

TEORÍA ESTADÍSTICA:
APLICACIONES Y MÉTODOS

Prólogo

Sobre teoría estadística se han escrito muchos libros, indudablemente más en el concierto internacional que en el nacional. Sin embargo, cada vez que un lector se enfrenta a una nueva publicación sobre el tema, él quisiera detectar qué es lo nuevo, diferente o atractivo que se presenta o desarrolla en la obra que tiene en sus manos. Desde esta premisa, es muy agradable presentar este libro en el cual se marcan diferencias importantes con respecto a muchos otros escritos sobre la materia. En las líneas siguientes explicaré estas características significativas, para usar un término muy «estadístico».

En virtud de la gran experiencia y habilidad en el manejo del lenguaje R por parte de los autores, el libro incluye muchos ejemplos ilustrativos de los conceptos fundamentales de la inferencia estadística, los cuales se han desarrollado con este lenguaje. Esto permite al lector comprender, por ejemplo entre muchas otras, la noción intuitiva de distribución muestral (o de muestreo). Se incluye la teoría estadística básica de la inferencia multivariada, crucial en el entendimiento del comportamiento probabilístico de un vector de variables aleatorias y de las relaciones entre ellas. No es usual encontrar un trabajo en donde se incluyan conjuntamente, los contextos univariado y multivariado de la inferencia estadística.

Este libro es un buen punto de partida para el conocimiento e interiorización de la teoría estadística, por parte de estudiantes de una carrera de estadística, en el entendido de hacer de la práctica estadística una profesión. Además, podrá ser un gran soporte para la realización de estudios de posgrado, bien sea a nivel de profundización de conocimientos o a nivel de investigación.

En forma muy general, se puede afirmar que en la presente obra, la teoría y sus aplicaciones son presentadas de manera muy coherente y equilibrada; es decir, sin profundizar en lo teórico más allá de lo necesario y sin exagerar en la inclusión de las aplicaciones. Por esto y todo lo expresado anteriormente, me siento muy complacido de presentar este libro y de recomendarlo a un amplio conglomerado de lectores o usuarios de la estadística.

Fabio Nieto, Ph.D.

Profesor titular

Departamento de Estadística

Universidad Nacional de Colombia

Prefacio

La estadística es una herramienta poderosa en manos del investigador y del profesional. En la vida práctica del profesional que utiliza la estadística, es bien sabido que no es posible realizar un trabajo apropiado sin tener las bases precisas que permitan el entendimiento y comprensión de las bases de esa herramienta. La teoría estadística da esas bases y este libro pretende ser un camino que permita el entendimiento y comprensión de las mismas.

Con el pasar del tiempo, el pensamiento estadístico se está convirtiendo en una cultura. Es una forma de razonar que los profesionales de la mayoría de las áreas del saber deben tener para ejercer exitosamente sus trabajos aplicados concernientes al análisis de datos. Y precisamente, el comienzo de la formación de este pensamiento se da en un curso de inferencia. Por lo anterior, es muy importante que el estudiante termine el curso estando preparado de la mejor forma posible, y para esto es fundamental contar con materiales adecuados para el aprendizaje tanto de la teoría como de los métodos prácticos. En el ámbito estadístico existen abundantes textos sobre el tema de inferencia, algunos desarrollando rigurosamente las teorías estadísticas formales; por otro lado, también se encuentran textos de la teoría estadística para disciplinas como ingeniería, veterinaria o biología, entre otras. Estos textos van dirigidos directamente a los usuarios finales de la estadística, algunos de ellos con muy poca formación en matemáticas, y por consiguiente los textos no son teóricos sino que se enfocan en la aplicación de la estadística en la vida real. Con este texto, nosotros quisimos centrarnos entre estas corrientes y hacer de la práctica un resultado directo del conocimiento teórico.

