Análisis de datos en psicología y educación

Análisis de datos en psicología y educación

Ana María Ruiz-Ruano García
Jorge López Puga

Mc Graw Hill | AULAMAGNA
PROYECTO CLAVE

Análisis de datos en psicología y educación

Primera edición: 2022

ISBN: 9788419187178
ISBN eBook: 9788419187574
Depósito Legal: SE 1115-2022

© del texto:
Ana María Ruiz-Ruano García
Jorge López Puga

© de esta edición:
Editorial Aula Magna, 2022. McGraw-Hill Interamericana de España S.L.
editorialaulamagna.com
info@editorialaulamagna.com

Impreso en España – Printed in Spain

Quedan prohibidos, dentro de los límites establecidos en la ley y bajo los apercibimientos legalmente previstos, la reproducción total o parcial de esta obra por cualquier medio o procedimiento, ya sea electrónico o mecánico, el tratamiento informático, el alquiler o cualquier otra forma de cesión de la obra sin la autorización previa y por escrito de los titulares del copyright. Diríjase a info@editorialaulamagna.com si necesita fotocopiar o escanear algún fragmento de esta obra.

Índice general

1 Introducción al análisis de datos **1**
- 1.1 Terminología esencial 1
- 1.2 Tipos de variables 6
- 1.3 Muestreo 11

2 Nociones de probabilidad **15**
- 2.1 Sucesos aleatorios 16
- 2.2 La probabilidad 18
- 2.3 Ventajas probabilísticas u *odds* 24
- 2.4 Cálculos de probabilidad 26
  - 2.4.1 Probabilidad de la unión de sucesos 26
  - 2.4.2 Probabilidad condicional 28
  - 2.4.3 Sucesos independientes 30
  - 2.4.4 Intersección de sucesos 31
  - 2.4.5 Teorema de la probabilidad total 33
  - 2.4.6 Teorema de Bayes 35

3 Combinatoria **39**
- 3.1 Variaciones sin repetición 40
- 3.2 Variaciones con repetición 42
- 3.3 Permutaciones sin repetición 43
- 3.4 Permutaciones con repetición 44
- 3.5 Combinaciones 45

4 Análisis estadístico y ciencia **47**
- 4.1 El método científico 48
- 4.2 El estado actual de la ciencia 56

5 Distribuciones de frecuencias **67**

5.1 Frecuencias absolutas . . . 68
5.2 Frecuencias relativas . . . 70
5.3 Frecuencias absolutas acumuladas . . . 71
5.4 Frecuencias relativas acumuladas . . . 73
5.5 Representación gráfica . . . 75
5.5.1 Gráfico de barras . . . 76
5.5.2 Gráfico de sectores . . . 76
5.5.3 Gráfico de puntos . . . 80
5.5.4 Histograma . . . 81
5.5.5 Polígono de frecuencias . . . 83
5.5.6 Polígono de frecuencias acumuladas . . . 84
5.5.7 Diagramas de líneas . . . 84
5.5.8 Diagrama de tallos y hojas . . . 87
5.6 Forma de las distribuciones de frecuencias . . . 90

6 Estadísticos de posición **93**
6.1 Idea de cuantil o *cuantila* . . . 94
6.2 Percentiles . . . 96
6.3 Otros cuantiles . . . 101
6.4 El diagrama de caja . . . 103
6.5 Diagramas de violín y frijoles . . . 107

7 Tendencia central **113**
7.1 Moda . . . 114
7.2 Mediana . . . 116
7.3 Media aritmética . . . 117
7.3.1 Cálculo vectorial . . . 119
7.3.2 Media aritmética ponderada . . . 120
7.3.3 Propiedades de la media . . . 121
7.4 Representación gráfica . . . 125

8 Estadísticos de dispersión **131**
8.1 Rango y amplitud . . . 133
8.2 Variaciones cuartílicas . . . 135
8.3 Desviación media . . . 139
8.4 Varianza . . . 142
8.4.1 Cálculo vectorial . . . 143
8.4.2 Varianza ponderada . . . 145
8.5 Desviación típica . . . 147

8.6 Coeficiente de variación 150
8.7 Representación gráfica 151

9 Estadísticos de forma **157**
9.1 Sesgo o asimetría 158
9.2 Curtosis 163
9.3 Representación gráfica 168

10 Transformación de variables **171**
10.1 Puntuaciones diferenciales 172
10.2 Puntuaciones típicas 175
10.3 Puntuaciones derivadas 178
10.4 Corrección del sesgo 180

11 Asociación, correlación y causación **183**
11.1 Asociación nominal 188
11.1.1 Tablas de contingencia y gráficos 188
11.1.2 Ji-cuadrado de Pearson 193
11.1.3 V de Cramer 201
11.2 Correlación lineal 202
11.2.1 Gráfico de dispersión 203
11.2.2 r de Pearson 205
11.2.3 Variables ordinales 217
11.2.4 Estimación puntual biserial 219
11.2.5 Variables dicotómicas 222
11.2.6 Miscelánea 225
11.3 Correlación parcial y semiparcial 232

12 Clasificación diagnóstica **237**
12.1 Matriz de confusión 238
12.2 Tasa de clasificaciones correctas 241
12.3 Sensibilidad y especificidad 242
12.4 Valores predictivos 244
12.5 Otros estadísticos 248
12.6 Razones de verosimilitud 252
12.7 Curvas ROC 254

