)

  Riesgo absoluto (diferencia de Berkson; método de Agresti-Caffo):
  d=0.254; 95%-IC(d)=(0.023, 0.459)

  Riesgo relativo:
  Rr=1.676; 95%-IC(Rr)=(0.969, 2.808)

  Riesgo atribuible:
  Ra=0.335; 95%-IC(Ra)= (-0.049, 0.578)

  Razón del producto cruzado (odds ratio):
  OR=2.821; 95%-IC(OR)= (1.070, 7.013)
```

Los fumadores presentan una OR de aproximadamente 2.8 de tener osteoporosis del cuello femoral respecto a los no fumadores. La asociación cruda es estadísticamente significativa ($p \approx 0.03$). Esta asociación se reevalúa con ajuste multivariable en Sección 11.8.3.

B.8.2. B.8.2 Tabla r×c: `tablarxc()`

Cuadro B.45 Signatura de tablarxc()

```
tablarxc(frecs, tablas = c("E", "F", "S"), fcat, ccat)
```

Ejemplo B.19: Osteoporosis por grupo de edad (tabla 3×2)

Cuadro B.46 tablarxc() — osteoporosis del cuello femoral por grupo de edad

```
# Reordenamos el factor cronológicamente para que las filas de la
# tabla coincidan con el orden de fcat.
osteo$grupo_edad <- factor(osteo$grupo_edad,
                           levels = c("< 25", "25-33", "> 33"))
tab <- table(osteo$grupo_edad, osteo$osteo_cue)
tablarxc(frecs = tab,
         fcat  = c("< 25", "25-33", "> 33"),
         ccat  = c("No", "Sí"))
```

```
# Test Chi-cuadrado para tablas RxC
# ---------------------------------
```

```
# Frecuencias observadas
          No   Sí Total
  < 25    27    5    32
  25-33    0    0     0
  > 33    22    8    30
  Total   49   13    62

# Test chi-cuadrado
  Validez: Frecuencia mínima esperada = 0
           2 frecuencias esperadas son menores a 1
           0 son menores a 5 (el 0% de la tabla)
   [!] El test   ²  NO es válido con estos datos
No se calcula el estadístico de contraste
```

B.9. B.9 Tabla resumen de funciones

Función	Capítulo	Propósito	Ejemplo de uso
`freq()`	1, 2	Tablas de frecuencias (absolutas, relativas, acumuladas)	`freq(osteo$edad, grf=FALSE)`
`grps()`	2	Estadísticas descriptivas por grupos	`grps(osteo$imc, osteo$sexo)`
`icm()`	7	IC bilateral para la media	`icm(m=8.5, s=1.8, n=94)`
`icp()`	7	IC para proporción (4 métodos)	`icp(x=41, n=94)`
`icl()`	7	IC para tasa Poisson (exacto)	`icl(x=12, n=1000)`
`nm()`	7	Tamaño muestral para media	`nm(m=8.5, s=1.8, d=0.4)`
`np()`	7	Tamaño muestral para proporción	`np(x=41, n=94, d=0.1)`
`nl()`	7	Tamaño muestral para tasa Poisson	`nl(x=3, d=1)`
`testnormal()`	9, 11	Test de Shapiro–Wilk + gráficos	`testnormal(osteo$hba1c)`
`testt()`	8, 9	t-test (1 muestra, 2 muestras, pareado)	`testt(m=osteo$hba1c, m0=7.5)`

Función	Capítulo	Propósito	Ejemplo de uso
`testf()`	9	Test F de igualdad de varianzas	`testf(s1, n1, s2, n2)`
`testp()`	8	Test de proporciones (1 o 2 grupos)	`testp(x=41, n=94, p0=0.35)`
`testwx()`	9	Wilcoxon–Mann–Whitney / rangos con signo	`testwx(m1=imc1, m2=imc2)`
`testmcnemar()`	11	Test de McNemar (datos apareados)	`testmcnemar(n11=35, n12=10, n21=5, n22=44)`
`rls()`	10	Regresión lineal simple con diagnósticos	`rls(hba1c ~ tevol, data=osteo)`
`rlm()`	10	Regresión lineal múltiple con diagnósticos	`rlm(hba1c ~ tevol + edad + imc, data=osteo)`
`rlogits()`	11	Regresión logística simple + AUC	`rlogits(osteo_cue ~ imc, data=osteo)`
`rlogitm()`	11	Regresión logística múltiple + Hosmer–Lemeshow + ROC	`rlogitm(osteo_cue ~ tabaco + edad + imc, data=osteo, grf=TRUE)`
`tabla2x2()`	11	Tabla 2×2: ^{2}, Fisher, OR/RR/RD + IC	`tabla2x2(fvar=tabaco, cvar=osteo_cue, o=osteo)`
`tablarxc()`	11	Tabla r×c: ^{2} con verificación de supuestos	`tablarxc(frecs=table(grupo_edad, osteo_cue))`

Recursos del paquete

- **Sitio web:** https://migariane.github.io/BioEstatR/
- **Repositorio GitHub:** https://github.com/migariane/BioEstatR
- **Viñeta introductoria:** https://migariane.github.io/BioEstatR/articles/BioEstatR-overview.html
- **Contacto:** `pfemia@ugr.es` (Pedro Femia Marzo)
- **Instalación:** `remotes::install_github("migariane/BioEstatR")`
- **Cita:** `citation("BioEstatR")`
- **Dataset `osteo`:** 94 pacientes diabéticos, 27 variables, Unidad de Diabetes, Facultad de Medicina, UGR.

Salidas en inglés

El paquete soporta también salida en inglés mediante:

```
options(BioEstatR.lang = "en")
rlm(hba1c ~ tevol + edad + imc, data = osteo, grf = FALSE)
options(BioEstatR.lang = "es")  # vuelve al castellano (valor por defecto)
```

Apéndice C

Apéndice C: Guía de R para Estadística Médica

Esta guía cubre los fundamentos de R orientados a la práctica de la bioestadística y la investigación médica: importación de datos clínicos, manejo de variables categóricas (exposición, enfermedad, grupo), descripción de cohortes, contraste de hipótesis y visualización. Complementa al Apéndice B (paquete `BioEstatR`) y al Apéndice A (repaso matemático).

> **Cómo usar este apéndice**
>
> El material está organizado como referencia rápida. Cada sección incluye ejemplos ejecutables con datos clínicos simulados y, donde procede, código no ejecutable (`eval: false`) que muestra la sintaxis para tareas habituales en investigación médica (importar SPSS, ajustar modelos de supervivencia, etc.).

C.1. C.1 Conceptos Fundamentales

C.1.1. Asignación, Tipos y Vectores

```
[1] "numeric"
```

```
[1] 5
```

```
[1] 22 60
```

C.1.2. Indexación: posicional y lógica

```
[1] 25
```

```
[1] 25 45 60
```

Cuadro C.1 Tipos básicos y vectores en R

```
# Asignación con <- (recomendado en R)
edad     <- 45L            # entero
peso_kg  <- 78.4           # numérico (double)
sexo     <- "Mujer"        # carácter
diabetes <- TRUE           # lógico

# Vectores con c()
edades       <- c(25, 30, 45, 22, 60)
diagnostico  <- c("HTA", "DM2", "HTA", "Sano", "DM2")
hospitalizado <- c(FALSE, TRUE, TRUE, FALSE, TRUE)

# Inspeccionar
class(edades);  length(edades);  range(edades)
```

Cuadro C.2 Indexación de vectores

```
# Posicional
edades[1]            # primer paciente
```

```
[1] 30 45 22 60
```

```
[1] 45 60
```

```
[1] "DM2" "HTA" "DM2"
```

```
[1] 30 60
```

C.1.3. Valores Perdidos: `NA`

En estudios clínicos los datos faltantes son la norma, no la excepción. R los representa con `NA`.

```
[1] NA
```

```
[1] 7.35
```

```
[1] FALSE FALSE  TRUE FALSE FALSE  TRUE
```

```
[1] 2
```

```
[1] 6.2 7.8 8.5 6.9
```

⚠ NA no es 0 ni cadena vacía

`NA == NA` devuelve `NA`, no `TRUE`. Usa siempre `is.na(x)` para comprobar valores perdidos. En análisis observacional, documenta el patrón de pérdida (¿MCAR, MAR o MNAR?) antes de elegir entre análisis de casos completos, imputación múltiple o ponderación inversa.

Cuadro C.3 Indexación de vectores

```
edades[c(1, 3, 5)]   # pacientes 1, 3 y 5
```

Cuadro C.4 Indexación de vectores

```
edades[-1]           # todos menos el primero
```

C.2. C.2 Estructuras de Datos

C.2.1. Data Frames y Tibbles

El **data frame** es la estructura central de R para datos tabulares. Un *tibble* (de `tibble`/`dplyr`) es un data frame con impresión más legible.

```
'data.frame':   6 obs. of  6 variables:
 $ id      : int  1 2 3 4 5 6
 $ edad    : num  25 30 45 60 55 72
 $ sexo    : chr  "M" "F" "F" "M" ...
 $ imc     : num  22.5 28 31.2 26.5 33.1 24.8
 $ hba1c   : num  5.4 6.2 7.8 6.8 8.5 5.9
 $ diabetes: num  0 0 1 1 1 0
```

```
  id edad sexo  imc hba1c diabetes
1  1   25    M 22.5   5.4        0
2  2   30    F 28.0   6.2        0
3  3   45    F 31.2   7.8        1
```

```
[1] 6 6
```

```
[1] "id"       "edad"     "sexo"     "imc"      "hba1c"    "diabetes"
```

C.2.2. Listas

Las listas almacenan objetos heterogéneos (el resultado típico de un test estadístico es una lista):

```
[1] 0.034
```

```
[1] 0.12 1.85
```

C.3. C.3 Factores y Variables Categóricas

En investigación médica, casi toda variable de exposición o desenlace es **categórica** (caso/control, expuesto/no expuesto, grupo de tratamiento). R las representa con **factores**.