Este libro nació del curso de inferencia estadística que dictamos en la Facultad de Estadística de la Universidad Santo Tomás en Bogotá, Colombia. Allí surgió la necesidad de fusionar la teoría y la práctica de una manera óptima. Por esta razón en este libro nosotros pretendemos abarcar ambos aspectos de la estadística: la teoría y la aplicación práctica. Por un lado, se desarrolla la teoría de la inferencia estadística con un alto grado de rigurosidad, y por otro lado se ilustra cómo es la aplicación de éstas técnicas y métodos en la práctica. Hemos optado por el uso del programa estadístico R para ilustrar mediante gráficas y códigos las teorías expuestas; de la misma forma, la mayor parte de los cálculos en las aplicaciones también fueron realizados en este software. Algunos de estos códigos computacionales están disponibles en el libro, otros, por limitación de espacio, no fueron incluidos pero los proveeremos en caso de ser requeridos.

El contenido del libro se divide principalmente en tres partes. A saber, inferencia univariada, tal como su nombre indica, estudia características de una variable aleatoria observada en una muestra; inferencia multivariada, que estudia conjuntamente varias variables y por último los apéndices que complementan el proceso natural del análisis de datos. Aunque las partes tienen temas y técnicas similares, se pueden considerar como dos partes independientes.

Este texto va dirigido a los profesionales y estudiantes que deban utilizar herramientas de inferencia estadística, y puede servir como libro guía para un curso de cuatro meses con intensidad horaria de seis horas por semana, si se desea abarcar tanto la parte univariada como la parte multivariada. Es importante aclarar que, dependiendo del curso y el énfasis, el docente debe enfocarse en la parte relevante del curso realizando algunas demostraciones teóricas, sin pretender cubrir todos los temas del libro, pero siempre enfatizando la aplicabilidad de los resultados teóricos y la relación estrecha que existen entre el sentido común y los resultados encontrados.

También el texto puede servir para un curso introductorio de estadística en un programa de especialización o maestría donde la mayoría de los estudiantes no son de profesión estadística, y necesitan simultáneamente adquirir formas de pensamiento estadístico y técnicas de análisis de datos en la práctica.

Al final de cada capítulo, se provee de ejercicios sobre el tema desarrollado en las distintas secciones que lo conforman. Algunos de estos ejercicios son teóricos y exigen que los estudiantes estén familiarizados con las herramientas vistas para derivar resultados. Otros ejercicios son de carácter práctico donde se describe un problema de la vida real que debe ser resuelto utilizando el sentido común, seguido del pensamiento estadístico y, entendiendo el contexto del problema, mediante el planteamiento de las preguntas que se deben resolver para, por último, aplicar las herramientas estadísticas apropiadas.

Agradecemos en primer lugar a Dios que nos dio la motivación y la perseverancia para escribir este libro, también el apoyo que encontramos en la Universidad Santo Tomás por medio del Centro de Investigaciones y Estudios Estadísticos (CIEES) y mediante la siempre elegante gestión administrativa de Sander Rangel en la decanatura de la Facultad de Estadística. Además agradecemos a los estudiantes del curso inferencia estadística de la Universidad Santo Tomás que colaboraron con la corrección del libro. De la misma manera, agradecemos a los profesores Yesid Rodríguez de la Universidad Santo Tomás y Sergio Calderón Villanueva y Luis Guillermo Díaz de la Universidad Nacional por los valiosos comentarios sobre las notas.

Los autores aclaran que la responsabilidad por los errores que pueden haber en el libro es única y exclusivamente de ellos y agradecen los comentarios, las correcciones y las posibles críticas constructivas sobre la obra. Éste es un producto del grupo de investigación en Muestreo y Marketing, adscrito al Centro de Investigaciones y Estudios Estadísticos (CIEES) de la Facultad de Estadística de la Universidad Santo Tomás.