Apéndices **260**

A Aximoas de la probabilidad según Kolmogorof **261**

B Un caso del teorema de la probabilidad total **265**

C Algunos cuantiles muestrales de proporción **269**

D Terremotos **273**

E Sumatorios y productorios **275**
E.1 Sumatorios 275
E.2 Productorios 277

F Valores predictivos positivos y negativos **279**

Referencias **285**

Índice de figuras

1.1 Escalas de medida según Stevens 9

2.1 Diagramas de Venn representando diferentes operaciones con sucesos 19
2.2 La frecuencia relativa como probabilidad de obtener una cara en diferente número de lanzamientos al lanzar una moneda 22
2.3 Relación entre probabilidad y *odds*. 25
2.4 Ejemplo de sucesos incompatibles y compatibles 27
2.5 Intersección de sucesos compatibles y universo reducido usando la probabilidad condicional 29
2.6 Representación gráfica de un experimento aleatorio con sucesos relacionados. 32
2.7 Teorema de la Probabiliad Total. 34

3.1 Variaciones con repetición al lanzar una moneda tres veces. 43

4.1 Ciclo simplificado del método científico 50
4.2 Esquema simplificado del sistema de publicación científica para una revista concreta 60
4.3 Diagrama de flujo en un trabajo científico pre-registrado . 65

5.1 Temperaturas máximas y producto estrella sustraído de un bazar en el mes de abril de 2018 en un barrio de la ciudad de Murcia 69
5.2 Gráfico de barras en dos formatos diferentes 77
5.3 Gráfico de sectores 77
5.4 Evaluación de la proporción 1 : 3 en gráficos de sectores y gráficos de barras 78
5.5 Magnitudes que son claramente percibidas y no percibidas en gráficos de barras y de sectores 79
5.6 Dificultad de los gráficos de sectores para clarificar los patrones más simples 80
5.7 Comparación de un gráfico de sectores y un gráfico de puntos 81
5.8 Ejemplos de histogramas 83
5.9 Ejemplos de polígonos de frecuencias 84
5.10 Ejemplos de polígonos de frecuencias acumuladas 85
5.11 Efecto del rango del eje y en un gráfico de líneas 86
5.12 Gráfico de líneas y de pasos con eje partido 87
5.13 Diámetro en centímetros de 30 flores de la familia de las malváceas y su correspondiente diagrama de tallos y hojas 89
5.14 Diámetro en centímetros de 30 flores (tipo B) de la familia de las malváceas y diagramas de tallos y hojas pareados . 90
5.15 Algunos tipos de distribuciones de frecuencias en histogramas 91

6.1 Algunos cuantiles importantes 102
6.2 Gráfico de puntos para la magnitud de 90 terremotos. . . 104
6.3 El diagrama de caja 106
6.4 Gráfico de puntos, diagrama de caja, diagrama de violín y gráfico de frijoles utilizados para comparar grupos de puntuaciones 108
6.5 Comparación entre histograma, gráfico de caja y diagrama de violín 109

6.6 Comparación del gráfico de frijoles asimétrico y diagramas de caja 110

7.1 Temperaturas máximas en un barrio murciano a lo largo de más de cuatro años 124
7.2 Comparación de la media, mediana y moda en el muestreo repetido 125
7.3 Gráfico de barras y de líneas para representar medias . . . 126
7.4 Diferentes distribuciones que generan el mismo gráfico de medias utilizando el diagrama de barras 127

8.1 Representación química de la dispersión 133
8.2 Amplitud intercuartílica representada en diagramas de caja 136
8.3 Efecto del cambio de escala en la amplitud intercuartílica 138
8.4 Barras de error 153
8.5 Problemas al usar barras de error en gráficos de barras . . 153
8.6 Gráfico de línea con barras de error en un estudio de caso único para tratar el hábito tabáquico 155

9.1 Histograma y gráfico de densidad de terremotos 159
9.2 Histogramas y gráficos de barras considerando las marcas de clase para tres tipos diferentes de de asimetría 161
9.3 Comportamiento de la media, mediana y moda en función del sesgo 163
9.4 Varable mesocúrtica, leptocúrtica y platicúrtica 165
9.5 Curtosis positiva y negativa en igualdad de varianzas respecto a la distribución normal 167
9.6 Curvas de Kaplansky para contrastar la curtosis y el apuntamiento frente a la distribución normal 168
9.7 Efecto del sesgo en el gráfico de caja 169

10.1 Corrección del sesgo positivo 181
10.2 Corrección del sesgo negativo 182

11.1 Modelos causales alternativos que podrían explicar la asociación entre consumo de café y cáncer de piel 186
11.2 Gráfico de barras agrupadas y apiladas 192
11.3 Gráfico de puntos y de mosaico para tablas de contingencia 193
11.4 Gráfico de dispersión 204

11.5 Gráfico de dispersión basado en puntos opacos y en círculos huecos 205
11.6 Efecto del grado de correlación en el gráfico de dispersión 215
11.7 Conjunto de Anscombe 216
11.8 Relación lineal sin covariación 217
11.9 Matriz de gráficos de dispersión 228
11.10Correlogramas 230
11.11Modelo hipotético de relación que se establece entre el consumo de café, el cáncer de piel y la exposición al sol . 234