```
[1] "No diabético" "Diabético"
```

Cuadro C.5 Indexación de vectores

```
# Lógica (filtros clínicos)
edades[edades >= 45]                # adultos mayores
```

Cuadro C.6 Indexación de vectores

```
diagnostico[hospitalizado]          # diagnósticos de hospitalizados
```

```
No diabético    Diabético
           3            3
```

C.3.1. Recodificación con `cut()` y `dplyr::case_when()`

```
       Bajo peso Normopeso Sobrepeso Obesidad
  <40          0         1         1        0
  40-64        0         0         1        2
  >=65         0         1         0        0
```

> **!** Nivel de referencia
>
> En regresión (lineal, logística, Cox), el **nivel de referencia** del factor define la categoría con la que se compara cada coeficiente. Por defecto R toma el primer nivel alfabético. Para forzar uno específico (p. ej. *Normopeso* como referencia):
>
> ```
> pacientes$imc_cat <- relevel(pacientes$imc_cat, ref = "Normopeso")
> ```

C.4. C.4 Importación de Datos Clínicos

En la práctica los datos llegan en CSV, Excel, SPSS o Stata. El paquete `readr` y `haven` cubren todos los formatos habituales en investigación biomédica.

> **Rutas robustas con `here::here()`**
>
> Evita rutas absolutas (`"C:/Users/.../"`). Usa proyectos de RStudio y `here::here("datos", "cohorte.csv")` para que el código funcione en cualquier máquina.

C.5. C.5 Manipulación de Datos con `dplyr`

Las cinco verbos esenciales de `dplyr` para preparar una cohorte:

Cuadro C.7 Indexación de vectores

```
edades[diagnostico == "DM2"]        # edad de diabéticos
```

Cuadro C.8 Manejo de valores perdidos

```
hba1c <- c(6.2, 7.8, NA, 8.5, 6.9, NA)

mean(hba1c)                # NA: cualquier NA contamina el resultado
```

Verbo	Acción
`select()`	elegir columnas
`filter()`	filtrar filas
`mutate()`	crear/modificar variables
`summarise()`	resumir (con o sin `group_by()`)
`arrange()`	ordenar

```
# A tibble: 2 x 5
  sexo_f     n edad_media imc_mediana pct_hba1c_alta
  <fct>  <int>      <dbl>       <dbl>          <dbl>
1 Hombre     2         66        25.6             50
2 Mujer      2         50        32.2            100
```

C.5.1. Uniones (`*_join`) y reestructuración (`pivot_*`)

C.6. C.6 Estadísticos Descriptivos

C.6.1. Variables Continuas

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.50   25.23   27.25   27.68   30.40   33.10

    media        de   mediana       RIC
27.683333  3.959503 27.250000  5.175000

    5%    25%    50%    75%    95%
23.075 25.225 27.250 30.400 32.625
```

C.6.2. Variables Categóricas

```
Bajo peso Normopeso Sobrepeso  Obesidad
        0         2         2         2

Bajo peso Normopeso Sobrepeso  Obesidad
  0.00000  33.33333  33.33333  33.33333

             obesidad
diabetes      FALSE TRUE
  No diabético     3    0
```

Cuadro C.9 Manejo de valores perdidos

```
mean(hba1c, na.rm = TRUE)  # 7.35: ignorar NAs
```

Cuadro C.10 Manejo de valores perdidos

```
is.na(hba1c)               # vector lógico
```

```
  No diabético 100.00000   0.00000
  Diabético     33.33333  66.66667
```

C.6.3. Resumen Estratificado

```
# A tibble: 2 x 6
  diabetes_f       n edad_media edad_de imc_mediana imc_RIC
  <fct>        <int>      <dbl>   <dbl>       <dbl>   <dbl>
1 No diabético     3       42.3   25.8         24.8    2.75
2 Diabético        3       53.3    7.64        31.2    3.30
```

C.7. C.7 Funciones Propias y Control de Flujo

C.7.1. Definir funciones

```
   imc categoria
1 19.0 Normopeso
2 23.5 Normopeso
3 30.0 Sobrepeso
```

C.7.2. Control de flujo: `if`, `for`, `sapply`

```
[1] "Normal"      "Prediabetes" "Diabetes"    "Diabetes"    "Diabetes"
```

> **Preferir vectorización sobre bucles**
>
> En R, `for` es más lento y verboso que la familia `apply`/`map` y que las funciones vectorizadas. Para un dataset clínico real (miles de filas), usa siempre `dplyr::mutate()` con `case_when()` o `sapply()`/`purrr::map_*()`.

C.8. C.8 Visualización con `ggplot2`

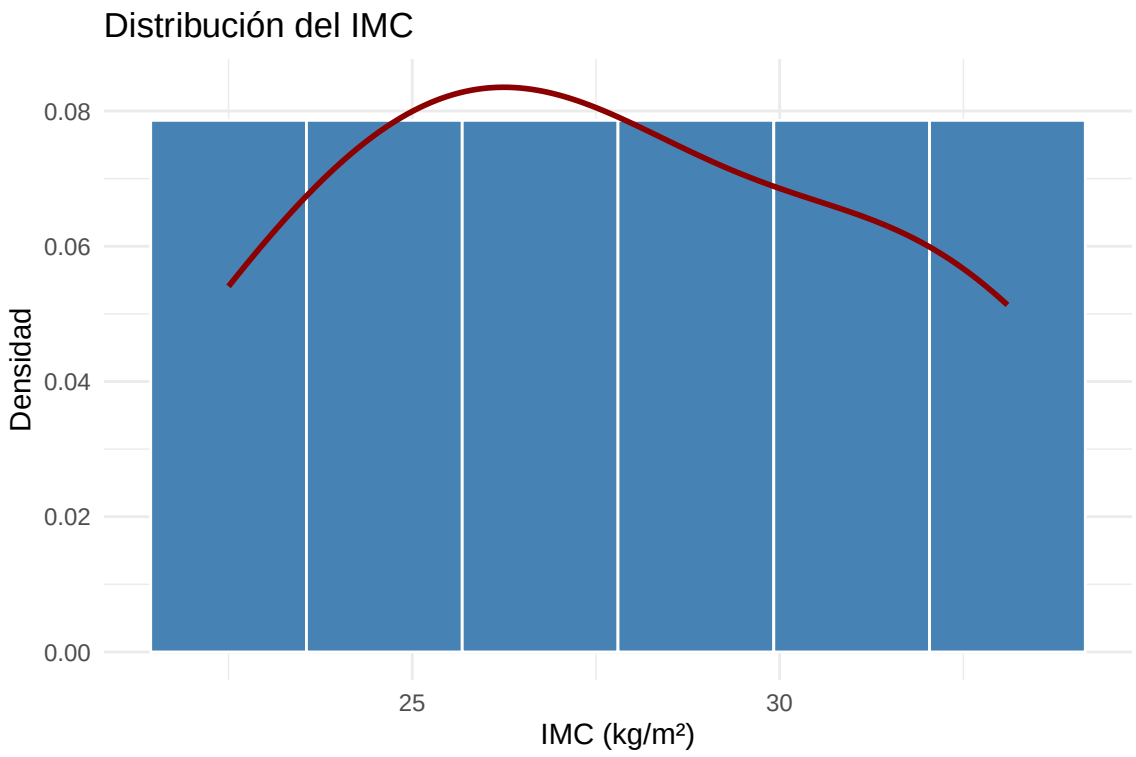
`ggplot2` es la herramienta estándar para visualización en publicaciones biomédicas.

Cuadro C.11 Manejo de valores perdidos

```
sum(is.na(hba1c))          # 2 valores perdidos
```

Cuadro C.12 Manejo de valores perdidos

```
hba1c[!is.na(hba1c)]       # eliminar NAs
```

