Universidad Autónoma de Ciudad Juárez

ESTADÍSTICAS DESCRIPTIVAS

con uso de Excel para la Investigación Educativa

HÉCTOR FRANCISCO PONCE RENOVA

UNIVERSIDAD AUTÓNOMA DE CIUDAD JUÁREZ

Juan Ignacio Camargo Nassar

Rector

Daniel Constandse Cortez

Secretario General

Alonso Morales Muñoz

Director del Instituto de Ciencias Sociales y Administración

Jesús Meza Vega

Director General de Comunicación Universitaria

Universidad Autónoma de Ciudad Juárez

ESTADÍSTICAS DESCRIPTIVAS

con uso de Excel para la Investigación Educativa

HÉCTOR FRANCISCO PONCE RENOVA

Avenida Plutarco Elías Calles 1210
Fovissste Chamizal, C. P. 32310
Ciudad Juárez, Chihuahua, México
Tels. +52 (656) 688 2100 al 09

Ponce Renova, Héctor Francisco
Estadísticas descriptivas: con uso de Excel para la Investigación Educativa / Héctor Francisco Ponce Renova.– Primera edición -- Ciudad Juárez, Chihuahua, México: Universidad Autónoma de Ciudad Juárez, 2023. 201 páginas; 22 centímetros.

ISBN: 978-607-520-458-1

Contenido: Presentación personal.— Agradecimientos y reconocimientos.— Introducción.— Resumen.—Objetivos de aprendizaje.— Términos clave.— Capítulo 1 Uso de las estadísticas en la Investigación Educativa.— Capítulo 2 Variables y sus distribuciones.— Capítulo 3 Organización y representación de los datos.— Capítulo 4 Medidas de tendencia central.— Capítulo 5 Medidas de dispersión.— Capítulo 6 Percentiles, rango de percentil, rangos, cuartiles, deciles, diagramas de caja (*box plots*) y valores estándar.— Apéndices.— Índice de tablas y figuras.

1. Estadísticas descriptivas — Investigación Educativa
2. Métodos estadísticos — Procesamiento de datos
3. Excel (Hoja de cálculo) — Estadísticas descriptivas

LC – QA276.45M53

La edición, diseño y producción editorial de este documento estuvieron a cargo de la Dirección General de Comunicación Universitaria, a través de la Subdirección de Editorial y Publicaciones

Coordinación editorial:
Mayola Renova González
Cuidado editorial:
Subdirección de Editorial y Publicaciones
Diseño de portada y diagramación:
Karla María Rascón

Primera edición, 2023
Disponible en: https://elibros.uacj.mx

Índice

Capítulo 2

Capítulo 3

Capítulo 4

Capítulo 5

Capítulo 6

Apéndices

Índice de tablas y figuras

Capítulo 1

Capítulo 2

Capítulo 3

Capítulo 4

Capítulo 5

Capítulo 6

Apéndices

Presentación personal

Héctor F. Ponce Renova, Ph. D., ha sido durante una década profesor e investigador de tiempo completo adscrito al Departamento de Humanidades de la Universidad Autónoma de Ciudad Juárez en el Instituto de Ciencias Sociales y Administración. Piensa que podemos obtener conocimiento de la psicometría y la estadística. Esta idea ha guiado su trayectoria de impartición de cursos de las ciencias antes mencionadas en los programas de Educación y Economía. Su educación incluye una Licenciatura en Artes con Especialización Dual en Economía y Periodismo por la Universidad de Texas en El Paso (UTEP); Maestría en Administración de Empresas por UTEP; Maestría en Ciencias en Recursos Humanos por la Universidad de Texas en Arlington; y Doctorado en Investigación en Educación con Especialización en Psicometría y Estadística por la Universidad del Norte de Texas. Su investigación tiene dos temas principales: las propiedades psicométricas de las pruebas y las encuestas, así como los enfoques metodológicos de la estadística. Uno de los objetivos de su investigación ha sido enseñar cómo utilizar las mejores prácticas posibles en la investigación cuantitativa en las Ciencias Sociales.

Ha publicado cuatro libros de estadísticas arbitrados por pares; ha editado un libro sobre investigación y la educación especial; ha publicado en español e inglés un total de 13 artículos sobre estadísti-

ca y psicometría en revistas arbitradas; ha ido a 7 congresos, donde también ha publicado 7 artículos arbitrados junto con sus estudiantes; tiene 7 capítulos de libro publicados; y ha dirigido exitosamente 10 tesis de maestría, así como 52 tesis de licenciatura en el Programa de Educación. Actualmente, se encuentra dirigiendo 2 tesis de maestría y 14 de licenciatura. Cree que la enseñanza y las publicaciones van de la mano en la vida académica y que se nutren mutuamente.

Ha participado en la actualización y rediseño de los programas de Maestría en Investigación Educativa, Maestría en Educación Especial y Licenciatura en Educación. En estos programas ha tenido la oportunidad de ajustar el currículo de las clases de Metodología de la Investigación, Estadística y Psicometría, para que cumplan con nuevos conocimientos y uso de *software*. Asimismo, ha creado una clase de Estadística Descriptiva para la Licenciatura en Educación.

Agradecimientos y reconocimientos

Le doy gracias a mi familia nuclear: Laura, Isabella Concepción alias *La Peque*, Prathiba y Simón, así como a la memoria de *Booguie*. Compartir con ustedes los días de altas y bajas de la vida, le ha dado un propósito a lo que escribo y a mi existencia. Igualmente, le doy las gracias al resto de mis queridos familiares y amigos que me han ayudado a formarme, tanto académica como personalmente, a través de décadas. Una persona no se forma en un vacío.

Reconozco la gran aportación de enseñar en una universidad para la creación de este libro de estadística. Por lo tanto, agradezco a las y el colega a cargo de los programas que me han dado la oportunidad de trabajar con ellos: Licenciatura en Educación, a cargo de Yeshica Márquez; Maestría en Economía, a cargo de Raúl Ponce; y Maestría en Investigación Educativa, a cargo de Beatriz Anguiano. Otra persona fundamental para que este libro, como los anteriores, haya visto la luz ha sido Cely Ronquillo, así que estoy muy agradecido con su gran trabajo.

Además, cada alumna y alumno con sus preguntas y comentarios me han ayudado a desarrollar, en primera instancia, materiales para las clases que luego se han vuelto libros como el presente. A veces, ni siquiera hacía falta que dijeran algo. Bastaba con la mirada de duda para que me esforzara más por explicar mejor el material de las clases.

No cabrían los nombres de todas y todos aquí, pero les dedico este trabajo que también es su legado para las próximas generaciones.

Por último, pero no menos importante, mi agradecimiento y reconocimiento a las autoridades administrativas de la UACJ encabezadas por el maestro Juan Ignacio Camargo Nassar por apoyar nuestra sencilla obra, pero de trabajo arduo. Ellas y ellos nos han provisto de los recursos necesarios, para que obras, como la mía, o más bien nuestra, se puedan publicar y difundir.

En memoria de ***Booguie***

(6 de junio de 2014 - 3 de junio de 2021)

Introducción

Resumen

La obra contiene seis capítulos de estadística descriptiva, en los cuales se usa Excel 2016:

1. Uso de las estadísticas en la Investigación Educativa;
2. Variables y sus distribuciones;
3. Organización y representación de los datos;
4. Medidas de tendencia central;
5. Medidas de dispersión; y
6. Percentiles, rango de percentil, rangos, cuartiles, deciles, diagramas de caja (*box plots*) y valores estándar.

Asimismo, cada capítulo contiene las siguientes secciones para profundizar en los temas:

» **Preguntas para resolver del capítulo.** Se muestra una serie de interrogantes para resolver o inferir las respuestas del contenido, con el propósito de reforzar los conocimientos y habilidades.

- **Problemas para resolver.** Están basados en el contenido del capítulo y se resuelven con Excel, y, en algunos casos, manualmente. Las soluciones *clave* a estos problemas están incluidas en el Apéndice G.
- **Preguntas para reflexionar.** Tratan de ir más allá del contenido propio de este libro, para que las y los lectores encuentren otros recursos para dar respuesta a estas interrogantes.
- **Opinión del Autor.** Esta trata de explicar y asociar el contenido de cada capítulo a la experiencia y perspectiva de este investigador/profesor.

Además de estas secciones, en este libro se muestra cómo resolver los ejemplos de los capítulos con Excel 2016 y manualmente, explicando cada paso y ecuación detalladamente. Esta obra contiene siete apéndices, donde se explican detalles y fórmulas para ampliar el contenido de los capítulos y resolver algunos de los problemas.

Descripción de la obra y su relación con la Investigación Educativa

En algunas áreas de la Investigación Educativa, se hace uso de la *medición* del aprendizaje, efectos de la pandemia por la COVID-19 en las escuelas, inteligencia, deserción, motivación, autorregulación y ausentismo, entre muchos otros *fenómenos*, para luego ser analizados por medio de estadísticas. En el caso de este manuscrito, estas estadísticas sirven para describir los datos con el uso de Excel 2016. De estos análisis descriptivos pueden surgir resultados que, a su vez, pueden ser utilizados para crear argumentos (para argumentación en estadística, véase: Hurley, & Watson, 2018) que puedan o no apoyar teorías del aprendizaje, como la *Teoría Cognoscitiva Social*, *Teoría del Constructivismo* y *Teoría del Procesamiento de la Información*, entre otras. Con esta obra, se trata de poder observar hasta dónde las mediciones del aprendizaje y otros fenómenos relacionados con la educación poseen datos con una posible distribución normal, así como sus estadísticas descriptivas, para después poder usar estadística inferencial con ayuda de otros textos. Además, algunas de las preguntas que se tratan de contestar con el apoyo de Excel 2016 son:

- ¿Qué es la estadística descriptiva y cómo se usa en la Investigación Educativa?
- ¿Cuáles son algunas de las posibles distribuciones de los datos?
- ¿Cómo se organizan y se representan los datos?
- ¿Cómo se utilizan las medidas de tendencia central y de dispersión?
- ¿Cómo se pueden organizar los datos en alguna jerarquía?