Contenido

| | |
|---|-----------|
| Prólogo | i |
| Prefacio | iii |
| I Inferencia estadística univariada | 1 |
| 1 Conceptos preliminares | 3 |
| 1.1 Variables aleatorias y distribuciones de probabilidad | 4 |
| 1.1.1 Distribuciones discretas | 6 |
| 1.1.2 Distribuciones continuas | 25 |
| 1.1.3 Percentiles | 55 |
| 1.2 Familia exponencial | 58 |
| 1.2.1 Familia exponencial uniparamétrica | 58 |
| 1.2.2 Familia exponencial multi-paramétrica | 60 |
| 1.3 Ejercicios | 61 |
| 2 Estimación puntual | 65 |
| 2.1 Introducción | 65 |
| 2.2 Conceptos básicos | 66 |
| 2.3 Estimaciones puntuales | 68 |
| 2.3.1 Método de máxima verosimilitud | 68 |
| 2.3.2 Método de los momentos | 86 |
| 2.3.3 Método de mínimos cuadrados | 99 |
| 2.4 Propiedades de estimadores puntuales | 100 |
| 2.4.1 Error cuadrático medio | 100 |
| 2.4.2 Suficiencia | 110 |
| 2.4.3 Estimadores UMVUE | 117 |
| 2.4.4 Completez | 131 |
| 2.4.5 Consistencia | 139 |
| 2.5 Comparación empírica de algunas propiedades | 140 |
| 2.6 Ejercicios | 144 |

| | | |
|-----------|---|------------|
| 3 | Estimación por intervalo de confianza | 151 |
| 3.1 | Introducción | 151 |
| 3.2 | Bajo normalidad | 154 |
| 3.2.1 | Problemas de una muestra | 154 |
| 3.2.2 | Problemas de dos muestras | 185 |
| 3.3 | Bajo distribuciones diferentes a la normal | 197 |
| 3.3.1 | Intervalos de confianza con distribución exponencial | 198 |
| 3.3.2 | Intervalos de confianza con distribución Bernoulli | 204 |
| 3.3.3 | Intervalos de confianza con distribución Poisson | 207 |
| 3.4 | Ejercicios | 208 |
| 4 | Pruebas de hipótesis | 213 |
| 4.1 | Conceptos preliminares | 213 |
| 4.2 | Una muestra bajo normalidad | 215 |
| 4.2.1 | Pruebas de hipótesis para la media poblacional | 215 |
| 4.2.2 | Pruebas de hipótesis acerca de la varianza poblacional | 247 |
| 4.3 | Dos muestras | 254 |
| 4.3.1 | Comparación entre dos medias | 254 |
| 4.3.2 | Comparación entre dos varianzas | 262 |
| 4.4 | k muestras | 267 |
| 4.4.1 | Igualdad de dos medias | 267 |
| 4.4.2 | Igualdad de varianzas | 271 |
| 4.5 | Muestras provenientes de la distribución Bernoulli y binomial | 273 |
| 4.5.1 | Una muestra | 273 |
| 4.5.2 | Dos muestras | 281 |
| 4.6 | Muestras provenientes de una distribución Poisson | 285 |
| 4.6.1 | Una muestra | 285 |
| 4.6.2 | Dos muestras | 288 |
| 4.7 | Muestras provenientes de la distribución exponencial | 291 |
| 4.7.1 | Una muestra | 291 |
| 4.7.2 | Dos muestra | 296 |
| 4.8 | Acerca del p -valor | 297 |
| 4.8.1 | Diversos puntos de vistas acerca del p -valor | 297 |
| 4.8.2 | p valores aleatorios | 299 |
| 4.8.3 | p valor no es una medida de soporte | 303 |
| 4.8.4 | Igualdad en la hipótesis nula | 304 |
| 4.9 | Ejercicios | 306 |
| II | Inferencia estadística multivariante | 309 |
| 5 | Distribuciones multivariantes | 311 |
| 5.1 | Vectores aleatorios | 311 |
| 5.2 | Algunas distribuciones multivariantes | 321 |
| 5.2.1 | Distribución multinomial | 321 |
| 5.2.2 | Distribución normal multivariante | 322 |