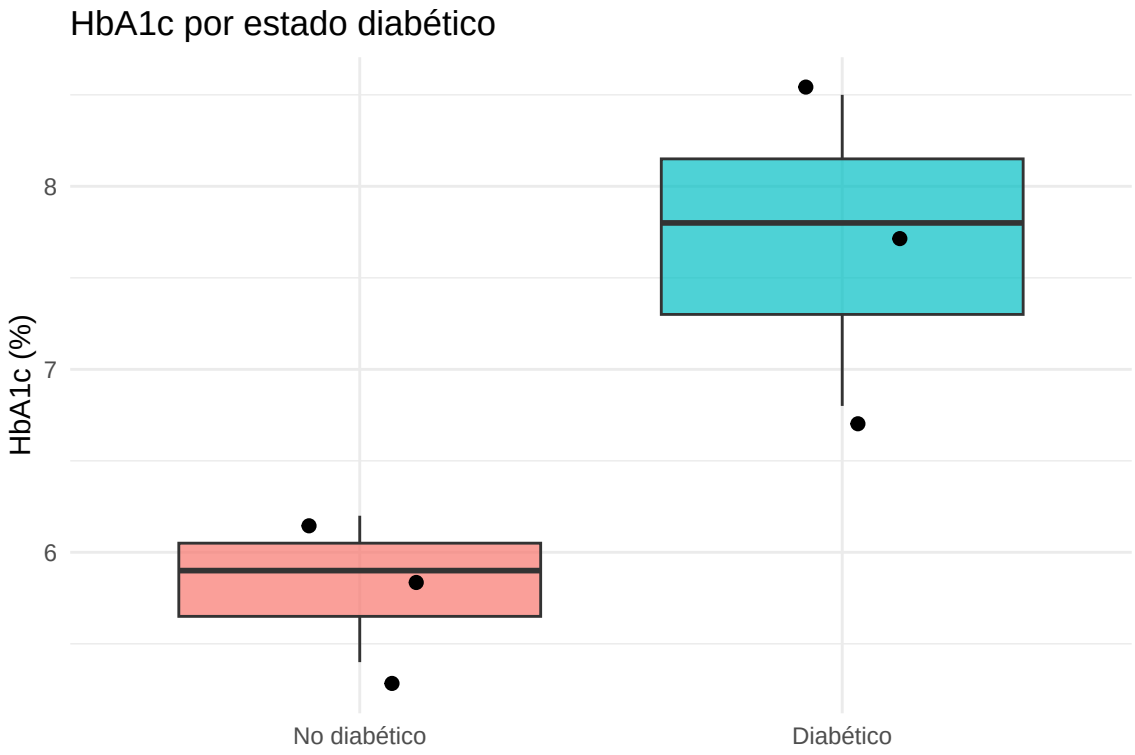


Cuadro C.13 Construcción de un data frame clínico

```r
pacientes <- data.frame(
  id        = 1:6,
  edad      = c(25, 30, 45, 60, 55, 72),
  sexo      = c("M", "F", "F", "M", "F", "M"),
  imc       = c(22.5, 28.0, 31.2, 26.5, 33.1, 24.8),
  hba1c     = c(5.4, 6.2, 7.8, 6.8, 8.5, 5.9),
  diabetes  = c(0, 0, 1, 1, 1, 0)
)

str(pacientes)        # estructura y tipos
```

Cuadro C.14 Construcción de un data frame clínico

```r
head(pacientes, 3)    # primeras 3 filas
```

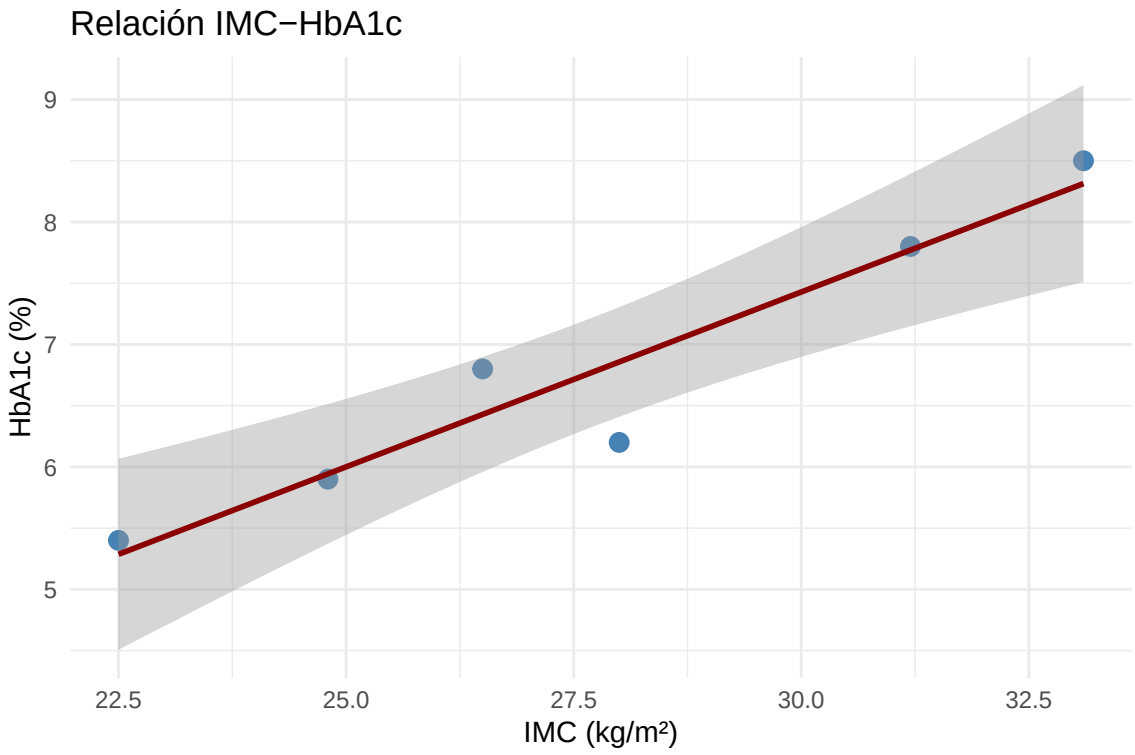


Cuadro C.15 Construcción de un data frame clínico

```
dim(pacientes)        # filas x columnas
```

Cuadro C.16 Construcción de un data frame clínico

```
names(pacientes)      # nombres de variables
```



C.9. C.9 Tests Estadísticos en R Base

Tabla de referencia con los contrastes más usados en investigación médica:

Pregunta clínica	Función de R
¿Difieren las medias de dos grupos?	`t.test()`
¿Difieren las medianas (no normal)?	`wilcox.test()`
¿Difieren tres o más medias?	`aov()`, `oneway.test()`
¿Asociación entre dos variables categóricas?	`chisq.test()`, `fisher.test()`
¿Datos apareados (antes/después)?	`mcnemar.test()`, `t.test(paired=T)`
¿Es la variable normal?	`shapiro.test()`
¿Hay correlación lineal?	`cor.test()`
Regresión lineal	`lm()`
Regresión logística / Poisson	`glm(family = binomial / poisson)`

Cuadro C.17 Listas: contenedores heterogéneos

```
resultado <- list(
  test       = "t-Student",
  p_valor    = 0.034,
  ic95       = c(0.12, 1.85),
  n_total    = 120
)
resultado$p_valor
```

Cuadro C.18 Listas: contenedores heterogéneos

```
resultado[["ic95"]]
```

```
    Welch Two Sample t-test

data:  hba1c by diabetes_f
t = -3.4207, df = 2.8523, p-value = 0.04515
alternative hypothesis: true difference in means between group No diabético and group Diabétic
95 percent confidence interval:
 -3.65529198 -0.07804135
sample estimates:
mean in group No diabético    mean in group Diabético
                  5.833333                   7.700000

    Fisher's Exact Test for Count Data

data:  table(pacientes$diabetes_f, pacientes$imc >= 30)
p-value = 0.4
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2031288       Inf
sample estimates:
odds ratio
       Inf

Call:
lm(formula = hba1c ~ imc + edad, data = pacientes)

Residuals:
      1       2       3       4       5       6
 0.2795 -0.4996  0.0762  0.2576  0.1586 -0.2723

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

Cuadro C.19 Factores, niveles y niveles de referencia

```
# Convertir a factor
pacientes$sexo_f <- factor(pacientes$sexo, levels = c("M", "F"),
                           labels = c("Hombre", "Mujer"))
pacientes$diabetes_f <- factor(pacientes$diabetes,
                               levels = c(0, 1),
                               labels = c("No diabético", "Diabético"))

levels(pacientes$diabetes_f)    # niveles
```

Cuadro C.20 Factores, niveles y niveles de referencia

```
table(pacientes$diabetes_f)     # frecuencia
```

```
(Intercept) -1.379357   1.316864  -1.047  0.37184
imc          0.279194   0.046683   5.981  0.00936 **
edad         0.008718   0.010235   0.852  0.45692
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4079 on 3 degrees of freedom
Multiple R-squared:  0.9288,    Adjusted R-squared:  0.8814
F-statistic: 19.57 on 2 and 3 DF,  p-value: 0.01899
```

```
                  2.5 %     97.5 %
(Intercept) -5.57020549 2.81149089
imc          0.13062701 0.42776027
edad        -0.02385378 0.04128988
```

```
                      OR        2.5 %      97.5 %
(Intercept) 0.0005718736 3.304375e-05 0.009897163
edad        1.0414789433 1.011409e+00 1.072443165
imc         1.1652214003 1.085568e+00 1.250719795
```

C.10. C.10 Paquetes Esenciales en Bioestadística

Selección curada de paquetes que el investigador clínico encuentra de forma recurrente. El bloque siguiente es informativo (no ejecutado).

Cuadro C.21 Categorizar variables continuas

```
library(dplyr)

pacientes <- pacientes %>%
  mutate(
    # Grupos de edad clínicos
    edad_grupo = cut(edad,
                     breaks = c(0, 40, 65, Inf),
                     labels = c("<40", "40-64", ">=65"),
                     right  = FALSE),
    # Clasificación IMC (OMS)
    imc_cat = case_when(
      imc <  18.5             ~ "Bajo peso",
      imc <  25               ~ "Normopeso",
      imc <  30               ~ "Sobrepeso",
      imc >= 30               ~ "Obesidad"
    ),
    imc_cat = factor(imc_cat,
                     levels = c("Bajo peso", "Normopeso",
                                "Sobrepeso", "Obesidad"))
  )

table(pacientes$edad_grupo, pacientes$imc_cat)
```

Citar los paquetes que usas

La revista te pedirá la versión de R y los paquetes. Usa `citation("survival")` y `sessionInfo()` en el cierre del análisis y guarda la salida.

C.11. C.11 Reproducibilidad en Investigación Médica

Un análisis médico debe poder reproducirse exactamente meses o años después (revisión por pares, auditoría, dudas del comité ético).

Sin semilla, sin reproducibilidad

Cualquier procedimiento estocástico (bootstrap, validación cruzada, imputación múltiple, simulación) **debe** ir precedido por `set.seed()`. Sin semilla, dos ejecuciones del mismo script darán p-valores e intervalos distintos — inaceptable en un análisis clínico.

Cuadro C.22 Importación desde distintos formatos