Al dar respuesta a estas preguntas, se espera que el panorama de las y los lectores se expanda para poder aplicar los conceptos y análisis mostrados en el libro, para llevar a cabo Investigación Educativa cuantitativa con estadística descriptiva y uso de Excel 2016.

Asimismo, el uso de la presente obra puede ser un inicio para investigadoras e investigadores educativos para ayudar, como primera aproximación, a entender y analizar la enorme cantidad de datos que se generan por las redes sociales, como Facebook, Instagram, YouTube y Twitter, entre otras fuentes, que se conoce como *Big Data*. Asimismo, datos recabados de la pandemia por la COVID-19 y sus efectos en la educación son otra fuente de *Big Data*, que engloba bases de datos *muy* grandes, como las que generan las ya mencionadas redes sociales para ser analizadas por medio de *softwares*. Un objetivo de los análisis es encontrar patrones, tendencias y asociaciones, especialmente relacionados con algunos comportamientos humanos e interacciones. Este concepto de *Big Data*, ligado a otro llamado Minería de datos (*Data mining*: es la práctica de examinar grandes bases de datos para generar nueva información), pueden ser de gran ayuda en la Investigación Educativa para resolver preguntas como las siguientes: ¿cómo comenzar a organizar y representar datos?; y ¿cuáles son las medidas de tendencia central y variación de los datos? Más precisamente, un ejemplo de *Big Data*, Minería de datos y la presente obra sería examinar el aprendizaje promedio y la desviación estándar de distintos grupos de estudiantes.

Dirigido a lectoras y lectores

Este libro está dirigido a las y los estudiantes, investigadoras e investigadores y profesoras y profesores que deseen entender la estadística descriptiva con el uso de Excel. La obra podría ser usada de manera independiente, porque puede ser considerada como el inicio de la estadística. Más al respecto, el presente libro forma parte de una serie de obras sobre estadística inferencial de Ponce-Renova (2019), Ponce-Renova (2020) y Ponce-Renova (2021).

Conocimientos previos

Se asume que las y los usuarios del presente libro tienen ciertos conocimientos elementales de álgebra e interés en usar Excel. La mayoría de los cálculos de la presente obra fueron ejecutados en Excel 2016 y en forma manual.

Aporte principal de la obra

La aportación principal de esta obra es tratar de sentar la base de la estadística descriptiva con el uso de Excel 2016. Otra aportación es que la mayoría de las fuentes del presente libro fueron publicadas originalmente en el idioma inglés y, con esta obra, se tradujeron para hacerlas más accesibles a aquellas personas que prefieren el idioma español.

Razón para la creación y publicación de la obra

Esta obra surgió de materiales didácticos de las clases de Métodos Cuantitativos para la Investigación Educativa y de Estadística para la Economía. Además, estos materiales didácticos fueron usados en seminarios de Tesis de Educación a nivel pre y posgrado. Durante casi una década, los materiales didácticos se han ido perfeccionando por las contribuciones de estudiantes y colegas para llegar a las versiones actuales. Ahora, esta obra se ha convertido en un libro de divulgación, ya que existen muchos programas educativos que lo podrían emplear como una referencia para sus tesis y clases.

Objetivos de aprendizaje

Al terminar el libro, la lectora o el lector será capaz de:

- Identificar las estadísticas descriptivas
- Diferenciar, a un nivel básico, las estadísticas descriptivas y las inferenciales
- Comprender cómo las estadísticas básicas sirven para organizar, resumir y presentar datos
- Calcular frecuencias, porcentajes, medidas de tendencia central y medidas de dispersión
- Desarrollar una noción básica de una variable, su escala de medición y su distribución
- Jerarquizar datos en percentiles, rangos, rangos de percentil, cuartiles, deciles, diagramas de caja y valores estándar
- Usar Excel 2016 para las estadísticas descriptivas

Términos clave

- Cuartil
- Decil
- Desviación estándar
- Diagrama de caja (*box plot*)
- Distribución
- Escala
- Escala de intervalo
- Escala de razón
- Escala nominal
- Escala ordinal
- Estadísticas descriptivas
- Estadísticas inferenciales
- Frecuencia
- Gráfica de barras
- Histograma
- Mediana
- Medidas de dispersión
- Medidas de tendencia central
- Moda
- Percentil
- Porcentaje
- Promedio
- Rango
- Rango de percentil
- Sigma
- Valor absoluto
- Valor estándar
- Variable
- Variable continua
- Variable discreta
- Variable nominal
- Varianza

CAPÍTULO 1

Uso de las estadísticas en la Investigación Educativa

"Todo lo que existe se manifiesta en alguna cantidad" es una frase de Edward Lee Thorndike (1918, p. 16), célebre teórico de psicología, creador de múltiples pruebas para medir el logro académico y miembro del Comité de Evaluación de Reclutas (*i. e.*, *Army Alpha and Beta Test*) durante la Primera Guerra Mundial. En esta guerra, se medía la inteligencia de los soldados para mejorar la selección de los mismos para su ubicación laboral dentro de las Fuerzas Armadas y su entrenamiento para ciertas tareas. Las estadísticas sirven en la Investigación Educativa cuantitativa para organizar, resumir, analizar y presentar estas cantidades que mencionó Thorndike, así como para describir y graficar datos, hacer comparaciones entre grupos (entre otras comparaciones) y relacionar variables. Otra idea que apoya la idea de utilizar la estadística descriptiva fue atribuida al Padre de la Inteligencia Artificial, John McCarthy, quien dijo: "Aquel quien se rehúsa a realizar aritmética está condenado a hablar sin sentido" (Rajaraman, 2014, p. 207). Para *cuantificar lo existente* y usar *aritmética* con datos, como propusieron los anteriores autores, se emplea Excel 2016 y cálculos manuales en el presente libro como una herramienta de la estadística descriptiva.

a. ¿Para qué sirve la estadística en la Investigación Educativa?

Una simple respuesta a esta interrogante puede ser: para conocer un *fenómeno* medible como el *aprendizaje*[1] (probablemente en la Investigación Educativa, el constructo[2] de mayor interés), entre muchos otros más. Aparte, hay algunos científicos que afirman que, sin medición,[3] no hay ciencia. En contraparte, aquí viene una advertencia atribuida a Albert Einstein, quien declaró: "No todo, lo que puede ser contado, cuenta, y no todo, lo que cuenta, puede ser contado" (Toye, 2015, p. 7). Por ello, hay que tener cuidado con lo que se mide, cómo se mide y lo que se deja de medir. En el contexto escolar del aprendizaje, este tema de la medición corresponde a la psicometría y, para ahondar, se recomienda consultar a Muñiz (2018); Rust, Kosinski, & Stillwell (2021); y Shultz, Whitney, & Zickar (2020). Está más allá de los objetivos del presente libro abordar las propiedades psicométricas de los instrumentos[4] para recolectar datos.

Por ejemplo, hay que hacer una medición del aprendizaje mediante alguna prueba o encuesta para darle un valor numérico, como una calificación. Una medición puede ser una variable *observable*, como la estatura y el peso de una/un estudiante. Por otro lado, las variables *no*-observables, como el aprendizaje y la inteligencia (variables

1 Es la adquisición de nueva información, comportamientos o habilidades después de practicar la observación u otras experiencias. Esta adquisición es evidenciada por un cambio en el comportamiento, conocimiento o funcionamiento del cerebro. El aprendizaje involucra, consciente o inconscientemente, el poner atención en los aspectos relevantes de una información entrante. Mentalmente se organiza la información en representaciones cognoscitivas coherentes con integración al conocimiento relevante y preexistente que fue activado desde la memoria a largo plazo (VandenBos, 2015, p. 594).

2 Un modelo para explicar un fenómeno basado en eventos o procesos que son medibles y empíricamente verificables —un constructo empírico— o un proceso inferido de los datos de este tipo, pero que no son observables directamente —un constructo hipotético (VandenBos, 2015, p. 239)—. En pocas palabras, es una variable que no se puede observar directamente como el aprendizaje, pero en teoría se manifiesta de alguna manera como los resultados de alguna prueba de conocimientos, entre otras.

3 Es un proceso sistemático para asignar números a ciertas características, de acuerdo con cierta regla (Hinkle, Wiersma, & Jurs, 2003, p. 8).