12.1 Matriz de confusión 240
12.2 Curva ROC 258

B.1 Teorema de la Probabilidad Total en una enfermedad. . . 266

Índice de tablas

5.1 Frecuencia absoluta de las temperaturas máximas diarias en un barrio murciano en el mes de abril de 2018 70
5.2 Número de artículos sustraídos con más frecuencia en un bazar de un barrio murciano en el mes de abril de 2018 . . 70
5.3 Frecuencia relativa de las temperaturas máximas diarias en un barrio murciano en el mes de abril de 2018 72
5.4 Frecuencia relativa de artículos sustraídos con más frecuencia en un bazar de un barrio murciano en el mes de abril de 2018 72
5.5 Frecuencia absoluta acumulada de las temperaturas máximas diarias en un barrio murciano en el mes de abril de 2018 73
5.6 Frecuencia absoluta acumulada de las temperaturas máximas diarias en un barrio murciano en el mes de abril de 2018 75

5.7 Temperaturas máximas diarias en un barrio murciano en el mes de abril de 2018 agrupadas en intervalos 82

6.1 Número de golosinas consumidas por diez niños en un día del fin de semana 95
6.2 Número de golosinas consumidas por diez niños en un día del fin de semana ordenadas de menor a mayor 95
6.3 Marcas de diez atletas en la prueba de 100 metros lisos . . 99
6.4 Marcas de diez atletas expresadas en centésimas de segundo ordenadas de menor a mayor con su correspondiente estadístico de orden 100

7.1 Número de notificaciones recibidas por dos personas en su red social preferida a lo largo de una semana 115
7.2 Cáluculo de la media ponderada 122

8.1 Número de tormentas tropicales a lo largo seis meses en dos regiones intertropicales 140
8.2 Cálculo de la desviación media 141
8.3 Cáluculo de la varianza ponderada 147

9.1 Distribuciones de frecuencias para tres variables que difieren en grado de asimetría 160

10.1 Número de emergencias acaecidas en un ambulatorio a lo largo de una semana 173
10.2 Puntuación diferencial respecto a la media para el número de emergencias acaecidas en un ambulatorio a lo largo de una semana 174
10.3 Puntuación diferencial y típcia respecto a la media para el número de emergencias acaecidas en un ambulatorio a lo largo de una semana 177
10.4 Puntuación derivada con media 100 y desviación típica 15 para el número de emergencias acaecidas en un ambulatorio a lo largo de una semana 179

11.1 Tabla de contingencia para frecuencias absolutas conjuntas 189
11.2 Tabla de contingencia para tipos de deporte y tipo de lesión 191
11.3 Tabla de contingencia 2×2. 197

11.4 Frecuencias esperadas para el tipo de lesión en función del deporte 200
11.5 Número de jugadas y puntos ganados en un juego por ocho personas 203
11.6 Puntuaciones diferenciales, su producto para el número de intentos y los puntos conseguidos en un juego por ocho personas 209
11.7 Variables ordenadas en en función de su *ranking* para calcular r_s 218
11.8 Salud cardiovascular y consumo de manzanilla 223
11.9 Tabla modelo para el cálculo de r_ϕ 223
11.10Matriz de correlaciones 226
11.11Matriz de correlaciones triangular 227
11.12Matriz de varianzas-covarianzas 231
11.13Relación entre el consumo de café y el cáncer de piel . . . 233
11.14Relación entre el consumo de café, el cáncer de piel y trabajar dentro de edificios 233
11.15Datos brutos para estudiar la relación que se establece entre consumo de café, cáncer de piel y exposición al sol . 234

Prólogo

Ya adelanto lo obvio, que lo importante de este libro está en el logos, así que, caro lector, puedes saltarte este prólogo sin ningún atisbo de culpa, nada importante te pierdes. En la ciencia, como en la vida, igual que le ocurría al arpa becqueriana, los datos yacen despreocupados, silenciosos y cubiertos de polvo, esperando la mano del metodólogo que sepa analizarlos y darles sentido. A veces somos nosotros mismos quienes generamos los datos exprofeso, mediante sofisticados experimentos, pergeñando osadas hipótesis en busca de confirmación o refutación. En ambos casos, y otros muchos intermedios, se necesita un cuidadoso análisis que permita extraer la información que los datos guardan celosamente, pues no hablan por sí mismos, tienden a la introversión, ni saben de dónde vienen, como bien nos enseñó Lord. Así como las manzanas contienen la sidra, lo datos albergan la información que necesitamos, pero no la ofrecen de manera obvia y directa. En el caso de las manzanas el lagar es un buen método de extracción, para los datos necesitamos un adecuado análisis, que variará en función de lo que persigamos, de la hipótesis que nos guíe. No hay análisis universales, válidos para cualquier situación, lo serán en función de lo que busquemos, unos mismos datos permiten análisis muy diferentes en función de nuestros intereses científicos y profesionales. Si los datos de partida son pobres, no hay metodología que los salve, los huevos de paloma no generan águilas, por muy buena que sea la incubadora, no conviene olvidarlo. Garantizar la calidad de los datos es previo a su análisis. Las revistas científicas están plagadas de artículos muy sofisticados metodológicamente, pero que

abordan problemas sustantivos anodinos, encontrar un equilibrio entre la relevancia sustantiva del problema estudiado y los tratamientos metodológicos es el reto fundamental de cualquier científico. Se trata de encontrar un equilibrio entre el contexto del descubrimiento y el de la justificación, por usar la terminología clásica de Reinchenbach. Es muy fácil desviarse hacia cualquiera de los dos polos, y tratar problemas relevantes con metodologías inadecuadas, o desarrollar un gran aparataje metodológico para indagar problemas irrelevantes.