```
# CSV (recomendado: readr es más rápido y robusto que read.csv)
library(readr)
datos <- read_csv("datos/cohorte.csv")

# Excel (.xlsx)
library(readxl)
datos <- read_excel("datos/cohorte.xlsx", sheet = "Visita1")

# SPSS (.sav) -- muy común en investigación clínica
library(haven)
datos <- read_sav("datos/cohorte.sav")
datos <- as_factor(datos)   # convertir variables etiquetadas a factor

# Stata (.dta)
datos <- read_dta("datos/cohorte.dta")

# Exportar a CSV con codificación correcta (acentos)
write_csv(datos, "datos/cohorte_limpia.csv")
```

C.12. C.12 Resolución de Problemas Comunes

Síntoma	Causa habitual	Solución
`mean(x)` devuelve `NA`	Hay valores perdidos	`mean(x, na.rm = TRUE)`
Coeficiente de un factor no aparece	Es el nivel de referencia	`relevel(f, ref = "...")`
El gráfico no se ve en RStudio	Ventana cerrada o tamaño nulo	`dev.off()` y volver a ejecutar
`cannot find function "..."`	Paquete no cargado	`library(paquete)` o `install.packages("paquete")`
`argument "x" is missing`	Olvidaste un argumento	Consulta `?nombre_funcion`
Lectura de CSV con acentos rotos	Codificación incorrecta	`read_csv(..., locale = locale(encoding = "Latin1"))`
`non-numeric argument to binary operator`	Tipo carácter donde esperabas numérico	`as.numeric(x)` (cuidado con factores: `as.numeric(as.character(x))`)
Resultados distintos en cada ejecución	Falta `set.seed()`	Fijar semilla antes del análisis
`object '...' not found`	Nombre mal escrito o variable de otro entorno	`ls()` para listar objetos, `names(df)` para variables

C.12.1. Comandos de ayuda

C.13. Resumen

Esta guía te ha llevado del vector más simple a un flujo completo de análisis clínico en R: importar

Cuadro C.23 Pipeline típico de preparación de una cohorte

```
library(dplyr)

resumen <- pacientes %>%
  filter(edad >= 40) %>%                                    # adultos
  mutate(hba1c_alta = if_else(hba1c >= 6.5, 1L, 0L)) %>%  #  6.5: DM
  group_by(sexo_f) %>%
  summarise(
    n             = n(),
    edad_media    = mean(edad, na.rm = TRUE),
    imc_mediana   = median(imc, na.rm = TRUE),
    pct_hba1c_alta = mean(hba1c_alta, na.rm = TRUE) * 100,
    .groups = "drop"
  )
resumen
```

Cuadro C.24 Unir tablas y pivotar a formato largo/ancho

```
library(dplyr); library(tidyr)

# Unir cohorte con tabla de laboratorio por id de paciente
cohorte_lab <- pacientes %>%
  left_join(laboratorio, by = "id")

# Pasar de ancho (una columna por visita) a largo (una fila por visita)
largo <- visitas %>%
  pivot_longer(cols = starts_with("pas_v"),
               names_to  = "visita",
               values_to = "pas_mmHg")
```

Cuadro C.25 Resumen univariante de variables continuas

```
summary(pacientes$imc)                                  # min, Q1, med, mean, Q3, max
```

Cuadro C.26 Resumen univariante de variables continuas

```
c(media = mean(pacientes$imc),
  de    = sd(pacientes$imc),
  mediana = median(pacientes$imc),
  RIC = IQR(pacientes$imc))
```

Cuadro C.27 Resumen univariante de variables continuas

```
# Cuantiles personalizados
quantile(pacientes$imc, probs = c(0.05, 0.25, 0.5, 0.75, 0.95))
```

Cuadro C.28 Frecuencias absolutas y relativas

```
# Frecuencias
table(pacientes$imc_cat)
```

Cuadro C.29 Frecuencias absolutas y relativas

```
# Proporciones
prop.table(table(pacientes$imc_cat)) * 100
```

Cuadro C.30 Frecuencias absolutas y relativas

```
# Tabla de contingencia 2x2
tabla <- table(diabetes = pacientes$diabetes_f,
               obesidad = pacientes$imc >= 30)
tabla
```

Cuadro C.31 Frecuencias absolutas y relativas

```
prop.table(tabla, margin = 1) * 100   #% por fila (riesgo en expuestos)
```

Cuadro C.32 Descriptivos por grupo (típico de Tabla 1)

```
pacientes %>%
  group_by(diabetes_f) %>%
  summarise(
    n          = n(),
    edad_media = mean(edad),
    edad_de    = sd(edad),
    imc_mediana = median(imc),
    imc_RIC    = IQR(imc),
    .groups = "drop"
  )
```

Cuadro C.33 Función propia: cálculo e interpretación clínica

```
# Cálculo del IMC con interpretación
clasificar_imc <- function(peso_kg, altura_m) {
  imc <- peso_kg / (altura_m^2)
  categoria <- cut(imc,
                   breaks = c(0, 18.5, 25, 30, Inf),
                   labels = c("Bajo peso", "Normopeso",
                              "Sobrepeso", "Obesidad"),
                   right = FALSE)
  data.frame(imc = round(imc, 1), categoria = categoria)
}

clasificar_imc(peso_kg = c(55, 72, 95),
               altura_m = c(1.70, 1.75, 1.78))
```

Cuadro C.34 Iteración vectorizada con sapply

```
# Vectorizado (preferido en R)
hba1c_vals <- c(5.6, 6.1, 6.5, 7.2, 8.0)

categoria_dm <- sapply(hba1c_vals, function(x) {
  if      (x <  5.7) "Normal"
  else if (x <  6.5) "Prediabetes"
  else               "Diabetes"
})
categoria_dm
```

Cuadro C.35 Histograma con curva de densidad

```
library(ggplot2)

ggplot(pacientes, aes(x = imc)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 6, fill = "steelblue", color = "white") +
  geom_density(color = "darkred", linewidth = 1) +
  labs(x = "IMC (kg/m²)", y = "Densidad",
       title = "Distribución del IMC") +
  theme_minimal()
```

Cuadro C.36 Boxplot estratificado: HbA1c por estado diabético

```
ggplot(pacientes, aes(x = diabetes_f, y = hba1c, fill = diabetes_f)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(width = 0.15, size = 2) +
  labs(x = NULL, y = "HbA1c (%)",
       title = "HbA1c por estado diabético") +
  theme_minimal() +
  theme(legend.position = "none")
```

Cuadro C.37 Diagrama de dispersión con línea de regresión

```
ggplot(pacientes, aes(x = imc, y = hba1c)) +
  geom_point(size = 3, color = "steelblue") +
  geom_smooth(method = "lm", se = TRUE, color = "darkred") +
  labs(x = "IMC (kg/m²)", y = "HbA1c (%)",
       title = "Relación IMC-HbA1c") +
  theme_minimal()
```

Cuadro C.38 Ejemplos: t-test, chi-cuadrado, regresión lineal

```
# 1) Comparar HbA1c entre diabéticos y no diabéticos
t.test(hba1c ~ diabetes_f, data = pacientes)
```

Cuadro C.39 Ejemplos: t-test, chi-cuadrado, regresión lineal

```
# 2) Asociación diabetes ~ obesidad (Fisher por n pequeño)
fisher.test(table(pacientes$diabetes_f, pacientes$imc >= 30))
```

Cuadro C.40 Ejemplos: t-test, chi-cuadrado, regresión lineal

```r
# 3) Regresión lineal: HbA1c ~ IMC + edad
modelo <- lm(hba1c ~ imc + edad, data = pacientes)
summary(modelo)
```

Cuadro C.41 Ejemplos: t-test, chi-cuadrado, regresión lineal

```r
confint(modelo)
```

Cuadro C.42 Regresión logística: probabilidad de diabetes

```r
# Pacientes simulados más amplios para que el modelo sea estimable
set.seed(123)
n <- 200
sim <- data.frame(
  edad = rnorm(n, 55, 12),
  imc  = rnorm(n, 28, 5)
)
sim$diabetes <- rbinom(n, 1,
                       plogis(-7 + 0.05*sim$edad + 0.12*sim$imc))

mod_log <- glm(diabetes ~ edad + imc, data = sim, family = binomial)
# Odds ratios con IC95%
exp(cbind(OR = coef(mod_log), confint.default(mod_log)))
```

Cuadro C.43 Paquetes recomendados según tarea

```
# --- Reportes clínicos y "Tabla 1" ----------------------------------------
library(gtsummary)   # Tablas publicables a partir de modelos / data frames
library(tableone)    # Tabla 1 demográfica clásica

tabla1 <- tbl_summary(pacientes,
                      by = diabetes_f,
                      missing = "no") %>%
  add_p() %>% add_overall()

# --- Análisis epidemiológico ----------------------------------------------
library(epiR)        # Medidas de asociación 2x2 (RR, OR, NNT)
library(epitools)    # Tablas, cálculo de IC para incidencia / prevalencia