4 Son encuestas (*e. g.*, satisfacción con una biblioteca) y pruebas (*e. g.*, conocimientos como la Prueba PISA) para medir constructos. Para la creación y evaluación de instrumentos, se recomienda ver la publicación de la American Educational Research Association, American Psychological Association y National Council on Measurement in Education (2014).

latentes o constructos), son medibles: por lo menos lo son en teoría.[5] Después de la medición de algún aprendizaje (*i. e.*, variable conceptual[6]) mediante calificaciones (*i. e.*, definición operacional[7]) como un ejemplo, se puede pasar a la *estadística descriptiva* en primera instancia. Con la estadística descriptiva, se pueden organizar, resumir, analizar y presentar gráficamente los datos. Luego, se puede pasar a la estadística inferencial para estimar valores (*i. e.*, parámetros[8]) de las poblaciones mediante las estadísticas (*i. e.*, valores) de las muestras y poner a prueba hipótesis para posibles generalizaciones (*cf.* Ross, 1997, p. 257). Además de buscar el conocimiento *per se*, la estadística puede tener usos prácticos en la educación (véase: Ravid, 2020) para la toma de decisiones ante problemas como la reprobación, ausentismo, abandono escolar, manejo de pandemias como la COVID-19, entre muchos otros.

b. Visión general de estadísticas descriptivas e inferenciales

En el sentido más amplio del término, las estadísticas se refieren a un rango de técnicas y procedimientos para analizar, interpretar, mostrar y tomar decisiones con base en los datos (Lane *et al.*, 2014). Otro aspecto de la Investigación Educativa cuantitativa implica el uso de probabilidad y estadística en los análisis de resultados, tanto en estudios experimentales como no-experimentales. La estadística es una rama de las matemáticas que transforma los números en información útil para los hacedores de decisiones (Berenson *et al.*, 2019). Además de informar a los hacedores de decisiones, en la presente obra se consid-

5 En general, una teoría puede ser definida como: Un principio o cuerpo de principios interrelacionados que pretende explicar o predecir una serie de fenómenos interrelacionados. Una segunda definición: En la filosofía de la ciencia, es un conjunto de hipótesis explicativas lógicamente relacionadas que son consistentes con un cuerpo de hechos empíricos y que pueden sugerir relaciones más empíricas. Véase explicación científica. Una manera llana de definir una teoría es: Un mecanismo que explica la relación de causa y efecto entre variables.

6 Un constructo como el aprendizaje, se puede definir por medio de una teoría: esta sería una definición conceptual de la variable aprendizaje (Ponce-Renova, 2019, p. 77).

7 Una definición operacional es una descripción de algo en términos de operaciones (procedimientos, acciones o procesos) que se pueden observar y medir (VandenBos, 2015, p. 735).

8 Ejemplos de parámetros son el promedio, la desviación estándar y la varianza, entre muchos otros más. Más adelante, en este capítulo, se habla sobre este tema.

era que las estadísticas son útiles para la búsqueda del conocimiento en sí mismo y como herramientas para investigadoras, investigadores y otras partes interesadas. Para muchas y muchos investigadores la estadística tiene *tres funciones* en general: describir, comparar y relacionar. Algunos ejemplos de estas funciones de la estadística son:

- *Descripciones*. Son representaciones gráficas de datos, medidas de tendencia central, medidas de dispersión, ordenamiento de datos, agrupaciones de variables (*i. e.*, análisis de componentes principales[9] y análisis exploratorio de factores[10] [Tabachnick, & Fidell, 2019]) y agrupaciones de personas, observaciones u objetos (*i. e.*, análisis de *clústeres*[11] [Hair *et al.*, 2019]), entre otros.
- *Comparaciones*. Son comparaciones de promedios de grupos (Pruebas *t* de estudiante [Ponce-Renova, 2021] y pruebas de análisis de varianza [Hinkle *et al.*, 2003]); y comparaciones entre datos observados y esperados (Prueba de chi cuadrada [Hinkle *et al.*, 2003]), entre otros.
- *Relaciones*. Son relaciones entre dos variables (correlaciones de producto-momento y rho de Spearman [Ponce-Renova, 2020]); relaciones lineales múltiples (regresión lineal múltiple y correlación canónica [Tabachnick, & Fidell, 2019]); y no-lineales (regresión logística [Hair *et al.*, 2019]), entre otros.

Las estadísticas que se cubren en este libro son descriptivas (para una introducción a las estadísticas inferenciales, véase: Ponce-Renova, 2019; Ponce-Renova, 2020; Ponce-Renova, 2021). En resumen, las estadísticas descriptivas[12] cubren aspectos de gráficas y tablas, tendencia central, dispersión, ordenamiento de datos y frecuencia cuando no

9 Sirve para combinar un gran número de variables y reducir su número a uno mucho menor (llamadas componentes); es parte de análisis multivariados (*i. e.*, varias variables).
10 Tiene la función de agrupar variables observables en constructos (factores), para observar si los datos apoyan un modelo teórico: usado como parte de análisis de propiedades psicométricas. También, es parte de análisis multivariados.
11 Muestra si un número de variables puede dividir a las personas, observaciones u objetos en dos grupos o más. Es uno más de los análisis multivariados descriptivos.
12 La estadística descriptiva es una rama de la estadística que trata con manipulaciones numéricas que pueden ser utilizadas para describir y resumir sets de datos (Russo, 2021, p. 13).

se desea generalizar, entre otros. En contraparte, con las estadísticas inferenciales se trata de generalizar. Se va de las estadísticas encontradas en una muestra a los parámetros de una población.

c. Estadísticas descriptivas

Son números que son usados para resumir y describir datos de un grupo (Lane *et al*., 2014). A un grupo se le suele llamar un conjunto de datos o set de datos. En la investigación educativa, los datos pueden tener diferentes orígenes, como instituciones gubernamentales[13] o privadas, encuestas, exámenes (*i. e*., pruebas), entre otras potenciales fuentes. A través de instrumentos como pruebas o encuestas, se recolectan datos para ser resumidos a través de estadísticas descriptivas. Para VandenBos (2015), la estadística descriptiva fue definida como:

> Procedimientos para representar los aspectos principales de una muestra de datos, sin necesariamente hacer una inferencia hacia la población. Describir estadísticas usualmente incluye el promedio, la mediana y la moda para indicar las medidas de tendencia central, así como el rango y la desviación estándar que revelan qué tan esparcidos están los puntajes dentro de su muestra. Las estadísticas descriptivas podrían también incluir tablas y gráficas, tales como distribución de frecuencias o histogramas, entre otras. (p. 301)

d. Tipos de teorías de probabilidad y estadística

Abordando el tema de las variantes de probabilidad y estadística, Ponce-Renova (2021) explicó que esta se divide en tres grandes ramas:

13 Secretaría de Educación Pública (México): https://www.gob.mx/sep; Instituto Nacional de Estadística y Geografía (INEGI): https://www.inegi.org.mx/
La Universidad de California en San Diego (UC San Diego) contiene un sitio donde se pueden encontrar una serie de páginas con información acerca de la educación: https://ucsd.libguides.com/data-statistics/education

a) *Teoría Clásica de la Probabilidad* (*Classical Probability Theory*), conocida como Teoría de la *Frecuencia*, que fue diseñada para largo plazo, lo cual implica que un solo resultado estadísticamente significativo en un solo estudio no es suficiente para probar un *efecto*,[14] porque habría que ver si existe un patrón al realizar una serie de réplicas a través del tiempo (para saber más de replicaciones, se recomienda consultar a Cumming, 2013);
b) La otra rama conocida como *verosimilitud*, llamada en inglés *likelihood* (véase: Rossi, 2018: este libro es para personas con conocimientos avanzados de matemáticas), que se refiere, a grandes rasgos, a la identificación del valor más probable en un conjunto de datos; y
c) La probabilidad y estadística relacionadas con el *Teorema de Bayes* (véase: Morris, 2016: esta obra es para principiantes), en el que, *grosso modo*, se declara una distribución de datos a manera de predicción educada de un fenómeno (*prior distribution*); se recaban los datos de la distribución de dicho fenómeno; y de ambas distribuciones, se obtiene una distribución más del fenómeno llamada posterior (*posterior distribution*). Se recomienda consultar a Russo (2021), porque cubre tanto la Teoría de la Frecuencia como la Bayesiana.

Cuando se hace alguna referencia a la estadística inferencial en el presente libro, se está refiriendo a la *Teoría Clásica de la Probabilidad*: Teoría de la Frecuencia.

e. Recursos del uso de Excel

En el Apéndice A-1 se muestra una serie de recursos de Excel 2016 y versiones más recientes de este programa para las y los usuarios de esta obra. Allí se pueden encontrar los enlaces a una serie de páginas

14 En general, un efecto sucede cuando una variable causa un cambio en otra (*cf.* VandenBos, 2015, p. 352). Por ejemplo, si al aplicar una serie de tutorías a un grupo de estudiantes, estos mejoran su aprendizaje respecto a como estaban antes de las tutorías, estas últimas tendrían un efecto en el aprendizaje; otras cosas siendo iguales. En otras palabras, las tutorías podrían tener un efecto en el aprendizaje de alguna materia.

de Microsoft, donde se explican desde temas básicos (crear un libro de trabajo con páginas de hojas de cálculo) hasta coautorías, entre otros, por medio de textos y videos. Se recomienda a las y los usuarios que no hayan sido muy expuestos a Excel, dar un vistazo a estas páginas antes de comenzar a usar este libro. De este modo, las indicaciones que se dan podrían tener más sentido para las y los lectores. Más al respecto, en el Apéndice A con la Tabla A-2 se muestran *análisis inferenciales* con sus videos en YouTube como una manera de facilitar referencias para las y los lectores. Aparte, YouTube (véase después de las *Referencias* del presente libro en los *Recursos de internet*) contiene una serie de videos para enseñar aspectos que contiene esta obra, así como otros temas relacionados con Excel 2016.

Por otro lado, existen alternativas a Excel para las estadísticas:

- Lado gratuito: JASP, que es *software* de fuente abierta patrocinado por la Universidad de Ámsterdam. JASP fue hecho como una alternativa a SPSS. Asimismo, JASP es una interfaz de *R*, que es un programa y un lenguaje gratuito de fuente abierta. Ambos *softwares* crecen constantemente con las aportaciones de docenas de contribuidores;
- Lado comercial: Están SPSS, Minitab y SAS como, posiblemente, los más conocidos.