Pues bien, este libro trata de todo ello, de cómo llevar a cabo los análisis de los datos de forma adecuada en cada situación. Los autores conocen bien los desarrollos metodológicos recientes y los presentan de forma clara y estructurada, ilustrándolos con numerosos ejemplos prácticos, gráficos y tablas, que permiten al lector una comprensión cabal de los distintos métodos y técnicas. El libro consta de doce capítulos, dedicado cada uno de ellos a un tema de interés, y seis apéndices en los que se tratan aspectos complementarios más especializados. En el primer capítulo se introducen de forma clara algunos de los conceptos fundamentales para entender bien el resto del libro, se repasan las taxonomías, y se subraya la importancia del muestreo para un adecuado análisis de los datos. El segundo capítulo introduce al lector en las ideas y conceptos fundamentales de la probabilidad, mientras que en el tercero se centra en la combinatoria. El cuarto es un capítulo de especial relevancia, que con frecuencia no se incluye en los manuales de análisis de datos, a pesar de ser clave para una comprensión cabal de la materia; trata de las relaciones entre el método científico y la estadística, de algunos aspectos relevantes de la actividad científica actual, y aborda con enjundia y buen conocimiento la problemática implicada en la publicación de manuscritos. Este tipo de información incluida en un libro introductorio de análisis de datos puede tener un importante valor motivador para los lectores en general y para los estudiantes en particular. El quinto capítulo expone diferentes aspectos de las distribuciones de frecuencia, muy ilustrado con excelentes gráficos que ayudan a la comprensión, amén de hacerlo más ameno. El sexto se centra en los estadísticos de posición, el séptimo en los de tendencia central, el octavo en los de dispersión, el noveno en aquellos dirigidos a evaluar la forma de las distribuciones, el décimo muestra las transformaciones de datos más frecuentes y su interpretación, el undécimo trata el concepto de correlación y sus relaciones con la causalidad y la asociación entre variables. Finalmente, en el capítulo doce los autores llevan a cabo un tratamiento muy pertinente de la clasificación diagnós-

tica, muy poco habitual en este tipo de libros, haciendo especial hincapié en la matriz de confusión diagnóstica y los estadísticos derivados de ella, y las curvas ROC.

Se trata, en suma, de un libro que introduce al lector en los arcanos del análisis de los datos. Lo hace con rigor, está muy bien escrito y editado, con gráficos, tablas y figuras que hacen su lectura más amigable y favorece la comprensión de los distintos conceptos y técnicas expuestas. Sigue la tradición de los excelentes libros sobre el tema publicados por autores españoles, que constituyen una herramienta fundamental en la iniciación de los estudiantes y profesionales al campo del análisis de los datos en las ciencias sociales y de la salud. Seguramente ello tiene mucho que ver con la pujanza en España de la metodología de las ciencias del comportamiento, y es que todo empieza por una buena base, por un adecuado *imprinting*. Hay que felicitar sinceramente a los autores por su trabajo, y felicitarnos a todos nosotros por disponer de un nuevo libro actualizado sobre análisis de datos que incluye los últimos avances y discusiones sobre el tema. Este primer paso seguro que abrirá camino a textos complementarios de problemas y ejercicios que ayuden a los principiantes a ir fijando y matizando los conceptos, a la vez que son capaces de tratar los análisis descritos de manera informatizada. Todo se andará, lo primero es lo primero, y el primer paso ya está dado, y de forma brillante.

Solo queda felicitar sinceramente a los autores por su concienzudo trabajo, y animar a los lectores a sumergirse en la lectura de un libro que les dará las claves para descifrar los secretos que guardan celosamente los datos.

José Muñiz
Catedrático de Psicometría
Universidad Nebrija

Prefacio

Este libro es fruto de nuestra experiencia investigadora y docente de los últimos 20 años en el ámbito de la metodología de las ciencias del comportamiento. Hemos tratado de proporcionar al lector mucho de lo que hemos aprendido en estas dos décadas con el ánimo de que pueda serle de utilidad en su quehacer académico y científico.

Inicialmente, este manuscrito fue concebido para ser usado como libro de texto en la asignatura de Análisis de Datos inserta en el Grado de Psicología. Por tanto, podría ser utilizado fácilmente como herramienta de trabajo para este fin. No obstante, pese a que trata aspectos básicos sobre estadística descriptiva, la lectura de esta obra también podría estar indicada para estudiantes de postgrado que se inician en la investigación científica en las áreas de psicología y educación. Queremos destacar, en este punto, que este libro es simplemente otro más que viene a complementar el magníficos elenco de manuales que ya están disponibles en nuestra lengua sobre este tópico. Por ejemplo, el libro de Amón (2000a) aborda la mayor parte de los tópicos aquí tratados y, además, introduce los modelos de regresión así como los modelos de relación no lineal entre variables. Otro manual que aborda muchos de los tópicos aquí tratados es el de (Solanas, Salafranca, Fauquet, y Núñez, 2005). Nos hubiese gustado haber incluido una introducción sobre estadística inferencial en este volumen, pero si lo hubiésemos hecho nos habríamos extendido demasiado. También existen magníficos volúmenes sobre esa temática en nuestro idioma (véase, por ejemplo, Amón, 2000b; Pardo y San Martín, 2008), aunque no es habitual que los libros introductorios a la estadística para