# --- Supervivencia --------------------------------------------------------
library(survival)
library(survminer)
fit_km  <- survfit(Surv(tiempo, evento) ~ grupo, data = cohorte)
ggsurvplot(fit_km, risk.table = TRUE, pval = TRUE)

fit_cox <- coxph(Surv(tiempo, evento) ~ edad + sexo + tratamiento,
                 data = cohorte)
summary(fit_cox)

# --- Inferencia causal / emparejamiento -----------------------------------
library(MatchIt)     # Propensity score matching
library(WeightIt)    # Ponderación inversa de probabilidad (IPTW)

# --- Datos perdidos -------------------------------------------------------
library(mice)        # Imputación múltiple (FCS)
library(naniar)      # Visualización del patrón de missing

# --- Modelos: limpieza y extracción de resultados -------------------------
library(broom)       # tidy(), glance(), augment() para modelos
library(performance) # Diagnósticos de regresión (colinealidad, supuestos)

# --- Meta-análisis --------------------------------------------------------
library(meta)        # Meta-análisis clásico (efecto fijo / aleatorio)
library(metafor)     # Meta-análisis y meta-regresión flexibles
```

Cuadro C.44 Buenas prácticas de reproducibilidad

```
# 1) Semilla aleatoria SIEMPRE antes de cualquier simulación o aleatorización
set.seed(20260101)

# 2) Proyecto de RStudio + rutas relativas con here
library(here)
datos <- read_csv(here("datos", "cohorte_v3.csv"))

# 3) Capturar el entorno: versión de R y paquetes
sessionInfo()

# 4) (Opcional) renv para bloquear versiones de paquetes
# install.packages("renv"); renv::init(); renv::snapshot()

# 5) Análisis dentro de un documento Quarto/Rmd
#    -> el código y los resultados viven juntos
```

Cuadro C.45 Sistema de ayuda de R

```
?mean                  # ayuda de una función
??"survival analysis"  # búsqueda por palabra clave
help(package = "dplyr")
vignette("dplyr")      # tutoriales largos del paquete
example(lm)            # ejemplos ejecutables

# Diagnóstico rápido del entorno
getwd();  setwd("ruta")
ls();     rm(list = ls())   # listar / limpiar objetos
sessionInfo()
```

Apéndice D

Apéndice D: Investigación Reproducible con Git, GitHub y RStudio

La **investigación reproducible** es el principio según el cual cualquier análisis cuantitativo —y, por extensión, cualquier resultado publicado— debe poder ser replicado por terceros a partir de los datos, el código y la documentación originales. En bioestadística, donde las decisiones clínicas pueden depender de un único valor p o de un único intervalo de confianza, la reproducibilidad no es un lujo sino una exigencia ética.

Este apéndice introduce las tres herramientas mínimas que todo investigador en ciencias de la salud debería dominar para producir análisis reproducibles: **Git** (control de versiones), **GitHub** (plataforma de colaboración basada en Git) y **RStudio** (entorno de desarrollo integrado para R que incorpora un cliente Git).

D.1. D.1 ¿Por qué investigación reproducible?

! La crisis de la replicabilidad

Múltiples estudios desde 2005 (Ioannidis, *PLoS Medicine*) han documentado que una fracción sustancial de los hallazgos publicados en biomedicina no logra replicarse cuando se repite el análisis con los datos originales o se realiza un estudio independiente. Esta **crisis de replicabilidad** ha motivado iniciativas internacionales (ICMJE, NIH, COS) que hoy exigen compartir datos y código como condición para publicar.

D.1.1. D.1.1 Niveles de reproducibilidad

Siguiendo la taxonomía de **Goodman, Fanelli & Ioannidis (2016)**:

Nivel	Significado	Requiere
Reproducibilidad de métodos	Otra persona obtiene los mismos resultados a partir de los datos y el código originales	Código + datos + entorno computacional
Reproducibilidad de resultados	Otro investigador, con datos nuevos, obtiene conclusiones consistentes	Diseño claro + protocolo público
Reproducibilidad inferencial	El conocimiento extraído del estudio se confirma en revisiones sistemáticas y meta-análisis	Reporting estandarizado (CONSORT, STROBE…)

D.1.2. D.1.2 Principios FAIR

Los datos y el código de un estudio reproducible deben ser **FAIR**: *Findable* (con identificador persistente, p. ej. DOI), *Accessible* (depositados en un repositorio público), *Interoperable* (formatos abiertos como CSV, parquet) y *Reusable* (con licencia explícita y documentación).

D.1.3. D.1.3 ¿Qué herramientas mínimas necesitamos?

- **Control de versiones** para registrar cada cambio del análisis (Git).
- **Plataforma de colaboración** para alojar el repositorio y compartirlo (GitHub, GitLab…).
- **Documentos computacionales** que mezclen código y narrativa (Quarto, R Markdown, Jupyter).
- **Gestión de entornos** que congele las versiones de los paquetes (`renv`, `conda`).

Este apéndice cubre los dos primeros, integrados en RStudio. La gestión de entornos se introduce brevemente en la sección D.5.

D.2. D.2 Control de versiones con Git

Git es un sistema distribuido de control de versiones: registra el historial completo de cambios en un conjunto de archivos (un *repositorio*) y permite recuperar cualquier estado anterior, comparar versiones, y trabajar en paralelo en líneas alternativas (*ramas*) que después se fusionan.

D.2.1. D.2.1 Conceptos básicos

Concepto	Significado
Repositorio	Carpeta del proyecto bajo control de versiones (`.git/` oculto en su raíz)
Commit	Instantánea (snapshot) del estado del proyecto en un momento dado, con un mensaje descriptivo
Branch (rama)	Línea de desarrollo paralela; `main` es la rama principal
Merge	Fusión de los cambios de una rama en otra
HEAD	Puntero al commit actualmente activo
Remoto (`origin`)	Copia del repositorio alojada en un servidor (típicamente GitHub)
`.gitignore`	Archivo que lista patrones que Git debe ignorar (datos crudos, caches, credenciales)

D.2.2. D.2.2 Instalación

Instalación de Git según el sistema operativo

- **macOS:** preinstalado o `xcode-select --install` (Command Line Tools).
- **Windows:** descargar **Git for Windows** desde https://git-scm.com/download/win (incluye *Git Bash*).
- **Linux (Debian/Ubuntu):** `sudo apt install git`.

Verificación:

```
git --version
# git version 2.42.0 (o superior)
```

D.2.3. D.2.3 Configuración inicial (una sola vez por máquina)

```
git config --global user.name  "Tu Nombre"
git config --global user.email "tu_email@ugr.es"
git config --global init.defaultBranch main
```

Equivalente desde R (recomendado para los lectores de este libro):

```
# install.packages("usethis")
usethis::use_git_config(
  user.name  = "Tu Nombre",
  user.email = "tu_email@ugr.es"
)
```

D.2.4. D.2.4 Flujo de trabajo básico (comandos esenciales)

```
git init                        # Crea un nuevo repositorio en la carpeta actual
git status                      # Muestra archivos modificados/nuevos
git add archivo.qmd             # Marca el archivo para incluirlo en el próximo commit
git add .                       # Marca TODOS los archivos modificados (cuidado con datos sensil
git commit -m "Mensaje claro"   # Crea un commit con los cambios marcados
git log --oneline               # Historial resumido de commits
git diff archivo.qmd            # Muestra los cambios sin hacer commit
git checkout -- archivo.qmd     # Descarta los cambios no comprometidos
```

D.2.5. D.2.5 Buenas prácticas para mensajes de commit

Reglas para escribir buenos mensajes

1. **Línea de asunto** en **imperativo** y < 70 caracteres: *"Añade tabla descriptiva de la cohorte"* (no *"añadiendo tabla"* ni *"añadido tabla"*).
2. **Una idea por commit:** un mensaje = un cambio coherente. Evita los *commits* monstruo que mezclan refactorizaciones con nuevas funcionalidades.
3. **Explica el porqué, no el qué.** El *qué* ya lo muestra `git diff`; el mensaje debe responder a *por qué* se hizo el cambio.
4. **Frecuencia:** mejor muchos *commits* pequeños y revertibles que pocos enormes.

Ejemplos:

- `Corrige criterio de inclusión: edad >= 18`
- `Añade IC bootstrap al modelo de mortalidad`
- `cambios` (vago)
- `WIP` (uso interno aceptable, pero no en `main`)

D.2.6. D.2.6 Ramas y fusiones (introducción)

Las **ramas** permiten experimentar sin tocar la versión "buena" del análisis:

```
git switch -c sensibilidad      # Crea y cambia a la rama 'sensibilidad'
# … trabajo, commits …
git switch main                 # Vuelve a la rama principal
git merge sensibilidad          # Fusiona los cambios de 'sensibilidad' en main
git branch -d sensibilidad      # Elimina la rama una vez fusionada
```

En proyectos colaborativos, las ramas se asocian habitualmente con *pull requests* en GitHub (ver D.3).

D.3. D.3 GitHub como plataforma colaborativa

GitHub (https://github.com) es un servicio en la nube que aloja repositorios Git, añadiendo funcionalidades de colaboración:

- **Repositorios públicos** (visibles y clonables por cualquiera) o **privados**.
- **Issues:** sistema de tickets para reportar errores, proponer mejoras o discutir decisiones.
- **Pull requests (PR):** mecanismo formal para revisar y fusionar cambios entre ramas, con discusión línea a línea.
- **Releases:** versiones etiquetadas del proyecto (`v1.0`, `v2.0-beta`…) con notas y artefactos descargables.
- **GitHub Pages:** publicación gratuita de sitios web estáticos (el propio libro que está leyendo se sirve mediante GitHub Pages).
- **GitHub Actions:** automatización (re-renderizar el libro con cada commit, lanzar análisis, enviar avisos…).

D.3.1. D.3.1 Crear una cuenta y un repositorio remoto

1. Crear cuenta en https://github.com (gratuita; usuarios académicos disponen del plan **GitHub Education** sin límite).
2. Pulsar **New repository**, asignar nombre (ej. `tfg-osteoporosis`), elegir entre **Public** o **Private**, y crear sin README inicial si va a conectar un repositorio local ya existente.

D.3.2. D.3.2 Autenticación: HTTPS + PAT o SSH

GitHub no permite contraseñas en línea de comandos desde 2021. Dos opciones modernas:

- **HTTPS + Personal Access Token (PAT):** sencillo para usuarios noveles. Crear en *Settings → Developer settings → Personal access tokens*. R lo facilita:

```
# install.packages(c("gitcreds", "usethis"))
usethis::create_github_token()  # Abre la página de GitHub con permisos preseleccionados
gitcreds::gitcreds_set()        # Guarda el token de forma segura en el llavero del sistema
```

- **SSH:** más cómodo a largo plazo. Generar par de claves con `ssh-keygen -t ed25519 -C "tu_email@ugr.es"`, copiar la clave pública (`~/.ssh/id_ed25519.pub`) y añadirla en GitHub → *Settings → SSH and GPG keys*.

D.3.3. D.3.3 Conectar un repositorio local con un remoto

```
# Una vez creado el repositorio vacío en GitHub:
git remote add origin git@github.com:tu_usuario/tfg-osteoporosis.git
git branch -M main
git push -u origin main          # 'origin' es el remoto, 'main' la rama
```

A partir de aquí:

```
git pull              # Trae los cambios del remoto
git push              # Sube tus commits al remoto
```

D.3.4. D.3.4 Pull requests: revisión por pares del código

El flujo típico para un cambio sustancial es:

1. Crear una rama: `git switch -c modelo-logistico`.
2. Hacer commits sobre esa rama.
3. Publicarla: `git push -u origin modelo-logistico`.
4. Abrir un **pull request** en GitHub desde la rama hacia `main`.
5. Discutir, revisar línea a línea, ajustar y, finalmente, fusionar.

Este flujo, equivalente a la revisión por pares de un artículo científico, mejora la calidad del código y deja registrado el porqué de cada decisión.

D.4. D.4 Integración con RStudio

RStudio incluye un **panel Git** que cubre el 90 % de las operaciones cotidianas sin abrir la terminal. Esto facilita enormemente la adopción de Git para usuarios de R que no se sienten cómodos en línea de comandos.

D.4.1. D.4.1 Crear un proyecto RStudio con Git desde cero

1. **File → New Project → New Directory → New Project**.
2. Marcar la casilla **Create a git repository**.
3. RStudio crea la carpeta, inicializa Git e incluye un `.gitignore` razonable (`.Rproj.user/`, `.Rhistory`, `.RData`).
4. El panel **Git** aparece en la esquina superior derecha.

D.4.2. D.4.2 Conectar un proyecto existente a Git

Para un proyecto RStudio sin Git:

```
# install.packages("usethis")
usethis::use_git()           # Inicializa Git y hace el primer commit
usethis::use_github()        # Crea el repositorio en GitHub y lo conecta (requiere PAT)
```

D.4.3. D.4.3 El panel Git de RStudio

Acción en el panel	Equivalente en línea de comandos
Marcar la casilla del archivo (*staged*)	`git add archivo`
Botón **Commit** + mensaje	`git commit -m "…"`
Botón **Push** (flecha verde arriba)	`git push`
Botón **Pull** (flecha azul abajo)	`git pull`
Botón **History** (reloj)	`git log` (con interfaz gráfica)
Botón **Diff**	`git diff` (visualización lado a lado)

Workflow diario en RStudio

1. Antes de empezar: **Pull** (sincronizar con el remoto).
2. Trabajar en el análisis (editar `.R`, `.qmd`, etc.).
3. Cada cambio coherente: **Stage** los archivos, **Commit** con un mensaje claro.
4. Al final de la sesión: **Push** (compartir con el equipo).

Repetir varias veces al día: *un commit, una idea.*

D.4.4. D.4.4 Visualización de cambios y resolución de conflictos

El botón **Diff** en RStudio muestra los cambios línea a línea con colores (rojo = eliminado, verde = añadido). Cuando dos personas modifican la misma línea de un archivo, Git señala un **conflicto** que se resuelve manualmente editando los marcadores `<<<<<<<`, `=======` y `>>>>>>>` antes de hacer commit.

D.5. D.5 Flujos de trabajo reproducibles para Bioestadística

D.5.1. D.5.1 Estructura recomendada de un proyecto bioestadístico

```
mi-estudio/
  README.md              # Propósito, autor, citación
  LICENSE                # CC-BY, MIT, GPL…
  .gitignore             # Archivos a ignorar
  mi-estudio.Rproj       # Proyecto RStudio
  data/
     raw/               # Datos originales (a menudo gitignored si son sensibles)
     processed/         # Datos derivados, reproducibles desde el script
  R/                     # Funciones reutilizables
  analysis/              # Cuadernos Quarto/Rmd con los análisis
  results/               # Tablas, figuras, modelos serializados (.rds)
  manuscript/            # Manuscrito en Quarto o LaTeX
  renv.lock              # Snapshot de las versiones de los paquetes (opcional)
```

D.5.2. D.5.2 `.gitignore` típico para un proyecto en R

```
# Artefactos de R
.Rhistory
.RData
.Ruserdata
.Rproj.user/

# Salida de Quarto / RMarkdown
*_files/
*_cache/
/_book/
/_site/

# Sistema operativo
.DS_Store
Thumbs.db

# Datos sensibles (¡nunca commitar PII!)
data/raw/*.csv
data/raw/*.xlsx
```

> ⚠ Datos sensibles y RGPD
>
> **Jamás** se debe subir a GitHub información identificable de pacientes (PII): nombres, DNI, fechas de nacimiento, códigos de historia clínica, dirección, etc. Aunque elimine el archivo en un commit posterior, **permanece en el historial**. Si por accidente sube datos sensibles, debe utilizar herramientas como `git filter-repo` o `BFG Repo-Cleaner` y rotar inmediatamente cualquier credencial expuesta. Para datos clínicos, la recomendación es trabajar con datos **anonimizados o seudonimizados** desde el primer commit.

D.5.3. D.5.3 Entornos reproducibles con `renv`

Para garantizar que el análisis se ejecuta dentro de cinco años con las mismas versiones de los paquetes:

```r
# install.packages("renv")
renv::init()      # Crea un entorno local + renv.lock
renv::snapshot()  # Actualiza renv.lock con las versiones actuales
renv::restore()   # Reinstala exactamente las versiones del lockfile
```

Conviene comprometer (`git commit`) el archivo `renv.lock`.

D.5.4. D.5.4 Publicar un libro o análisis con GitHub Pages

Este libro mismo está alojado en GitHub Pages. El flujo es:

1. En `_quarto.yml`: `project: type: book` con `output-dir: docs`.
2. Renderizar localmente: `quarto render` (se genera la carpeta `docs/`).
3. Comprometer y subir: `git add docs/ && git commit -m "Publica versión X" && git push`.
4. En GitHub → *Settings* → *Pages*, seleccionar **branch: main**, **folder: /docs**.
5. La URL `https://tu_usuario.github.io/nombre-repo/` queda activa.

Para automatizar la renderización en cada *push* se utiliza **GitHub Actions** (ver el archivo `.github/workflows/publish.yml` del repositorio de este libro como ejemplo).

D.6. D.6 Ejercicio guiado: tu primer repositorio reproducible

Crear un repositorio público con un análisis del dataset `osteo` que se publique automáticamente como página web.

Pasos

1. Crear cuenta y proyecto:

- Crear cuenta en https://github.com.
- En RStudio: **File → New Project → New Directory → Quarto Project** (Book/Website/Document).
- Marcar **Create a git repository**.

2. Configurar Git e instalar BioEstatR:

```
usethis::use_git_config(user.name = "Tu Nombre", user.email = "tu@ugr.es")
remotes::install_github("migariane/BioEstatR")
```

3. Crear un análisis con Quarto:

Archivo `analisis.qmd` (contenido sugerido):

````
---
title: "Análisis exploratorio del dataset osteo"
author: "Tu Nombre"
date: today
format: html
---

```{r}
library(BioEstatR)
data(osteo)
````
````

```
freq(osteo$tabaco)
grps(osteo$imc, osteo$sexo)
tabla2x2(fvar = osteo$tabaco, cvar = osteo$osteo_cue, o = osteo)
rls(hba1c ~ tevol, data = osteo)
**4. Renderizar localmente:** `quarto render analisis.qmd` produce `analisis.html`.

**5. Primer commit:**

```bash
git add .
git commit -m "Primer análisis exploratorio del dataset osteo"
```

6. Crear el repositorio en GitHub y conectarlo:

```
usethis::use_github(private = FALSE)
```

7. Activar GitHub Pages en *Settings* → *Pages* → *main / docs* (o `/root` si el HTML está en la raíz).
8. Verificar: abrir `https://tu_usuario.github.io/nombre-repo/` y comprobar que el análisis es accesible públicamente.