Excel, al igual que todos estos *softwares* estadísticos antes mencionados, ayuda a minimizar el tiempo y los conocimientos de algoritmos[15] para analizar datos; entonces, uno se podría enfocar en maximizar *destrezas analíticas* al evaluar los resultados (*cf.* Levine, Stephan, & Szabat, 2021). Estas destrezas analíticas podrían interpretarse como *pensamiento crítico*.[16] Hasta cierto punto, se utiliza el pensamiento

15 Es una secuencia finita de instrucciones bien definidas que se pueden implementar por computadora, generalmente para resolver una clase de problemas o para realizar un cálculo.

16 Es una forma directa, un pensamiento enfocado en el problema, en el cual el individuo pone a prueba ideas o posibles soluciones para observar los posibles errores e inconvenientes. Es esencial para estas actividades como examinar la validez de las hipótesis o en la interpretación del significado de los resultados (VandenBos, 2015, p. 267). De una manera más coloquial, el pensamiento crítico busca encontrar explicaciones a los eventos que suceden en lugar de atribuirle la relación de causa y efecto a explicaciones mágicas.

crítico en la presente obra al obtener evidencias para sacar conclusiones. Por esta razón, este libro muestra una serie de ejemplos para utilizar Excel en cierto contexto y obtener respuestas a una serie de *preguntas* acerca de datos.

Hablando del contexto, Libman (2010) explicó cómo integrar el proceso de investigación con conjuntos de datos reales a un curso de estadística descriptiva. Su propósito fue instruir métodos de enseñanza alternativos para propiciar que las y los estudiantes tomaran un rol más activo en su propio aprendizaje y participaran en el proceso de evaluación de su propio aprendizaje. En la presente obra, al final de cada capítulo, se trata de emular la parte de Libman (2010) acerca de incentivar el rol activo de las y los estudiantes con la sección *Preguntas para reflexionar*.

f. Las matemáticas y la estadística

Se recomienda consultar el Apéndice B para revisar las *Matemáticas elementales para la estadística*, ya que esto puede resultar útil al momento de hacer cálculos en los siguientes capítulos. Por ejemplo, se usan varios símbolos y operaciones matemáticas, tanto para describir los datos como para analizarlos en la estadística: Excel contiene varias funciones para llevar a cabo estas operaciones (Apéndice B). Por esta razón, en el Apéndice B se explica el símbolo griego Σ (sigma), que significa suma, junto con algunas de las operaciones elementales y reglas que se practican con él. Además, se explica cómo organizar los datos en una hoja de Excel (Tabla B-1 y Tabla B-2): *i. e.*, columnas para las variables y renglones para los participantes, objetos u observaciones. Aparte, un recurso muy empleado en estadística es el valor absoluto: *e. g.*, $|3|$ (véase: Apéndice B). Asimismo, operaciones de aritmética, indicadores especiales (raíz cuadrada, exponenciales), orden de las operaciones aritméticas y redondeo son mostrados tanto manualmente como en Excel.

g. Estadísticas vs. parámetros

Esta es la única parte de la presente obra que trata con cierta profundidad la estadística inferencial, porque se asume que las y los lectores eventualmente abordarán este tema cuando analicen con más profundidad el origen de sus datos (*i. e.*, una muestra de una población). En concreto, esta sección describe algunos principios de *estadísticas inferenciales* para expandir el panorama, pero no está dentro de los objetivos de este libro el ahondar en estas estadísticas. Dentro del contexto de la estadística inferencial, cuando se mencionan las *estadísticas* se está refiriendo a los valores que se obtienen de las muestras que son parte de una población. Paralelamente, cuando se mencionan los *parámetros* es que se está hablando de poblaciones. Para definir estos dos conceptos, Hinkle *et al.* (2003, p. 738) declararon: Una población incluye a todos los miembros de un grupo definido. Una población puede estar formada por personas (*e. g.*, estudiantes), observaciones (*e. g.*, calificaciones) u objetos (*e. g.*, escuelas). Por ejemplo, los estudiantes universitarios de México en 2019 (en este caso, se especificó el lugar y el tiempo). Una muestra es un subset de la población (*e. g.*, una porción de la población de estudiantes universitarios de México en 2019). Asimismo, una muestra puede estar formada por personas, observaciones u objetos.

Una muestra es representativa de una población de cierto tamaño cuando se toma al azar bajo los criterios de un nivel de confianza,[17] un intervalo de confianza[18] y se calcula con una fórmula cierto tamaño

17 El nivel de confianza dice qué tan seguro se puede estar de un evento (*i. e.*, es algo que sucede como al lanzar una moneda al aire). Se expresa como un porcentaje y representa la frecuencia con la que el porcentaje real de la población que elegiría una respuesta, se encuentra dentro del intervalo de confianza. El nivel de confianza del 95% significa que puede estar seguro al 95% (si se toman 100 muestras, 95 de ellas contendrán el parámetro de la población y 5 de ellas no lo harán); el nivel de confianza del 99% significa que puede estar seguro al 99% (si se toman 100 muestras, 99 de ellas contendrán el parámetro de la población y 1 de ellas no lo hará). Más aún, se ignora cuál de las muestras en particular no contendrá el parámetro de la población. La mayoría de las y los investigadores utilizan el nivel de confianza del 95%.

18 El intervalo de confianza (llamado margen de error) es la cifra más o menos que generalmente se informa en los resultados de las encuestas de opinión en periódicos o la televisión. Por ejemplo, si se utiliza un intervalo de confianza de 4 puntos para un resultado del 47% obtenido de una muestra, se puede estimar que si se hubiera formulado la pregunta a toda la población relevante entre el 43% (47 – 4) al 51% (47 + 4) habría elegido esa respuesta.

de la muestra (*n*). Uno de los propósitos de obtener una muestra representativa, es tener la posibilidad de generalizar en la población lo que se encuentre en esta primera (*e. g.*, promedio). En la Tabla 1.1 se muestran algunos símbolos de los parámetros de poblaciones y de estadísticas de muestras, así como sus respectivos cálculos en Excel. Está más allá del propósito del presente libro el ahondar en el muestreo; por ello, se recomienda consultar a Dattalo (2008) para análisis multivariados[19] y Blair y Blair (2015) para una introducción al tema.

Tabla 1.1 Símbolos para poblaciones y muestras

Parámetro/ Estadística	Parámetros de una población representada por letras griegas	Estadísticas de una muestra representada por símbolos y palabras	Excel 2016 (población): *fx*	Excel 2016 (muestra): *fx*
Promedio	μ (letra griega que se pronuncia mi o miu)	$\bar{x}$	= AVERAGE ($Celda_1$: $Celda_n$)	= AVERAGE ($Celda_1$: $Celda_n$)
Varianza	σ^2	Varianza	= VAR.P ($Celda_1$: $Celda_n$)	= VAR.S ($Celda_1$: $Celda_n$)
Desviación estándar	σ (letra griega que se pronuncia sigma)	*SD*	= STDEV.P ($Celda_1$: $Celda_n$)	= STDEV.S ($Celda_1$: $Celda_n$)
Tamaño de la muestra	*N*	*n*	Véase el Capítulo 3 para calcular las frecuencias totales	Véase el Capítulo 3 para calcular las frecuencias totales

Nota: *n* representa hasta qué celda va a abarcar la fórmula de Excel. Se recomienda ver el Apéndice A para familiarizarse con los recursos de Excel y sus fórmulas ofrecidas por Microsoft. Los temas de Promedio (Capítulo 4), Varianza y Desviación estándar (Capítulo 5), se tratan más adelante.

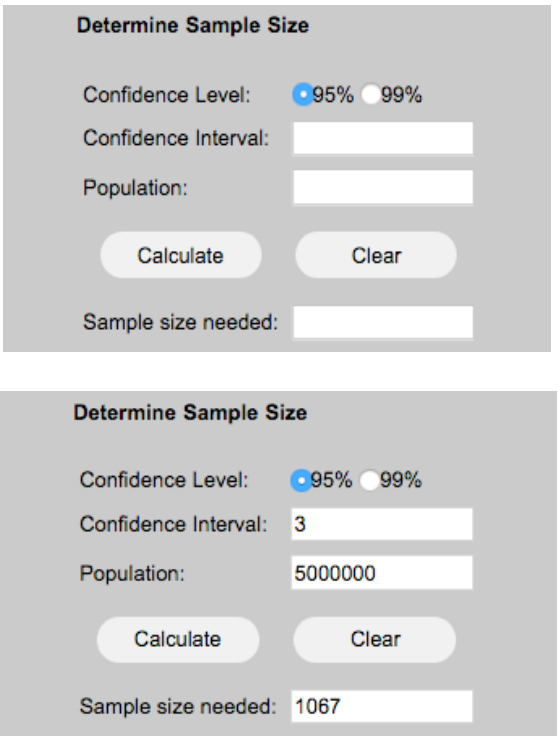
Un ejemplo de muestreo

Por ejemplo, se desea saber si las clases en línea son consideradas útiles para las y los estudiantes universitarios y se les pregunta con la previa instrucción de responder *sí* o *no*: ¿Son las clases en línea útiles para ti? En 2019, había una población de 5,000,000 de estudiantes universitarios en México, aproximadamente, según la ICEX España Exportación e Inversiones (2019). Para estimar el tamaño de una muestra de esta población de estudiantes, se va a la calculadora en línea (Creative

19 Los análisis multivariados involucran múltiples variables, que pueden ser varias variables independientes y una dependiente o varias variables independientes y dependientes, así como variables donde no existe relación de dependencia, sino que solo están relacionadas (véase: Hair *et al.*, 2019; Tabachnick, & Fidell, 2019).