la psicología o para educación contengan notas sobre el cálculo de probabilidades. En este trabajo hemos introducido, precisamente, algunos elementos del cálculo de probabilidades que son esenciales para que la persona que se aproxima al análisis de datos asimile eficientemente tópicos más avanzados relacionados con la inferencia estadística. Por ejemplo, no es común que los libros que actualmente hay en el mercado incluyan secciones sobre combinatoria, algo que sí está contemplado aquí. Habíamos pensado incluir alusión a los Factores de Bayes como hemos hecho en obras previas (véase, por ejemplo, Ruiz-Ruano y Puga, 2018b) pero hemos acabado dándonos cuenta de que hubiese sido demasiado. En cualquiera de los casos, hemos intentado sentar las bases para que el lector pueda afrontar con éxito el estudio de la estadística bayesiana teniendo en cuenta la relevancia que está cobrando en la actualidad (O'Hagan y Luce, 2003; Puga, Krzywinski, y Altman, 2015a). Además, siguiendo con este perfil orientado al análisis probabilístico, hemos incluido un capítulo sobre estadísticos de clasificación diagnóstica. Tampoco es frecuente encontrar estos contenidos en los libros introductorios sobre análisis de datos. Por último, tampoco es común que aparezcan capítulos sobre el proceso de comunicación científica y nosotros hemos incluido algunas páginas para tratar este asunto.

Hemos tratado de utilizar un estilo ameno y cercano en las páginas que siguen para que cualquier persona que se aproxime a su lectura pueda seguir los argumentos sin mayores problemas. Como se puede apreciar a simple vista, abundan las ecuaciones matemáticas, pero hemos intentado que ello no comprometa la inteligibilidad del texto. Es más, en realidad, muchas secciones del texto están destinadas, única y exclusivamente, a dar explicaciones detalladas de cómo y por qué las ecuaciones son como son. Creemos que el lenguaje matemático es útil para entender nuestra realidad y hemos tratado de dar sentido verbal a esas formulaciones con el afán de que sean comprendidas lo más eficientemente posible. De este modo creemos estar contribuyendo a desarrollar un verdadero pensamiento estadístico en vez de favorecer «rituales» improductivos (Gigerenzer, 1998, 2004, 2018).

Este manuscrito lleva asociado un proyecto[1] en la *Open Science Framework* (OSF). El objetivo es triple. Por un lado, en ese proyecto hemos depositado algunos bloques de código para `R` que pueden utilizarse para replicar ciertos análisis o gráficos que hemos incluido en el libro. Algunos

[1] Proyecto del libro: `https://osf.io/pr4bv/`

datos que han sido utilizados a lo largo de los ejemplos del libro también pueden encontrarse en este repositorio. En segundo lugar, el libro será publicado en escala de grises en formato papel, pero no queríamos sacrificar el valor añadido que aportan los colores. Dado que en muchas figuras del manuscrito aparecen ilustraciones a color, hemos creado un proyecto específico, dentro del señalado anteriormente, en el que se encuentran todas las imágenes (a máxima resolución) que no pueden prescindir del color[2]. De este modo, contribuimos a la sostenibilidad del sistema y confiamos en que la persona que consiga un ejemplar en papel de este trabajo pueda disfrutar de los colores que hemos utilizado para crear los gráficos e ilustraciones del libro. Es por ello que, bajo muchas de las figuras del libro aparece la etiqueta «`VCI:`» (referido a «versión a color de la imagen:») seguida de una dirección URL. Esta dirección es, precisamente, el lugar en el que se aloja la versión en PDF de la figura correspondiente a todo color. Para terminar, en este repositorio público también puede encontrarse un archivo que contiene la «Fe de erratas» del libro[3]. Así, a medida que sean detectadas erratas en el texto publicado, podremos ir actualizando las correcciones en ese documento.

Esperamos que nuestro trabajo pueda contribuir humildemente a favorecer un pensamiento estadístico crítico, reflexivo y constructivo para la humanidad.

Ana María Ruiz-Ruano García
Departamento de Psicología Evolutiva y del Desarrollo
Universidad de Granada

Jorge López Puga
Departamento de Personalidad, Evaluación y Tratamiento Psicológico
Universidad de Granada

Compilado con LaTeX el 25 de mayo de 2022.

[2]Versión a color de las imágenes: `https://osf.io/8nvp7/`
[3]Fe de erratas: `https://osf.io/ph27j/`

Agradecimientos

TODO lo que digamos en esta sección sobre el profundo agradecimiento que sentimos hacia José Muñiz es poco. Por expresarlo con una analogía, podríamos decir que la gratitud que sentimos hacia José Muñiz es algo que no podemos expresar con palabras del mismo modo que hay cosas que no pueden ser pagadas con dinero. El escritor irlandés Oscar Wilde, supuestamente, lo expresó del siguiente modo: «hoy en día la gente conoce el precio de todo y el valor de nada»; y nos gustaría apropiarnos de esa oración para apelar a esa misma idea. Lo que queremos decir es que, para nosotros, su contribución a esta obra es de VALOR inestimable (y, por encima de todo, incalculable). Es más, esta obra ha sido llevada a término, en cierto modo, para poder contar con su contribución en ella ya que nuestras orientaciones docentes se han visto tangencialmente desviadas del análisis estadístico de datos (puro). Como decimos, no sabemos cómo podremos compensar a José Muñiz por prologar este libro. Está claro que no con dinero, pero trataremos de hacer todo lo posible por trascender su legado académico a lo largo de lo que nos queda de vida profesional.