Criterios de éxito

- El repositorio aparece en su perfil de GitHub con el archivo `analisis.qmd` y el commit inicial.
- Existe un `README.md` que describe el propósito del repositorio y cita el paquete `BioEstatR`.
- El archivo `.gitignore` excluye `*_files/`, `.Rhistory`, `.Rproj.user/`.
- La página renderizada es accesible desde la URL de GitHub Pages.
- Cualquier persona puede clonar el repositorio (`git clone …`) y reproducir el análisis ejecutando `quarto render analisis.qmd` tras `remotes::install_github("migariane/BioEstatR")`.

D.7. D.7 Recursos adicionales

Lecturas recomendadas

Libros y guías (gratuitos y en abierto):

- **Bryan, J.** *Happy Git and GitHub for the useR*. https://happygitwithr.com — guía orientada a usuarios de R.
- **Chacon, S. & Straub, B.** *Pro Git* (2.ª ed.). https://git-scm.com/book — referencia exhaustiva de Git.
```

- **Wickham, H. & Bryan, J.** *R Packages* (2.ª ed.). https://r-pkgs.org — incluye flujos de trabajo con Git y GitHub para desarrollar paquetes de R.
- **Xie, Y., Dervieux, C. & Riederer, E.** *R Markdown Cookbook*. https://bookdown.org/yihui/rmarkdown-cookbook/.
- **Quarto Documentation.** *Publishing with GitHub Pages*. https://quarto.org/docs/publishing/github-pages.html.

Artículos clave sobre reproducibilidad en biomedicina:

- Ioannidis, J. P. A. (2005). *Why most published research findings are false*. PLoS Medicine, 2(8), e124.
- Goodman, S. N., Fanelli, D. & Ioannidis, J. P. A. (2016). *What does research reproducibility mean?* Science Translational Medicine, 8(341).
- Peng, R. D. (2011). *Reproducible research in computational science*. Science, 334(6060), 1226–1227.

Iniciativas y manifiestos:

- **TOP guidelines** (Transparency and Openness Promotion): https://www.cos.io/initiatives/top-guidelines
- **FAIR Principles:** https://www.go-fair.org/fair-principles/
- **EQUATOR Network** (reporting guidelines para investigación biomédica): https://www.equator-network.org

Cinco hábitos para empezar hoy

1. **Inicializa Git en cada proyecto nuevo** (incluso TFG y prácticas).
2. **Commit pequeño y frecuente** con mensaje claro en imperativo.
3. **Sube a GitHub** tus análisis y código (manteniendo los datos sensibles fuera).
4. **Documenta el README**: propósito, datos, instrucciones de reproducción.
5. **Cita el código y los datos** con un DOI permanente al publicar (Zenodo se integra automáticamente con GitHub).

La reproducibilidad no es un destino sino un hábito. Cada commit es un compromiso con la transparencia, y cada repositorio público es una contribución al conocimiento compartido.

Apéndice E

Examen Final

Asignatura: Matemáticas para la Estadística **Institución:** Universidad de Granada **Duración:** 2 horas 30 minutos **Puntuación Total:** 125 puntos (seleccionar 4 de 5 partes = 100 puntos, o responder las 5 para nota extra)

E.1. Instrucciones

- Este examen cubre los 5 bloques: Análisis Exploratorio, Probabilidad, Inferencia, Regresión Lineal y Datos Categóricos
- Se permite material de consulta: calculadora, tablas de distribuciones y tablas de la distribución χ^2
- No se permite: apuntes, Internet, comunicación con otros estudiantes
- Muestre todo el trabajo y cálculos intermedios
- Las respuestas deben estar claramente identificadas

E.2. Parte I: Análisis Exploratorio de Datos (25 puntos)

E.2.1. Pregunta 1.1 (12 puntos)

Se recopilaron datos de altura (cm) de 50 estudiantes de primer año:

```
165, 168, 171, 160, 175, 172, 169, 166, 173, 170, 167, 164, 176, 168, 172,
165, 169, 171, 174, 168, 170, 166, 173, 167, 175, 169, 168, 172, 171, 165,
174, 170, 167, 173, 169, 166, 172, 168, 171, 175, 170, 169, 173, 167, 174,
168, 172, 170, 175, 171
```

a) Calcule la media, mediana, desviación estándar y rango intercuartílico.
b) Construya una tabla de frecuencias con 5 intervalos de clase.
c) Interprete la distribución (simetría, dispersión).

Mostrar respuesta 1.1 (solo instructor)

a) Estadísticos descriptivos: - Media: $\bar{x} = \frac{\sum x_i}{n} \approx 170.1$ cm - Mediana: $Q_2 \approx 170$ cm (valor central de los 50 datos ordenados) - Desviación estándar: $s \approx 3.8$ cm - Rango intercuartílico: $IQR = Q_3 - Q_1 \approx 173 - 167 = 6$ cm

b) Tabla de frecuencias (5 intervalos, amplitud 4 cm):

Intervalo	Frecuencia	Frecuencia relativa	Frecuencia acumulada
160-164	3	0.06	0.06
164-168	12	0.24	0.30
168-172	22	0.44	0.74
172-176	13	0.26	1.00

c) Interpretación: La distribución es aproximadamente simétrica, centrada alrededor de 170 cm, con una concentración principal en 168-172 cm (44 % de los datos). La dispersión es moderada (s 3.8 cm). No hay valores atípicos evidentes.

E.2.2. Pregunta 1.2 (13 puntos)

Se estudia la relación entre número de horas de estudio (X) y calificación en examen (Y) de 10 estudiantes:

Horas	2	3	2.5	4	3.5	5	1.5	4.5	3	2
Nota	4	5	4.5	6	5.5	7	3	6.5	5	4

a) Calcule el coeficiente de correlación de Pearson entre X e Y.
b) Interprete el resultado.
c) ¿Pueden haber valores atípicos?

Mostrar respuesta 1.2 (solo instructor)

a) Coeficiente de Pearson:

$$\text{Primero: } \bar{x} = 3.1, \quad \bar{y} = 5.0$$

$$s_x = 1.23, \quad s_y = 1.27$$

$$\text{Covarianza: } \text{Cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \approx 1.40$$

$$r = \frac{\text{Cov}(X,Y)}{s_x s_y} = \frac{1.40}{1.23 \times 1.27} \approx 0.90$$

b) Interpretación: Existe una correlación positiva muy fuerte (r 0.90) entre horas de estudio y nota del examen. A mayor número de horas, mayor calificación. Aproximadamente 81 % (r^2) de la variación en notas se explica por las horas de estudio.

c) Valores atípicos: La relación es lineal sin desvíos aparentes. No hay evidencia de outliers en los datos.

E.3. Parte II: Probabilidad y Variables Aleatorias (25 puntos)

E.3.1. Pregunta 2.1 (12 puntos)

Un examen tiene 20 preguntas de verdadero/falso. Un estudiante responde al azar (50 % de probabilidad de acierto).

a) ¿Cuál es la probabilidad de acertar exactamente 12 preguntas?
b) ¿Cuál es la probabilidad de acertar 15 o más preguntas?
c) ¿Cuál es el número esperado de aciertos y su desviación estándar?

Mostrar respuesta 2.1 (solo instructor)

Sea $X \sim \text{Binomial}(n = 20, p = 0.5)$

a) P(X = 12):

$$P(X = 12) = \binom{20}{12}(0.5)^{12}(0.5)^{8} = \binom{20}{12}(0.5)^{20}$$

$$= 125970 \times (0.5)^{20} \approx 0.1201$$

b) P(X 15):

$$P(X \geq 15) = \sum_{k=15}^{20} \binom{20}{k}(0.5)^{20}$$

Por tabla o software: $P(X \geq 15) \approx 0.0207$

c) Esperanza y desviación:

$$E[X] = np = 20 \times 0.5 = 10$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{20 \times 0.5 \times 0.5} = \sqrt{5} \approx 2.24$$

E.3.2. Pregunta 2.2 (13 puntos)

El tiempo de atención en una consulta médica sigue distribución normal con media 15 minutos y desviación estándar 3 minutos.

a) ¿Qué porcentaje de consultas duran más de 20 minutos?
b) ¿Cuál es el percentil 90 del tiempo de consulta?
c) Si se observan 100 consultas, ¿cuántas esperamos que duren entre 12 y 18 minutos?

Mostrar respuesta 2.2 (solo instructor)

Sea $X \sim \mathcal{N}(\mu = 15, \sigma = 3)$

a) P(X > 20):

$$Z = \frac{20 - 15}{3} = 1.67$$

$$P(X > 20) = P(Z > 1.67) \approx 0.0475 \text{ o } 4.75\,\%$$

b) Percentil 90:

Necesitamos $z_{0.90} \approx 1.28$ (de tabla normal estándar)

$$x_{90} = 15 + 1.28 \times 3 = 15 + 3.84 = 18.84 \text{ minutos}$$

c) Número de consultas en [12, 18]:

$$Z_1 = \frac{12 - 15}{3} = -1 \quad Z_2 = \frac{18 - 15}{3} = 1$$

$$P(12 < X < 18) = P(-1 < Z < 1) \approx 0.6826 \text{ o } 68.26\,\%$$

Esperadas: $100 \times 0.6826 \approx 68$ consultas

E.4. Parte III: Inferencia Estadística (25 puntos)

E.4.1. Pregunta 3.1 (12 puntos)

Se realizó un estudio de presión arterial sistólica en una muestra de 40 pacientes hipertensos. La media muestral fue 145 mmHg y la desviación estándar muestral 12 mmHg. Asuma distribución normal.

a) Calcule un intervalo de confianza del 95 % para la media poblacional.
b) Interprete el intervalo.
c) ¿Cuál sería el tamaño muestral necesario para un margen de error de 3 mmHg?

Mostrar respuesta 3.1 (solo instructor)

a) Intervalo de confianza 95 %:

Con $n = 40$, $\bar{x} = 145$, $s = 12$, $\alpha = 0.05$:

Grados de libertad: $df = 39$

$t_{0.025,39} \approx 2.023$ (de tabla t)

Margen de error: $E = t_{\alpha/2} \times \frac{s}{\sqrt{n}} = 2.023 \times \frac{12}{\sqrt{40}} = 2.023 \times 1.897 \approx 3.84$

$$IC_{95\,\%} = [145 - 3.84, 145 + 3.84] = [141.16, 148.84] \text{ mmHg}$$

b) Interpretación: Con 95 % de confianza, la presión arterial sistólica media poblacional se encuentra entre 141.16 y 148.84 mmHg. Si repitiéramos el estudio muchas veces, el 95 % de los intervalos construidos contendrían el parámetro verdadero.

c) Tamaño muestral para E = 3:

$$n = \left(\frac{t_{\alpha/2} \times s}{E}\right)^2 = \left(\frac{2.023 \times 12}{3}\right)^2 \approx \left(\frac{24.276}{3}\right)^2 \approx 65.4$$

Se necesitan aproximadamente **n = 66 pacientes**.

E.4.2. Pregunta 3.2 (13 puntos)

Un laboratorio afirma que el 90 % de sus pruebas de COVID tienen sensibilidad 95 %. Para verificar, se prueban 50 muestras positivas y se detectan correctamente 47.

a) Plantee las hipótesis nula y alternativa.
b) Calcule el estadístico de prueba y el p-valor.
c) ¿Rechaza la afirmación del laboratorio al nivel = 0.05?

Mostrar respuesta 3.2 (solo instructor)

a) Hipótesis:

- $H_0 : p = 0.95$ (la sensibilidad es del 95 %)
- $H_A : p < 0.95$ (la sensibilidad es menor al 95 %, test unilateral)

b) Estadístico de prueba:

Proporción muestral: $\hat{p} = \frac{47}{50} = 0.94$

Error estándar bajo H_0: $SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.95 \times 0.05}{50}} \approx 0.0308$

$$z = \frac{\hat{p} - p_0}{SE} = \frac{0.94 - 0.95}{0.0308} \approx -0.325$$

p-valor (unilateral): $P(Z < -0.325) \approx 0.372$

c) Conclusión: Con p-valor = $0.372 >$ $= 0.05$, **no rechazamos** H_0. No hay evidencia suficiente para rechazar la afirmación del laboratorio de que la sensibilidad es del 95 %.

E.5. Parte IV: Regresión Lineal (25 puntos)

E.5.1. Pregunta 4.1 (25 puntos)

Se estudia la relación entre precio de venta (Y, en miles €) y tamaño de la vivienda (X, en m²) en 8 pisos vendidos:

Tamaño (m²)	80	100	90	120	110	95	85	130
Precio (k€)	150	185	170	210	200	180	165	230

a) Estime el modelo de regresión lineal simple $Y = \beta_0 + \beta_1 X + \epsilon$
b) Interprete los coeficientes
c) Calcule R^2 e interprete la calidad del ajuste
d) Prediga el precio para una vivienda de 105 m²

Mostrar respuesta 4.1 (solo instructor)

a) Estimación del modelo:

Primero calculamos estadísticos:

- $\bar{x} = 102.5$, $\bar{y} = 186.25$
- $S_{xx} = \sum (x_i - \bar{x})^2 = 1512.5$
- $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = 3057.5$
- $S_{yy} = \sum (y_i - \bar{y})^2 = 6337.5$

Coeficiente de regresión:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{3057.5}{1512.5} \approx 2.021$$

Ordenada en el origen:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 186.25 - 2.021 \times 102.5 \approx -22.41$$

Modelo estimado:

$$\hat{Y} = -22.41 + 2.021X$$

b) Interpretación de coeficientes: - $\hat{\beta}_0 = -22.41$: Es el precio teórico para X = 0 (sin interpretación práctica, pues no hay viviendas de 0 m²) - $\hat{\beta}_1 = 2.021$: Por cada m² adicional, el precio aumenta aproximadamente 2.021 mil euros, o sea, 2021 € por m²

c) Coeficiente de determinación R²:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{3057.5}{\sqrt{1512.5 \times 6337.5}} \approx 0.9815$$

$$R^2 = r^2 \approx 0.963$$

Interpretación: El 96.3 % de la variabilidad en los precios se explica por el tamaño de la vivienda. El ajuste es excelente.

d) Predicción para X = 105 m²:

$$\hat{Y}(105) = -22.41 + 2.021 \times 105 = -22.41 + 212.205 \approx 189.80 \text{ k€}$$

Precio predicho: aproximadamente **189,800 €** para una vivienda de 105 m².

E.6. Parte V: Análisis de Datos Categóricos (25 puntos)

E.6.1. Pregunta 5.1 (12 puntos)

En un estudio sobre la asociación entre consumo de tabaco y enfermedad cardiovascular se obtiene la siguiente tabla:

	Enfermedad cardiovascular	Sin enfermedad	Total
Fumador	55	95	150
No fumador	25	125	150

a) Calcule las frecuencias esperadas bajo H_0 de independencia.
b) Calcule el estadístico χ^2 e identifique la distribución de referencia.
c) Con $\alpha = 0.05$ ($\chi^2_{0.05,1} = 3.84$), ¿existe asociación estadísticamente significativa?

d) ¿Se cumplen las condiciones de aplicación del test Chi-cuadrado?

Mostrar respuesta 5.1 (solo instructor)

a) Frecuencias esperadas:

Marginales: fila 1 = 150, fila 2 = 150, columna 1 = 80, columna 2 = 220, $n = 300$.

$$E_{11} = \frac{150 \times 80}{300} = 40, \quad E_{12} = \frac{150 \times 220}{300} = 110$$
$$E_{21} = \frac{150 \times 80}{300} = 40, \quad E_{22} = \frac{150 \times 220}{300} = 110$$

b) Estadístico ²:

$$\chi^2 = \frac{(55-40)^2}{40} + \frac{(95-110)^2}{110} + \frac{(25-40)^2}{40} + \frac{(125-110)^2}{110}$$

$$= \frac{225}{40} + \frac{225}{110} + \frac{225}{40} + \frac{225}{110} = 5.625 + 2.045 + 5.625 + 2.045 = 15.34$$

Distribución de referencia: $\chi^2_{(2-1)(2-1)} = \chi^2_1$.

c) Decisión:

$\chi^2 = 15.34 > 3.84 = \chi^2_{0.05,1}$, $p < 0.001$. **Rechazamos** H_0: existe asociación estadísticamente significativa entre tabaquismo y enfermedad cardiovascular.

d) Condiciones: Mínima frecuencia esperada = $40 > 5$. Condición cumplida.

E.6.2. Pregunta 5.2 (13 puntos)

Un estudio caso-control investiga la asociación entre obesidad (IMC 30) e infarto de miocardio. Se reclutan 100 casos (infarto) y 100 controles. De los casos, 60 son obesos; de los controles, 30 son obesos.

a) Construya la tabla 2×2 con la notación estándar (a, b, c, d).
b) Calcule el Odds Ratio (OR).
c) Calcule el intervalo de confianza del 95 % para el OR.
d) Interprete el resultado en contexto clínico. ¿Sería correcto calcular el Riesgo Relativo (RR) en este diseño? Justifique.

Mostrar respuesta 5.2 (solo instructor)

a) Tabla 2×2:

	Caso (infarto)	Control	Total
Obeso	60	30	90

	Caso (infarto)	Control	Total
No obeso	40	70	110
Total	100	100	200

$a = 60$, $b = 30$, $c = 40$, $d = 70$.

b) Odds Ratio:

$$\widehat{OR} = \frac{ad}{bc} = \frac{60 \times 70}{30 \times 40} = \frac{4200}{1200} = 3.50$$

c) Intervalo de confianza del 95 %:

$$SE_{\ln OR} = \sqrt{\frac{1}{60} + \frac{1}{30} + \frac{1}{40} + \frac{1}{70}} = \sqrt{0.0167 + 0.0333 + 0.0250 + 0.0143} = \sqrt{0.0893} = 0.299$$

$$IC_{95\,\%} = \exp(\ln 3.50 \pm 1.96 \times 0.299) = \exp(1.253 \pm 0.586) = [e^{0.667},\, e^{1.839}] = [1.95,\, 6.29]$$

d) Interpretación: Las personas obesas tienen **3.5 veces más odds** de infarto de miocardio que las no obesas ($OR = 3.50$; $IC_{95\,\%}$: 1.95–6.29). El intervalo no contiene el 1, confirmando asociación estadísticamente significativa.

RR en caso-control: No es correcto calcular el RR porque en un estudio caso-control el número de casos y controles es fijado por el investigador (100 y 100 respectivamente), y no refleja la prevalencia o incidencia real de infarto en la población. Los totales de fila no son proporcionales a los totales poblacionales, por lo que las "proporciones" calculadas como $a/(a + c)$ no son estimaciones válidas del riesgo. La única medida estimable de forma válida en este diseño es el OR.

E.7. Instrucciones Finales

- Revise sus respuestas antes de entregar
- Indique claramente los pasos de sus cálculos
- Si usa software o calculadora, muestre los resultados intermedios
- La precisión y claridad contribuyen a la calificación

Fin del examen