Research Systems, 2012). Se puede introducir el tradicional nivel de confianza (*confidence level*) del 95%, con un intervalo de confianza de 3 puntos (*confidence interval*), y se presiona: *calculate* (Figura 1.1).

Figura 1.1 Tamaño de una muestra



El resultado es una muestra tomada al azar de 1,067 estudiantes universitarios. Suponiendo que a esta *n*, se le aplica la encuesta y el 70% contesta que *no*: *i. e.*, no considera las clases en línea útiles. Por el contrario, el resto (30%) responde que *sí*. Para la interpretación de estos resultados, se toma en cuenta el nivel de confianza y el intervalo de confianza. En este caso, el intervalo de confianza (*IC*), que se conoce como *margen de error*, son los puntos que se suman o se restan al porcentaje que se da (véase: Ponce-Renova, 2019; y Ponce-Renova, 2020, para estimaciones de *IC*; así como a Cumming, 2013). Esto es, se tenía un 70% de estudiantes (estadística) que dijeron *no* con un *IC* de

3 puntos, así que se esperaría que el parámetro de la población de estudiantes estaría entre 67% y 73% al contestar la pregunta antes mencionada. Adicionalmente, el nivel de confianza dice qué tan seguro se está. Esto es, si se tomaran 100 muestras de 1,067, un número de 95 de ellas hubiera capturado el parámetro de la población. Colocando estos dos conceptos juntos, se está con un 95% de confianza de que el verdadero valor de la población de estudiantes está entre 67% y 73%.

Excel 2016 y el muestreo

Excel 2016 *no* tiene una función[20] en automático que calcule un tamaño de la muestra como lo hace la calculadora de Creative Research Systems (2012). En contraparte, se puede tener un conjunto de datos que represente a una población de la cual se va a extraer una muestra de cierto tamaño al azar. Para ello, se usa la función: = RAND ().

Nota: Con esta función no es necesario especificar la celda, así que se queda el paréntesis vacío.

Con esta función, se pueden crear una serie de números al azar, ordenarlos y tomar una muestra de cierto tamaño. Por ejemplo, una docente tiene cinco alumnas y alumnos con su respectiva calificación de alguna materia (Figura 1.2). Solo tiene dos libros que le gustaría obsequiar al azar, así que utiliza Excel para crear cinco números al azar para cada calificación, para luego ordenarla de menor a mayor y regalar los libros a las o los estudiantes con el número al azar más pequeño (Figura 1.2, sexto paso en negrillas). En la Figura 1.2 se muestran los pasos para hacer esto.

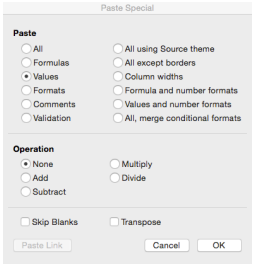
20 Una función (*fx*) es una fórmula predefinida que realiza cálculos utilizando valores específicos en un orden particular. Excel incluye muchas funciones comunes que se pueden emplear para encontrar rápidamente la suma, el promedio, el recuento, el valor máximo y el valor mínimo para un rango de celdas (véase el Apéndice A para más información de cómo usar las funciones de Excel y crear las propias).

Figura 1.2 Función para crear números al azar

1.° Escribir la función para obtener un número al azar/aleatorio. Luego, deslizar esta celda de la función para que calcule números al azar para el resto del set. Cada celda de Excel implica una coordenada definida por una letra (A, B, C,…, *n*) y un número (1, 2, 3,…, *n*).

	A	B	"B"
1	70	= RAND ()	0.7152
2	71	= RAND ()	0.3046
3	72	= RAND ()	0.9723
4	73	= RAND ()	0.1989
5	74	= RAND ()	0.8379

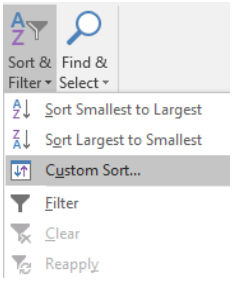
2.° Seleccionar y copiar las columnas A y B. Hacer clic derecho y seleccionar *Paste special*. Se abre la ventana mostrada aquí, se selecciona *Values* y se presiona OK.



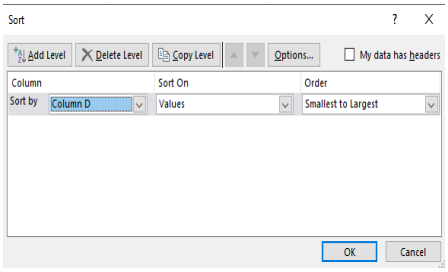
3.° Luego, las columnas A y B se pegan en otras como la C y la D. Asimismo, se seleccionan ambas columnas para ordenarlas con base en los números generados al azar (columna D).

	C	D
1	70	0.7152
2	71	0.3046
3	72	0.9723
4	73	0.1989
5	74	0.8379

4.° Para ordenar las dos columnas, se va a *Sort & Filter* y se selecciona *Custom sort*.



5.° Se abre esta ventana y se selecciona la columna D, así como del valor más pequeño al más grande: *Smallest to largest*.



6.° Una vez ordenados de menor a mayor, se pueden seleccionar las dos primeras calificaciones.

C	D
73	**0.1989**
71	**0.3046**
70	0.7152
74	0.8379
72	0.9723

Preguntas para resolver del Capítulo 1

- ¿Para qué sirve la estadística en la Investigación Educativa?
- ¿Qué es el aprendizaje?
- ¿Qué es un constructo?
- ¿Cuáles son las tres funciones de la estadística?
- ¿Para qué sirve la estadística descriptiva?
- ¿Cuáles son algunas de las diferencias que dividen a las estadísticas descriptivas de las inferenciales?
- ¿Cuál de estas dos estadísticas se podría usar primero?

Problema para resolver

Problema.[21] El escenario es que se tienen las calificaciones de estudiantes (N = 200) y se desea obtener una muestra (n) con un nivel de confianza (*confidence level*) del 95% y un intervalo de confianza (*confidence interval*) de 2 puntos, usando la calculadora de Creative Research Systems (2012).

1. ¿Cuál resultó ser el tamaño de la muestra?
2. ¿Cuál es la interpretación de este tamaño de la muestra?
3. ¿Qué pasa con el tamaño de la muestra si el nivel de confianza (*confidence level*) del 95% cambia a 99%?
4. Se mantiene el nivel de confianza (*confidence level*) al 95%, pero se cambia el intervalo de confianza (*confidence interval*) de 2 puntos a 5 puntos. ¿Qué pasa?
5. Hay que seleccionar 185 calificaciones usando la función = RAND (); obtener el promedio (véase el Capítulo 4 para detalles de las medidas de tendencia central) = AVERAGE (Celda$_1$: Celda$_n$); y la desviación estándar (véase el Capítulo 5 para detalles de las medidas de dispersión) = STDEV.S (Celda$_1$: Celda$_n$). Una vez calculadas estas estadísticas, se comparan con las de la población: $\mu = 45.85$ y $\sigma = 28.33$. ¿Son las estadísticas diferentes a los parámetros?, ¿qué significa si hay una diferencia?

21 Este problema corresponde al área de la estadística inferencial, porque se trata de inferir el posible valor de una población a partir de la estadística de una muestra.

Tabla P-1.1 Calificaciones de la población

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	33	39	21	39	93	73	51	40	29	64	48	71	63	78	67	7	16	32	14	6
2	24	3	71	57	93	67	67	44	46	62	24	34	31	58	18	73	51	61	10	31
3	9	81	22	55	45	89	36	3	65	13	11	88	91	22	51	22	41	0	90	65
4	44	24	46	56	25	36	24	81	3	34	25	9	9	79	72	17	9	80	43	17
5	15	25	60	92	12	99	45	56	90	80	25	65	25	29	78	34	35	84	7	94
6	84	56	16	18	68	67	13	51	11	41	59	20	27	65	98	55	67	12	53	99
7	50	24	65	4	96	3	78	41	70	99	65	54	88	63	69	43	64	48	16	42
8	25	1	90	39	21	83	4	63	33	45	14	72	10	97	40	31	42	98	26	66
9	26	64	2	45	88	72	84	7	38	60	63	8	12	45	4	6	1	97	54	60
10	68	2	31	56	15	79	18	90	33	61	43	15	25	47	16	99	23	81	73	6

Preguntas para reflexionar

- ¿En qué circunstancias sería apropiado hacer un estudio con solo estadísticas descriptivas?
- ¿En qué circunstancias no sería apropiado hacer un estudio con solo estadísticas descriptivas?
- ¿Cómo se puede usar la literatura para discutir los resultados de una investigación con estadísticas descriptivas?
- Si solo se hicieran investigaciones con estadísticas descriptivas, ¿qué podría pasar con el conocimiento?

Opinión del Autor

Las estadísticas descriptivas nos permiten tener una primera aproximación a un fenómeno como el aprendizaje, entre muchos otros más. Creo que las estadísticas descriptivas son una parte esencial en una investigación educativa para ser presentadas en nuestros manuscritos,[22] reportes, entendimiento de una situación y para tomar decisiones. Como posible resultado, estas nos ayudan a *convertirnos en uno mismo con los datos*. ¡Suena a canción! Con uno mismo quiero decir que

22 Un manuscrito se puede mejorar bastante al atender una serie de recomendaciones y guías que contiene el libro de la APA (2019), entre otras publicaciones.

uno va conociendo sus datos al saber cuántas observaciones hay o al identificar las variables y las escalas en las que fueron medidas; también, como darse cuenta de los datos perdidos[23] (los datos que por alguna razón debieron de aparecer, pero por alguna razón no están). Las estadísticas descriptivas nos permiten observar estas particularidades de nuestros datos.