| | | |
|----------|---|------------|
| 5.2.3 | Distribución Wishart | 334 |
| 5.2.4 | Distribución T^2 de Hotelling | 336 |
| 5.3 | Ejercicios | 336 |
| 6 | Inferencia multivariante | 339 |
| 6.1 | Inferencia en la distribución multinomial | 340 |
| 6.1.1 | Una muestra | 340 |
| 6.1.2 | Dos muestras | 345 |
| 6.1.3 | k muestras | 349 |
| 6.2 | Inferencia en la distribución normal multivariante | 351 |
| 6.2.1 | Estimador de máxima verosimilitud | 351 |
| 6.2.2 | Propiedades de los estimadores de máxima verosimilitud | 355 |
| 6.3 | Región de confianza y pruebas de hipótesis para el vector de medias | 358 |
| 6.3.1 | Σ conocida | 359 |
| 6.3.2 | Σ desconocida | 362 |
| 6.4 | Inferencia para una combinación lineal de medias | 364 |
| 6.5 | Juzgamiento de hipótesis para la matriz de varianzas y covarianzas | 368 |
| 6.6 | Ejercicios | 375 |
| A | Breve historia del desarrollo estadístico | 379 |
| B | Herramientas de bondad de ajuste | 389 |
| B.1 | Gráficas QQ plot | 389 |
| B.1.1 | QQ plot para una distribución exponencial | 390 |
| B.1.2 | QQ plot para una distribución normal | 392 |
| B.1.3 | QQ plot para una distribución Weibull | 395 |
| B.1.4 | QQ plot para una distribución Gamma pos-estimación | 397 |
| B.1.5 | QQ plot para una distribución Beta pos-estimación | 399 |
| B.1.6 | QQ plot para Normal multivariante | 402 |
| B.2 | Pruebas de bondad de ajuste | 403 |
| B.2.1 | Prueba de normalidad de Shapiro-Wilk | 404 |
| B.2.2 | Prueba de Kolmogorov Smirnov | 405 |
| B.2.3 | Prueba de Mardia | 407 |
| C | Transformación de Box-Cox | 411 |
| C.1 | Definición de la transformación Box-Cox | 411 |
| C.2 | Casos particulares de la transformación Box-Cox | 411 |
| C.2.1 | Transformación logarítmica | 412 |
| C.2.2 | Transformación raíz cuadrada | 412 |
| C.3 | Estimación de máxima verosimilitud de λ | 413 |
| D | Repaso matricial | 417 |
| D.1 | Matriz y vector | 417 |
| D.2 | Suma y producto entre matrices | 418 |
| D.3 | Transpuesta de una matriz | 418 |
| D.4 | Determinante | 419 |

| | |
|--|------------|
| D.5 Inversa | 420 |
| D.6 Traza | 420 |
| D.7 Valores y vectores propios | 420 |
| D.8 Formas cuadráticas y matices semidefinidas | 421 |
| D.9 Descomposición espectral y raíz cuadrada de una matriz | 421 |
| D.10 Matriz particionada | 421 |
| D.11 Derivadas matriciales | 422 |
| E Inferencia en tablas de contingencia | 425 |
| F Tablas de percentiles de distribuciones | 429 |

Índice alfabético

- Corrección de continuidad, 274
- Cota de Cramer-Rao, 126
- Desigualdad de Cramer-Rao, 126
 - estimador insesgado, 128
- Distancia de Mahalanobis, 323
- Distribución
 - Bernoulli, 9
 - Beta, 52
 - binomial, 11
 - chi cuadrado, 43
 - exponencial, 33
 - F, 50
 - Gamma, 28
 - hipergeométrica, 13
 - multinomial, 321
 - propiedades, 322
 - normal, 37
 - normal estándar, 40
 - normal multivariante, 322
 - esperanza condicional, 332
 - propiedades, 329, 332
 - Poisson, 17
 - t-student, 47
 - t-student no central, 48
 - T2 de Hotelling, 336
 - uniforme continua, 25
 - uniforme discreta, 6
 - Weibull, 35
 - Wishart, 334
 - propiedades, 335
- Eficiencia relativa, 143
- Error
 - tipo I, 214
 - tipo II, 214
- Error cuadrático medio, 101
- Espacio paramétrico, 6, 213
 - alterno, 213
 - nulo, 213
- Esperanza condicional, 116, 117
- Estadística, 66
 - auxiliar, 133
- Estadística de prueba
 - normal
 - media poblacional, 218
- Estadística de prueba, 214
- Estimación, 67
- Estimador, 67
 - asintóticamente insesgado, 104
 - completo, 131
 - Bernoulli, 131
 - familia exponencial, 132
 - uniforme, 132
 - consistente, 139, 140
 - invarianza, 139
 - media muestral, 139
- de máxima verosimilitud, 70, 339
 - Bernoulli, 71
 - exponencial, 71
 - Gamma, 79
 - hipergeométrica, 79
 - invarianza, 82
 - multinomial, 340, 341, 345
 - normal, 74
 - normal multivariante, 351, 355
 - Poisson, 70
 - uniforme, 81
- de mínimos cuadrados, 99, 355
 - media poblacional, 99
- de momentos, 86
 - Beta, 92
 - Gamma, 89