Por otro lado, queremos dar las gracias a los editores y a la Editorial Universitaria Aula Magna - Mc Graw Hill. Su apoyo ha sido crucial para que hayamos terminado de desarrollar esta obra. Somos conscientes de la relevancia internacional de este sello editorial y confiamos que nuestro trabajo sirva para proyectar, aún más, el prestigio de la marca. Nosotros hemos intentado transmitir todo nuestro buen saber-hacer a las páginas que aparecen a continuación con el afán de rayar en la excelencia. En este

tipo de trabajos, como en otros de índole artística o manual, siempre se detectan errores u omisiones a *posteriori*, pero confiamos en que sean mínimos y siempre estaremos dispuestos a enmendarlos.

Como es habitual en nuestras obras de esta calaña, queremos agradecer a toda la comunidad de programadores del mundo su esfuerzo y dedicación para poner piezas de código a libre disposición de la humanidad. Este trabajo ha sido desarrollado casi íntegramente con *software* libre. Entre otros, hemos usado TEX, LATEX, **MiKTEX**, `R`, RStudio e Inkscape. Nuestra humilde contribución, como hemos dicho anteriormente, a esta causa, va a permitir que se pueda acceder libremente a todas las imágenes vectorizadas y a color incluidas en este libro. Es más, si alguien desea acceder a nuestro código fuente o a las imágenes en cualquier otro formato, estamos a su disposición.

Agradecemos, como no podría ser de otro modo, a la Universidad de Granada (UGR) la ayuda que nos ha proporcionado para sacar a la luz este trabajo. A modo de homenaje, debemos decir que los colores de la identidad visual corporativa de la UGR plagan este manuscrito. Los colores de los números de cada capítulo, de los títulos de cada capítulo y de los encabezados de página son, respectivamente, el verde, el azul y el rojo de la UGR[4]. Los hiperenlaces a las versiones a color de cada ilustración están también en el azul de la identidad corporativa de la UGR. También agradecemos a nuestros respectivos departamentos (Psicología Evolutiva y de la Educación y Personalidad, Evaluación y Tratamiento Psicológico) la confianza que depositan en nosotros día tras día. Vaya por delante, por tanto, un agradecimiento a todo el equipo humano con el que tenemos la fortuna de compartir, de manera más cercana, nuestras experiencias en la UGR.

Para terminar, debemos dar las gracias a nuestros hijos Julia, Álvaro (él ha diseñado el *hueso nazarí* de la portada utilizando el programa LibreCAD e inspirado por la Alhambra de Granada), Cristina y Sandra. Sin la energía que nos transmiten nunca habríamos tenido el valor y la determinación para construir este libro. Tenemos que agradecerles, del mismo modo, su comprensión y paciencia cuando, aunque haya sido en contadas ocasiones, no les hemos dedicado el tiempo con la calidad merecida. Gracias a ellos mejoramos como personas y como docentes día a día. Sin su apoyo emocional y logístico estaríamos perdidos. Por consiguiente, además de darles las gracias queremos dedicarles este trabajo.

[4] `https://secretariageneral.ugr.es/pages/ivc`

Capítulo 1
Introducción al análisis de datos

OBJETIVOS

- Presentar conceptos esenciales del análisis de datos.
- Presentar distintas taxonomías para las variables estadísticas.
- Mostrar la relevancia del muestreo para el análisis de datos.

1.1. Terminología esencial

El análisis estadístico de datos es considerado hoy día como una herramienta clave para resolver problemas de diferente índole. Tanto a nivel de políticas públicas como a nivel de gestión empresarial, el análisis estadístico de datos se presenta como una herramienta esencial para mejorar y corregir procesos de producción o servicios. Y cómo no, el análisis estadístico cobra un papel fundamental en el ámbito de la investigación científica. Gracias a la estadística los equipos de investigación en diferentes áreas de conocimiento son capaces de identificar si un fármaco es útil o no para tratar tal o cual afección o identificar qué tipo de terapia ocupacional reduce los síntomas que se asocian con ciertos trastornos psicológicos.

Para retornar a los orígenes de la estadística, tal y como la conocemos hoy en día, habría que retrotraerse al surgimiento del Estado Moderno (Fernández-Abellán y Gómez, 2002). En esta época fue cuando el filósofo, historiador, economista y jurista alemán Gottfried Achenwall, introdujo el término «estadística» (*statistik*) en un sentido parecido al que se le da hoy en día. Sin embargo, como señalan Ruiz-Ruano y Puga (2018b), los orígenes de la estadística podrían remontarse mucho más atrás. Por ejemplo, no sería descabellado pensar que los egipcios hace más de 5500 años tuviesen algún tipo de análisis publiométrico-estadístico que permitiese predecir las crecidas o sequías del Nilo. De hecho, Monteagudo, Paz, y Cámara (1997) indican que la aplicación de la estadística se remonta a unos 3050 años a. C. con motivo del inicio de la construcción de las pirámides. O, incluso, tampoco sería una barbaridad pensar que las comunidades humanas *pre*-históricas, inicialmente nómadas, se dedicasen a registrar excedentes o deficiencias de alimentos cuando dominaron la agricultura y la ganadería con el fin de prevenir y/o evitar hambruna. De hecho, hallazgos arqueológicos han concluido que el Hueso de Ishango (considerado como el vestigio más antiguo de la contabilidad humana) tiene una antigüedad de unos 20000 años (Mateos, 2019).