Una recomendación para investigadoras e investigadores nuevos es que empiecen sus proyectos de investigación con preguntas descriptivas:

- ¿Qué es esto?
- ¿Cuáles son los niveles?
- ¿Cuáles son las frecuencias?
- ¿Cuántos valores están perdidos?
- ¿Cuáles son las medidas de tendencia central y de variación?
- ¿Qué muestran las gráficas?
- ¿Cuál es la distribución de los datos?, entre algunas preguntas más.

Esto puede ser aplicable tanto a diseños experimentales[24] como no-experimentales[25] que se estén considerando emplear. Luego, y conforme estén más familiarizadas y familiarizados con los datos y la literatura, pueden hacer preguntas de efecto: *e. g.*, ¿cuál es el efecto de más horas de estudio en las calificaciones? Dependiendo de los recursos para el proyecto, se pueden hacer interrogantes teóricas: ¿por qué pasa este fenómeno? Otra alternativa es poner a prueba una teoría al tratar de replicar algún modelo[26] teórico (*i. e.*, estructura de

23 Los datos perdidos (*missing data*) pueden alterar drásticamente las estadísticas descriptivas e inferenciales, así que se recomienda ver a Enders (2010) para los posibles remedios a esta situación.

24 Un diseño experimental involucra el comparar dos o más grupos diferentes, o un solo grupo, en dos o más ocasiones cuando se ha aplicado algún tipo de tratamiento (*e. g.*, tutorías, talleres, entre otros).

25 Un diseño no-experimental no cuenta con un tratamiento, pero sí se pueden hacer comparaciones (véase: Ponce-Renova, 2021), así como correlaciones (*e. g.*, Ponce-Renova, 2020). *Nota:* Para ambos diseños, se puede consultar a Hernández-Sampieri, & Mendoza (2018).

26 En general, un modelo es una representación de un concepto de forma gráfica, teórica o de otro tipo, que puede ser usado para varios propósitos investigativos y demostrativos, tales como la ampliación del conocimiento de un concepto o proceso al proponer hipótesis o mostrar rela-

las diferentes variables) por medio de *modelos de ecuaciones estructurales*[27] (*i. e.*, un modelo predeterminado que indica qué tan bien una teoría explica los datos. Incluso, se dice qué tan bien la teoría cabe en los datos: *how well a theory fits the data*).

Más en detalle, muchas veces ya tenemos una teoría preconcebida de cómo funcionan las cosas y forzamos para que estas entren en nuestras explicaciones sin considerar mucho todos los datos o hasta eliminando algunos de ellos. Un ejemplo de esto me sucedió en una defensa de tesis cuando una estudiante de maestría sustentaba su tesis y decía que cierto tratamiento dado a un grupo de niños había funcionado. Le pregunté que por qué tenía como tamaño de la muestra a 30 estudiantes de primaria en partes de su documento, pero en sus resultados solo aparecían 10. Ella contestó que solo había colocado en esa sección a los 10 estudiantes a los que el tratamiento sí les había funcionado y había omitido a aquellos a los que no les había funcionado. En pocas palabras, ella solo hablaba de un tercio de la muestra y deliberadamente dejó fuera aquellos resultados que no apoyaban su conclusión. Es decir, la estudiante hizo que los datos entraran en su teoría de que el tratamiento era efectivo; pero debió de ser lo contrario: la teoría debería de caber en/explicar todos los datos considerados.

En adición, creo que un buen procedimiento para las estadísticas, y en particular para las descriptivas, es el sugerido por Levine *et al.* (2021, p. 3):

a) Define los datos que uno quiere estudiar para resolver un problema o alcanzar un objetivo;
b) Recolecta los datos de la fuente apropiada;
c) Organiza los datos recolectados al desarrollar tablas;

ciones o algún patrón (*cf.* VandenBos, 2015, pp. 661-662). Por ejemplo, un modelo podría ser tan sencillo como: por cada hora de estudio por parte de un grupo de estudiantes, estos aumentarán sus calificaciones en un .50 de un punto. Matemáticamente esto sería: calificación = .50 + horas de estudio. Una alumna/alumno que estudió 7 horas tendría la siguiente calificación = .50 + 7 = 7.50.

27 Los modelos de ecuaciones estructurales son representaciones gráficas y matemáticas elaboradas con métodos estadísticos y algoritmos de computadora, para observar hasta dónde los modelos teóricos caben en los datos. También, se les conoce como modelos causales (véase: Byrne, 2016).

d) Visualiza los datos recolectados al desarrollar figuras (representaciones visuales); y
e) Analiza los datos recolectados, llega a conclusiones y presenta resultados.

CAPÍTULO 2

Variables y sus distribuciones

Hace ya algún tiempo un investigador estaba estudiando el fenómeno de la deserción escolar y el embarazo para su tesis de maestría. Tenía una base de datos donde aparecían los nombres de las jóvenes y su escolaridad. Él especulaba que el embarazo tenía un efecto en la deserción de mujeres menores de 18 años de edad. Entonces, el embarazo a temprana edad se *asociaba*, posiblemente, a que las jóvenes abandonaran la escuela antes de terminar la preparatoria. De hecho, no había identificado una teoría que explicara la posible relación entre las variables antes mencionadas. Esto último no quiere decir que no exista tal teoría en la literatura. Obtuvo la información de una base de datos construida por el personal de hospitales públicos antes de su proyecto de investigación. Él pudo identificar en la base de datos cuándo habían desertado las jóvenes, porque habían indicado los años de escolaridad, y cuándo habían dado a luz. Una variable era la escolaridad medida en años. La otra variable se midió por el embarazo: *i. e.*, todas las participantes estaban embarazadas, así que la codificó con el número 1. Luego calculó las medidas de tendencia central y de dispersión de ambas variables. La variable de la deserción le dio como resultado 8.5 años de escolaridad con una desviación estándar de 2.1 años, aproximadamente. Cuando quiso calcular las estadísticas del embarazo, se dio cuenta de que el promedio era 1 y la desviación estándar era 0. Este promedio solo le decía que todas las participantes estuvieron embarazadas, lo cual él ya sabía. La base de datos solo contenía jóvenes embarazadas. La varianza de 0 apoyaba el promedio de que todas estaban embarazadas: *i. e.*, ¡no había variación! Entonces, la relación aparente era que todas estaban embarazadas, así que todas desertaron. Esto no tenía mucho sentido, así que después de reflexionar concluyó que la variable del embarazo no era útil para el modelo, porque no variaba. Él necesitaba otra variable, como la edad a la que quedaron embarazadas, donde hubiera alguna variación para relacionarla con el grado académico en el que ocurrió la deserción (véase: Ponce-Renova, 2020, para más detalles sobre la correlación).

a. Variables

Las variables son propiedades o características de algún evento, objeto o persona, que pueden tomar diferentes valores o cantidades (Shultz *et al.*, 2020; *e. g.*, estatura, peso, ingreso, etcétera). En oposición a la variable que sí cambia, esta es una *constante*[1] que no lo hace. La varianza de una constante sería cero y la de una variable sería un número *positivo* mayor que cero. Como ya se había mencionado en el Capítulo 1, hay muchas variables observables en el ámbito de la educación (*e. g.*, edad y grado escolar, entre otras) y no son observables directamente (aprendizaje: constructo). Detrás de estas variables no-observables hay teorías. De una manera coloquial, se puede considerar a una teoría como un manual que explica el mecanismo de una máquina (como ejemplos de estas, se recomienda leer las Teorías del Aprendizaje: *e. g.*, Illeris, 2018; Schunk, 2012). En otras palabras, una teoría explica las relaciones de causa y efecto de un set de variables (véase: Byrne, 2016, para observar cómo se pone a prueba una teoría mediante modelos de ecuaciones estructurales).

Una clasificación de variables

Se pueden clasificar en tres grandes ramas: a) Categóricas (*i. e.*, nominales); b) Discretas; y c) Continuas:

> ***Variables categóricas***[2] *(nominales):* Expresan un atributo, como color, religión, género, etcétera. Dentro de este tipo de variable (género) hay dos niveles: *e. g.*, mujeres (un nivel) u hombres (otro nivel). No hay una jerarquía absoluta. Cuando se emplean números para codificar este tipo de variables solo se usan para obtener frecuencias, porcentajes o ciertas fracciones:

1 Una constante tiene un valor que no cambia, como 1, 2, 3, …*n*, entre otros; *n* indica el valor más grande del set.

2 Cuando se utilizan variables categóricas en ciertos análisis, como los de comparaciones de grupos (*e. g.*, Prueba *t*; véase: Ponce-Renova, 2021), es necesario que los grupos se excluyan mutuamente: *i. e.*, o se está en un grupo (*e. g.*, mujeres) o se está en el otro (hombres).