- invarianza, 96
 - media poblacional, 86
 - normal, 87
 - Poisson, 87
 - uniforme, 96, 98
 - varianza poblacional, 86
- función del parámetro, 135
- insesgado, 101
 - media muestral, 102
- sobreestimación, 101
- subestimación, 101
- suficiente, 110
 - Bernoulli, 113–115
 - Beta, 113
 - en familia exponencial, 114, 115
 - exponencial, 113
 - normal, 113, 115
 - Poisson, 110, 112
- UMVUE, 128, 133, 134
 - normal, 134
 - Poisson, 129, 134
- Familia exponencial
 - multi-paramétrica, 60
 - uniparamétrica, 58
- Función de potencia, 223
 - exponencial
 - una muestra, 292
 - normal
 - igualdad de dos medias, 257
 - igualdad de dos varianzas, 265
 - media poblacional, 241, 243
 - media poblacional, 224, 231, 232
 - varianza poblacional, 248, 252, 253
- Función de verosimilitud, 69, 339
- Gráficas QQ plot, 73, 389
 - distribución Beta, 399
 - distribución exponencial, 390
 - distribución Gamma, 397
 - distribución normal, 76, 392
 - distribución normal multivariante, 402
 - distribución Weibull, 395
- Hipótesis
 - alterna, 213
 - compuesta, 217
 - nula, 213
 - simple, 217
- Información de Fisher
 - binomial, 121
 - en una muestra, 119
 - en una variable, 118
 - estimador suficiente, 124
 - normal, 120
 - Poisson, 122
- Intervalo de confianza
 - Bernoulli, 204
 - Agresti-Caffo, 206
 - Newcombe, 207
 - Wald, 205, 206
 - bilateral, 151
 - escogencia, 153
 - exponencial
 - aproximado, 201
 - exacto, 198
 - función del parámetro, 162
 - longitud, 153
 - esperada, 153
 - varianza, 153
 - normal, 154
 - cociente de varianzas, 193, 195
 - coeficiente de variación, 181
 - diferencia de medias, 186, 188, 191
 - media poblacional, 154, 164, 166, 169
 - varianza poblacional, 172, 173, 178
 - Poisson, 207
 - unilateral, 152
- Lema de Neyman-Pearson, 236
- Máximo de una muestra, 66
 - función de densidad, 108
 - función de distribución, 108
- Método de Delta, 136
- Método de la variable pivote, 154
- Método de los momentos, 86