Hoy en día la situación es considerablemente más compleja, al menos en lo que se refiere al volumen de datos que se manejan. En nuestros días ha cobrado gran fama lo que se conoce como *Big Data*, una especie de híbrido entre las ciencias de la computación y la estadística tradicional. Esta disciplina trabaja con grandes volúmenes de datos procedentes de bases de datos descomunales que se actualizan masivamente en tiempo real. Por ejemplo, imagina la cantidad de información que genera toda la comunidad de internautas a lo largo de todo el mundo. Piensa en el texto que producen en las redes sociales, los clics que hacen sobre otros usuarios u opciones que les ofrecen las interfaces gráficas, las páginas que visitan, el tiempo que invierten en cada una de ellas, etc. Todos estos datos (y muchos, muchísimos más) son recogidos, ordenados, analizados e interpretados con diferentes propósitos. Uno de los propósitos más directo y claro es el interés económico. Es decir, la red se ha convertido en un espacio de comercio y, por tanto, analizar el comportamiento de los usuarios puede reportar ventaja competitiva a ciertos negocios. Como consecuencia de la relevancia que tiene la estadística en nuestros días, algunos investigadores consideran que sería conveniente que se produjese una cierta *alfabetización estadística* (p. e., Arias de Blois, 1993; Bond, 2009; Seldmeier y Gigerenzer, 2001) que redundase en la calidad de vida

de las personas, dada la complejidad actual del mundo.

Si tuviésemos que describir con pocas palabras cuál es el propósito del análisis estadístico de datos, podríamos decir que lo que pretende es «organizar la información para hacerla inteligible frente a la toma de decisiones». Como señala Arias de Blois (1993), el propósito de la estadística es recoger, procesar, representar e interpretar información con fines prácticos. Por ejemplo, supongamos que un ayuntamiento estuviese interesado en mejorar el servicio de gestión de residuos orgánicos con el fin de que causen los menores problemas (por ejemplo, de olores) a los ciudadanos del municipio. Para poder implementar medidas eficientes, una de las primeras cosas que se podrían hacer es elaborar un informe estadístico de las cantidades de basura que se acumulan en los contenedores en función del día de la semana. Así, si, por ejemplo, las basuras se acumulan en mayor cantidad en los días del fin de semana; entonces podría ser sensato reforzar la recogida en esos días.

Imaginemos ahora la preocupación que podría tener un empresario del gremio del transporte con autobuses. Supongamos que tuviese una plantilla de 40 chóferes y que recientemente está observando que la tasa de absentismo laboral está aumentando en los últimos años. Esto, además de ser una preocupación por la organización de los servicios, estaría haciendo perder dinero a la empresa. ¿Qué podría hacer para identificar los elementos que están propiciando la alta tasa de absentismo entre sus trabajadores? Pues quizá podría estudiar algunas variables laborales de sus chóferes. Por ejemplo, podría ver cuántos fines de semana libran, en promedio al mes, sus trabajadores. También podría ver el número de horas, en promedio, que sus trabajadores están fuera de sus hogares. O, podrían analizarse el promedio de horas de descanso que transcurre entre un servicio y otro. Y todo ello podría ser segmentado entre trabajadores con alto y bajo absentismo. Si se observase que los trabajadores con bajo absentismo son aquellos que más fines de semana libran al mes, quizá habría que considerar ese dato para reorganizar los servicios con la intención de reducir el absentismo.

Cuando realizamos estadísticas tenemos que tener claro la diferencia que existe entre una población y una muestra. Una *población* es un conjunto de elementos (personas, animales, objetos, etc.) que comparten algunas características. Por su parte, una *muestra* es un subconjunto de una población. La idea de población estadística no se limita necesariamente al concepto territorial. Es decir, una población podría estar referida al conjunto de personas que viven en una determinada ciudad, pero

también podría referirse al total de personas que sufren cierta dolencia médica (por ejemplo, aquellas personas que tienen un colesterol alto). O, incluso, una población podría ser el total de tornillos producidos por una fábrica en un mes. Es decir, una población es, sencillamente, un conjunto de elementos que son susceptibles de ser estudiados estadísticamente.

De manera general, el número de elementos que componen una población suele representarse con la letra ene mayúscula (N). Por su parte, el tamaño muestral suele vincularse a la letra ene minúscula (n). Esto no es siempre así en todos los contextos, pero en algunas ocasiones se utilizan indistintamente la ene mayúscula y minúscula para referirse al tamaño muestral. Sin embargo, en este manual siempre consideraremos que la ene mayúscula está referida al tamaño poblacional mientras que la ene en minúscula se referirá al tamaño muestral. Al procedimiento por medio del cual se obtiene una muestra a partir de una población se denomina «muestreo».

Definición 1.1: Población y muestra

La **población** está referida a un conjunto de elementos (personas, animales u objetos) que comparten ciertas características y sobre la cual cabe hacerse alguna pregunta estadística. Por su parte, la **muestra** es un subconjunto de la población que es recogida y analizada con el fin de sacar conclusiones sobre alguna o algunas características de la población.