» Existen 505 estudiantes en la universidad con beca
» El 50% de las y los alumnos de posgrado están casadas y casados
» Hay tres veces más varones en un salón que damas (3:1; o 3/1)

Variables discretas: Implican secuencia y jerarquía. Una variable discreta, por ejemplo, es el número de alumnas y alumnos en un salón, porque solo puede medirse en números enteros. En este ejemplo un número como 2.5 estudiantes no tendría sentido, porque no puede haber media o medio estudiante. En pocas palabras, las variables discretas solo pueden tomar valores enteros. Sin embargo, tienen niveles que pueden tender al infinito.
Variables continuas: Implican jerarquía. Ahora, una variable continua puede ser el promedio de calificación por alumna o alumno en una materia. María puede tener 86.91 de promedio en matemáticas. Las variables continuas miden números, como calificaciones, edad, peso, etcétera. Los niveles de una variable continua son infinitos, porque siempre hay un número entre otros dos por pequeños que sean: *e. g.*, entre 2.01 y 2.02 está 2.011.

Otras escalas para las variables

Existen diferentes formas de medir variables a través de escalas. A continuación, se presenta un modo de medir variables de cuatro maneras: nominal, ordinal, intervalo y razón. En la Tabla 2.1 se muestran algunos de los alcances y limitaciones de estas escalas.

Escalas nominales: Se usan con variables categóricas (nominales). Las escalas nominales son para simplemente nombrar o categorizar variables, como género, religión, estado de residencia, etcétera. Estas escalas no implican un orden. De estas variables se puede decir que son el nivel más básico de medición. Se pueden utilizar en una encuesta para recabar información de estudiantes. Por ejemplo, una serie de preguntas para medir nominalmente una variable serían:

» ¿Cuál fue el lugar de nacimiento?

- ¿En qué colonia/fraccionamiento vive?
- ¿Qué tipo de familia tiene?, ¿extendida o nuclear?

Entre sus propiedades están (Hinkle *et al*., 2003):

- Los niveles se *excluyen* mutuamente: *e. g.*, género: o se es mujer u hombre.
- Los niveles no tienen un orden lógico.

Escalas ordinales: Los elementos en esta escala están ordenados de una manera jerárquica. La distancia entre los niveles no es necesariamente la misma (*i. e.*, un *ranking* donde hay un primero, segundo, tercer lugar, etcétera, pero la distancia entre estos puede ser diferente). Este *ranking* se emplea en las competencias, como las carreras, para otorgar las medallas de oro, plata y bronce, sin importar el tiempo que separó a los participantes de su llegada a la meta. Una pregunta de esta escala sería:

- ¿En qué lugar de la Olimpiada de Matemáticas quedaron las y los participantes de una secundaria?

Entre sus propiedades están (Hinkle *et al*., 2003):

- Los niveles se excluyen mutuamente: Se está en el primer lugar o en algún otro.
- Los niveles tienen un orden lógico: Primero, segundo, tercero, etcétera.
- Los niveles corresponden a cierta cantidad de algo: Un puntaje para estar en cierto lugar.

Un ejemplo de este tipo de escalas son los percentiles y rangos de percentil (véase: Capítulo 6).

Escalas de intervalo: Son escalas numéricas en las cuales los intervalos tienen la misma distancia entre uno y otro punto. No tienen un cero *verdadero*. Un ejemplo de esta escala sería una de coeficiente intelectual, en la cual una diferencia entre 90 y 100

puntos sería la misma que entre 120 y 130. Se podría pensar que existe un cero en esta escala, el cual significaría que una persona no tiene inteligencia; ¿cómo podría ser eso? Por ello, estas escalas no tienen un cero verdadero. Otro ejemplo: al medir una variable como la satisfacción con la escuela de una hija/hijo, una escala de este tipo puede mostrar 5 niveles: Muy insatisfecho, Insatisfecho, Neutral, Satisfecho y Muy satisfecho. Esto permite comparaciones en un grado (*e. g.*, satisfacción). Se supone que una escala así de la satisfacción subyace en una escala continua. Algunas preguntas de esta escala serían:

- ¿Cuál fue la calificación en lectura?
- ¿Cuál fue el puntaje en el examen de admisión?

Entre sus propiedades están (Hinkle *et al*., 2003):

- Los niveles se excluyen mutuamente.
- Los niveles tienen un orden lógico.
- Los niveles corresponden a cierta cantidad de algo.
- Existen diferencias iguales entre los niveles.
- El punto cero es solo otro punto en la escala.

Escalas de razón: Son las escalas que más información proveen. Tienen un cero verdadero que significa la total ausencia de alguna propiedad (*e. g.*, distancia, peso, edad, etcétera). Por ejemplo, el dinero destinado a algún gasto en la educación puede alcanzar el cero una vez que se haya terminado por completo. También, tienen un orden (jerarquía) y las distancias entre los intervalos son iguales. Algunas preguntas sobre esta escala serían:

- ¿Cuántos años tiene la o el estudiante?
- ¿Cuánto es el ingreso familiar?

Entre sus propiedades están (Hinkle *et al*., 2003):

- Los niveles se excluyen mutuamente.
- Los niveles tienen un orden lógico.

- Los niveles corresponden a cierta cantidad de algo.
- Existen diferencias iguales entre los niveles.
- El punto cero refleja la total ausencia de cierta característica.

Tabla 2.1 Resumen de las escalas

Característica/Escala	Nominal	Ordinal	Intervalo	Razón
El orden de los valores es conocido		√	√	√
Se puede contar: *i. e.*, frecuencias	√	√	√	√
*Se puede obtener la moda	√	√	√	√
*Se puede obtener la mediana		√	√	√
*Se puede obtener el promedio			√	√
Se puede cuantificar la diferencia entre cada valor			√	√
Se pueden sumar o restar valores			√	√
**Se pueden multiplicar y dividir valores				√
Tiene un cero verdadero				√
Ejemplos:	Lugar de nacimiento Nombre de escuelas, estados Tipo de clases: Español, Matemáticas Género	Nivel socioeconómico: (Trabajador, medio y alto) Grado en la escuela: (Primaria, secundaria, preparatoria, universidad) *Ranking* en el que están clasificadas las escuelas: 1.°, 2.° y 3.°	Las escalas de las encuestas están clasificadas como intervalo: Escala tipo Likert (1 = Totalmente en desacuerdo; 2 = En desacuerdo; 3 = Neutral; 4 = De acuerdo; 5 = Totalmente de acuerdo) Calificaciones La temperatura: Celsius o Fahrenheit Coeficiente intelectual	Peso, estatura, edad, distancia y tiempo

Fuente: http://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/
Nota: *Promedio, mediana y moda se ampliarán en el Capítulo 4. ***De facto*, sí se usan las multiplicaciones y divisiones con las escalas de intervalo y se recomienda consultar a Harpe (2015) para una discusión más a fondo. Para obtener estadísticas (*i. e.*, moda) de las escalas nominales, hay que codificar sus niveles a números para obtener frecuencias (véase: Capítulo 3).

b. Conversión de las escalas

Las escalas se pueden convertir unas en otras cuando se hace de una forma jerárquica. En la jerarquía, el nivel más alto de la medición como primer lugar sería para la escala de razón; el segundo, para la de intervalo; el tercero, para la de orden; y el cuarto, para la nominal. Como resultado, se puede ir de un nivel más alto a uno más bajo, pero no al revés. Por ejemplo, las escalas de razón se pueden convertir en escalas de intervalo, ordinales o nominales. Con algunos ejemplos sobre la conversión, en la Tabla 2.2 se muestra cómo se podrían llevar a cabo algunas transformaciones y poder ser usadas en la Investigación Educativa, así como el sentido en el que se podrían convertir unas escalas en otras.

Tabla 2.2 Conversión de variables

Jerarquía	El más alto / 1.°	2.°	3.°	El más bajo / 4.°
Variable	Razón	Intervalo	Orden	Nominal
Estatura de un salón de clases	172 cm 179 cm 183 cm		Menos alto = 1 Mediano = 2 Alto = 3	
*Encuesta de satisfacción con una escuela		1 = Muy insatisfecho 2 = Insatisfecho 3 = Neutral 4 = Satisfecho 5 = Muy satisfecho		Insatisfecho Satisfecho
Edad	12 años 14 años 17 años 25 años			Adolescente Adulto
Sentido de la conversión	→			

Nota: Solo se emplean tres participantes para la simplicidad del argumento. *Se recomienda que las encuestas tengan escalas de, por lo menos, cinco niveles, para poder realizar más tipos de operaciones matemáticas y análisis estadísticos (véase: Bandalos, & Finney, 2010).

c. Variables dependientes e independientes

Las variables dependientes[3] e independientes[4] pueden tener una relación causal, donde la independiente puede tener un efecto en la dependiente (para más información sobre causa y efecto, véase el Capítulo 10 de Hurley, & Watson, 2017). Por ejemplo, en un diseño experimental se pueden aplicar algunas tutorías para mejorar el aprendizaje de matemáticas en un grupo de estudiantes (grupo tratamiento) y a otro grupo no se le aplica el tratamiento (control). Se les compara al final del estudio con un examen para ver si han mejorado las y los alumnos que recibieron las tutorías. En este caso, la variable independiente fueron los grupos en conjunción con las tutorías que se aplicaron a uno de los grupos, mientras que la dependiente serían los posibles diferentes niveles de aprendizaje de ambos grupos.

Otra manera en la que se puede relacionar una variable independiente con una dependiente, es mediante una correlación o regresión en un diseño no-experimental.[5] Por ejemplo, si se tiene como variable independiente las horas que se estudia y la variable dependiente es medida por una calificación. Más específicamente, si se incrementan las horas de estudio, aumenta la calificación. En este caso, se tiene una relación positiva entre las variables: si una se incrementa, la otra también (véase: Ponce-Renova, 2020, para las correlaciones). Este tipo de relación entre una variable dependiente y una independiente puede ser representada gráficamente en un plano cartesiano.