- Método de máxima verosimilitud, 68
- Método de mínimos cuadrados, 99
- Mínimo de una muestra, 66
- Matriz, 417
- Matriz de correlación, 317
 - propiedades, 318
- Matriz de información, 122
 - normal, 122
- Matriz de varianzas y covarianzas, 315
 - propiedades, 316
- Momento, 86
- Momento muestral, 86
- Muestra aleatoria, 66
- Multiplicador de Lagrange, 156, 340
- Nivel de confianza, 151
- Nivel de significación, 216
- p valor, 219, 297
 - Bernoulli
 - dos muestras, 282
 - una muestra, 273, 276, 279
 - exponencial
 - una muestra, 291, 292
 - multinomial
 - k muestras, 350
 - dos muestras, 347
 - una muestra, 342
 - normal
 - igualdad de k medias, 270
 - igualdad de dos medias, 257
 - igualdad de dos varianzas, 263
 - media poblacional, 240, 243
 - media poblacional, 222, 229, 232
 - varianza poblacional, 248, 251, 253, 254
 - normal multivariante
 - independencia, 371
 - Poisson
 - dos muestras, 289
 - prueba binomial, 276
- Parámetro, 6
 - de escala, 197
 - de localización, 197
- Percentil, 55
- Probabilidad de cobertura, 151
- Prueba
 - de razón generalizada de verosimilitudes
 - multinomial, 346
 - normal, 267, 271
 - normal multivariante, 368, 370, 372
 - Poisson, 285, 289
 - binomial, 275
 - de Bartlett, 271
 - de chi cuadrado de independencia, 284, 426
 - de razón de verosimilitud, 233
 - distribución asintótica, 278
 - normal, 235, 250, 253
 - regla de decisión, 234
 - de razón generalizada de verosimilitudes
 - distribución asintótica, 278
 - de razón generalizada de verosimilitudes, 237
 - Bernoulli, 278
 - exponencial, 292, 297
 - multinomial, 341
 - normal, 238, 254, 259
 - exacta de Fisher, 284, 425
- Prueba de bondad de ajuste, 403
 - Kolmogorov-Smirnov, 405
 - Mardia, 407
 - Shapiro-Wilk, 404
- Prueba de hipótesis, 213
 - Bernoulli, 273, 281
 - dos muestras, 281
 - una muestra, 273, 275
 - exponencial, 291
 - dos muestras, 296
 - una muestra, 291
 - multinomial
 - k muestras, 349
 - dos muestras, 345, 346
 - una muestra, 341
 - normal, 254
 - igualdad de k medias, 267
 - igualdad de k varianzas, 271
 - igualdad de dos medias, 254, 259, 262

- igualdad de dos varianzas, 262
 - media poblacional, 216, 228, 231, 239, 242
 - varianza poblacional, 247, 250, 253
 - varianza poblacional, 252
 - normal multivariante, 362
 - combinación lineal de medias, 364
 - independencia, 370, 372
 - matriz de varianzas, 368
 - vector de medias, 364
 - Poisson, 285
 - dos muestras, 288
 - una muestra, 285
 - Tamaño, 216
- Región de confianza
- normal multivariante
 - vector de medias, 358, 360, 364
- Región de rechazo, 219
- normal
 - media poblacional, 240
 - media poblacional, 219, 229
 - varianza poblacional, 251
- Regla de decisión, 214
- Bernoulli
 - dos muestras, 282
 - una muestra, 273, 279
 - exponencial
 - dos muestras, 297
 - una muestra, 291, 292
 - multinomial
 - k muestras, 350
 - dos muestras, 347
 - una muestra, 342
 - normal
 - igualdad de k medias, 269
 - igualdad de k varianzas, 271, 272
 - igualdad de dos medias, 256, 261, 262
 - igualdad de dos varianzas, 262
 - media poblacional, 239, 243
 - media poblacional, 218
 - varianza poblacional, 248, 251–253
 - normal multivariante
 - combinación lineal de medias, 365, 366
 - independencia, 371, 374
 - matriz de varianzas, 369
 - Poisson
 - dos muestras, 289
 - una muestra, 285, 286
 - Regla de decisión
 - normal
 - media poblacional, 232
 - Regla de decisión
 - normal
 - igualdad de dos varianzas, 263
 - Sesgo, 101
 - Tablas de contingencia, 425
 - Teorema
 - de Basu, 133
 - de factorización de Fisher-Neyman, 111, 113
 - del límite central, 41
 - Rao-Blackwell, 116
 - Transformación
 - Box-Cox, 411
 - logarítmica, 412
 - raíz cuadrada, 412
 - Variable pivote, 154, 166, 198
 - normal, 154
 - Vector, 417
 - Vector aleatorio, 311
 - esperanza, 313
 - función de densidad, 312
 - función de densidad marginal, 313
 - función de distribución, 312
 - función generadora de momentos, 319
 - independencia, 319