La estadística se ha dividido en dos grandes bloques tradicionalmente. Por un lado, tendríamos lo que se conoce como «estadística descriptiva» y, por otro, lo que se denomina como «estadística inferencial» (Krzywinski y Altman, 2013a). La *estadística descriptiva* está interesada en analizar los datos muestrales sin pretender gereralizarlos a la población. Por su parte, la *estadística inferencial* pretende ir más alla de los datos muestrales observados para tratar de sacar conclusiones que puedan ser generalizadas a la población. O, como señala Jeffreys (1931), la inferencia estadística consistiría en «derivar información sobre sensaciones que aún no han sido experimentadas» (p. 1).

Por ejemplo, supongamos que un oficial de policía estuviese interesado en conocer el número de accidentes de tráfico mortales que se producen en una carretera concreta de su municipio. Si tomase una muestra aleatoria de 50 expedientes de accidentes acaecidos en los últimos seis meses en esa carretera y calculase la proporción de accidentes por semana, el agente estaría realizando un análisis estadístico descriptivo. Lo que esta-

ría obteniendo sería un resumen de lo que ha sucedido en esa carretera en los últimos seis meses. Sin embargo, si el agente estuviese interesado en conocer el intervalo de accidentes que se producen con cierta probabilidad (por ejemplo, con el 95 % de confianza) en esa carretera cada semana, lo que estaría haciendo es un análisis estadístico inferencial. En este segundo caso, el agente de policía estaría tratando de ir más allá de los datos observados y pretendería generar un modelo que permitiese encuadrar la proporción de accidentes en un rango que tuviese sentido probabilístico. Es decir, estaría tratando de identificar el número de accidentes que cabría esperar en una semana concreta con cierto grado de confianza. Dicho de otro modo, cuando se realiza estadística inferencial se está tratando de obtener información que nos permita reducir la incertidumbre natural que está asociada a los datos muestrales que recogemos con el objetivo de generalizar esa información a la población.

Definición 1.2: Estadística descriptiva e inferencial

La **estadística descriptiva** se limita a analizar los datos sin pretensión de generalizar los datos a la población de la que se extrajo la muestra, mientras que la **estadística inferencial** pretende reducir la incertidumbre natural contenida en los datos muestrales con el objetivo de hacer generalizaciones sobre la población.

Cuando se estudia una propiedad de la realidad (por ejemplo, el número promedio de hijos que tienen las personas) se hace una diferenciación en función de si lo estamos estudiando a nivel poblacional o a nivel muestral. Cuando calculamos el promedio de hijos sobre el total de una población lo que se obtiene es lo que se denomina *parámetro poblacional*. Los parámetros poblacionales suelen referirse con letras griegas (por ejemplo, μ, β, o π entre otras). En el caso del promedio, se utiliza la letra griega mu (μ). Por su parte, cuando calculamos el promedio de hijos en una muestra aleatoria de ciudadanos, lo que obtenemos se denomina «estadístico muestral». En este caso, el estadístico muestral denominado como «promedio aritmético» se representa con la expresión $\bar{x}$.

Definición 1.3: Muestreo

El **muestreo** es el procedimiento por medio del cual se obtiene una muestra a partir de una población.

1.2. Tipos de variables

En la sección anterior hemos indicado que el propósito de la estadística es recoger información, organizarla, interpretarla y tomar (o sugerir) decisiones al respecto. Pues bien, en estadística, cuando nos referimos a información, esta se presenta en forma de variables. Una *variable* es sencillamente una característica de la población que varía, es decir, que no es constante. Dicho de otro modo, una variable es una característica que asume, como mínimo dos posibles valores. Por ejemplo, el que guste o no la música clásica, podría considerarse como una variable. Es decir, si preguntamos a muchas personas si les gusta la música clásica veríamos que algunas personas responderían que sí les gusta mientras que otras responderían que no les gusta. Esto es, existiría variabilidad en las respuestas. Por ello, la respuesta a esa pregunta es una variable y no una constante (imagina que todo el mundo dijese que no le gusta la música clásica, eso no sería una variable sino una constante).

Definición 1.4: Variable

Una **variable** es una característica de una población que puede asumir diferentes posibles valores.

Desde el punto de vista tradicional, las variables se han clasificado en tres grandes grupos. Por un lado, existen las *variables cuantitativas*, que son aquellas cuyos posibles valores son el resultado de contar o cuantificar algún fenómeno. Dentro de las variables cuantitativas podemos distinguir, a su vez, otros dos tipos de variables: las variables cuantitativas discretas y las variables cuantitativas continuas. Las *variables cuantitativas discretas* son aquellas que no asumen valores intermedios entre dos valores dados. O, dicho de otra forma, son aquellas que solo asumen números naturales ($\mathbb{N}$, o enteros, $\mathbb{Z}$) como valores admisibles. Un ejemplo de variable continua discreta es el número de hijos. El número de hijos solo puede asumir valores del conjunto de los números naturales[1] ($\mathbb{N} = 0, 1, 2, 3, \cdots$) o el valor cero, y no es posible que el número de hijos esté comprendido entre dos números naturales. Es decir, no es

[1] Véase que aquí utilizamos una de las formas posibles de considerar a los números naturales incluyendo al cero dentro de ese conjunto de numérico. En realidad, no existe consenso entre los expertos en matemáticas al considerar si el número cero forma parte de los números naturales o no (Maestre, 2019).