3 El resultado que se observa pasa o cambia después de la intervención o variación de la variable independiente en un experimento. También, es el efecto que se desea predecir o explicar en una relación correlacional. La variable dependiente puede estar relacionada casualmente con la independiente (VandenBos, 2015, p. 298).

4 Es la variable en un experimento que es específicamente manipulada o se observó que pasó antes que la variable dependiente para evaluar su efecto o influencia. La variable independiente puede estar relacionada casualmente con la dependiente (VandenBos, 2015, p. 533).

5 Un diseño no-experimental es aquel en el cual no se manipulan las variables: *i. e.*, no existe un tratamiento que se haya implementado para cambiar algo (*cf.* Hernández-Sampieri, & Mendoza, 2018).

d. Escalas, variables, análisis y gráficas

Además de la descriptiva, los análisis en estadística se podrían colocar en dos grandes categorías cuando se habla de estadística paramétrica:[6] comparaciones de promedios de grupos y relaciones entre variables. Está más allá de los objetivos del presente libro el cubrir comparaciones de promedios de grupos y relaciones entre variables. Para tener una rápida introducción a estos temas recién mencionados, se recomienda consultar a Ponce-Renova (2020) y Ponce-Renova (2021). Más al respecto, en la Tabla A-2 del Apéndice A se muestra una serie de análisis inferenciales con sus correspondientes variables dependientes e independientes, así como las escalas en las que pueden aparecer estas. Asimismo, se indica si el análisis se puede llevar a cabo en Excel 2016, así como un enlace a un video de YouTube, donde se puede encontrar un ejemplo al respecto.

Preguntas para resolver del Capítulo 2

- ¿Qué son las variables y las constantes?
- ¿Qué son las variables categóricas, discretas y continuas?
- ¿Qué tienen en común y en qué se diferencian las escalas nominales, ordinales, de intervalo y de razón?
- ¿Qué diferencia a una variable independiente de una dependiente?

Problemas para resolver

Problema 1. Se tiene el siguiente set de calificaciones: 7.2; 7.5; 8.3; 8.6; 9.1; y 10. Ahora hay que colocarlas en una escala ordinal en Excel. Comenzando con la calificación más alta hasta la más pequeña.

1. ¿Cómo queda el orden?

6 Procedimientos estadísticos que se basan en los supuestos de la distribución de atributos de una población a la cual se está poniendo a prueba (*e. g.*, existe una distribución normal; VandenBos, 2015, p. 759).

Problema 2. Se tiene el siguiente set de calificaciones: 5.2; 6.3; 6.9; 7.2; 7.5; 8.3; 8.6; 9.1; y 10. Ahora hay que colocarlas en una escala nominal (Aprobatoria ≥ 7.0 vs. No-aprobatoria < 7.0). Para fines de simplicidad, se podrían codificar: 1 = Aprobatoria; y 0 = No-aprobatoria.

2. ¿Cuántas fueron calificaciones aprobatorias?

Preguntas para reflexionar

- ¿Qué variables son las más usadas en la Investigación Educativa?
- ¿En qué escalas comúnmente se miden estas variables?
- ¿En qué escalas serían más útiles estas variables?
- ¿Qué evidencia apoya el uso de los puntajes de encuestas como de intervalo?
- ¿Por qué los puntajes de las encuestas se utilizan como si fueran de razón en algunas publicaciones?

Opinión del Autor

Cuando se va a iniciar un proyecto de investigación, hay que tener muy en cuenta las escalas que se van a usar para medir las variables. Dependiendo de las escalas que se escojan serán las operaciones que se puedan ejecutar. Recomiendo que se trate de utilizar escalas de intervalo o de razón, porque se pueden hacer más análisis que con las otras escalas. Además, las escalas de razón y de intervalo se pueden convertir en nominales y ordinales con relativa facilidad.

Una vez tuve un estudiante de tesis que quería medir el ingreso de las y los alumnos universitarios, pero creyó que las y los estudiantes no le iban a decir la cantidad por cuestiones de seguridad. Él decidió que les preguntaría en forma ordinal: *e. g.*, de uno a dos salarios mínimos por día; de tres a cuatro salarios mínimos por día; de cinco a seis salarios mínimos por día; etcétera. Tal vez tenía razón en suponer que las y los participantes no se lo iban a decir exactamente, pero sus análisis requerían una escala de razón como hubiera sido la cantidad de ingreso. Como no fueron datos en una escala de razón, solo los pudo describir, sin que pudiera hacer los análisis. En pocas palabras, perdió

la oportunidad de emplear una escala de razón que era crucial para su estudio de regresión múltiple lineal.

Mi sugerencia es que, si es un tema sensible como el dinero, se haga una encuesta piloto para ver cómo reaccionan las y los participantes antes de descartar otras posibilidades. Después de todo se es una investigadora/investigador y hay que tener cierta evidencia empírica antes de tomar una decisión como perder información por no usar la escala apropiada para la medición.

CAPÍTULO 3

Organización y representación de los datos

"Aprender primero lo que uno puede hacer ayudará a trabajar más fácilmente y efectivamente" (Tukey, 1977, p. v). Lo anterior dependería de la motivación, conocimientos y recursos que se tengan para organizar y representar los datos. Posiblemente, si se tiene una idea de cómo organizar y representar datos, se puede trabajar más fácilmente con un conjunto de datos. También, habría que considerar que no todos los datos que se tengan pueden servir para cierto propósito y siempre se van a quedar variables sin medir. Por ello, y además de la creatividad personal, habría que ver los antecedentes de la literatura al respecto y usar libros metodológicos como el presente, entre otras posibles fuentes.

a. Codificación de los datos

El primer paso es tener los datos en una hoja de Excel (Figura 3.1). Ya sea que se haya adquirido el conjunto de datos (*i. e.*, se le conoce como base de datos) de una fuente externa o se haya construido por parte de la o el investigador, hay que conocerlo bien. Usualmente, los datos se organizan en: *columnas* para

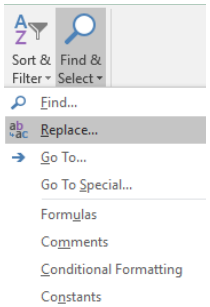
las variables y *renglones* para las y los participantes, objetos u observaciones (para más detalles, véase: Tabla B-2 del Apéndice B).

Una manera de codificar los datos es dar valores numéricos a los niveles nominales. Por ejemplo, en la siguiente tabla/hoja de cálculo (Figura 3.1) se muestra cómo hacerlo en Excel, donde se pueden codificar miles de datos con un solo clic: Según The Windows Club (2021), el límite de una hoja de Excel es 1,048,576 renglones y 16,384 columnas, pero habría maneras de salvar aún más datos en otras hojas. En este caso, solo una columna, la del género, fue codificada: Mujer = 1 y Hombre = 0. Esto no quiere decir que ya la escala nominal del género, se convirtió en otra escala. Solo se hace para simplificar y, más tarde, se mostrará cómo hacer frecuencias con estas codificaciones.

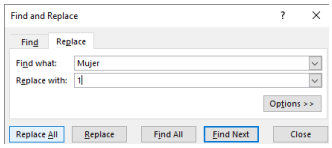
Figura 3.1 Codificando datos

	A	B	C	D	E	F
1	Nombre	Género	Edad	Examen de admisión	Parcial 1	Parcial 2
2	Susana	Mujer	19	120	74	76
3	María	Mujer	20	110	71	73
4	José	Hombre	18	145	75	78
5	Amelia	Mujer	17	135	74	77
6	Beto	Hombre	18	140	80	83

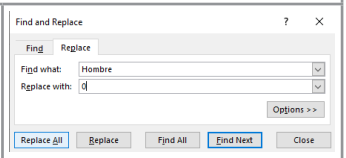
Datos originales



Hay que ir a *Find & Select* y seleccionar *Replace*.



Luego se coloca la palabra que se desea reemplazar: En este caso fue Mujer (reemplazada por 1) y se oprime *Replace all*.



Después se hace lo mismo con la palabra Hombre (reemplazada por 0).

Continúa...

	A	B	C	D	E	F
1	Nombre	Género	Edad	Examen de admisión	Parcial 1	Parcial 2
2	Susana	1	19	120	74	76
3	María	1	20	110	71	73
4	José	0	18	145	75	78
5	Amelia	1	17	135	74	77
6	Beto	0	18	140	80	83

Se reemplazó el género por números.

Nota: El conjunto de datos se considera una matriz de 6 x 6. El 6 es por el número de renglones (estudiantes, en este caso) y el 6 es el número de columnas (variables: Nombre, Género, Edad, Examen de admisión, Parcial 1 y Parcial 2).

Más allá de este ejemplo, otras variables se podrían codificar para resumir información e ir conociendo los datos. Por ejemplo, si una de las variables es el país de origen (escala nominal) y se tienen 50 naciones, puede ser que se clasifiquen por continentes (*e. g.*, con cinco continentes sería del 1 al 5; o con siete sería del 1 al 7) si esto tiene algún significado para la investigación en cuestión. O, tal vez, se pueda codificar si el país se considera desarrollado o no-desarrollado, entre otras muchas maneras. Estas escalas no dejarán de ser nominales, aunque se les represente con números.

b. Representaciones con tallos y hojas

Otra manera de representar los sets de datos es mediante una representación de tallos y hojas (*steams and leafs*). Esto da una idea de la distribución de los datos: *e. g.*, se concentran en el centro del set o en alguno de los extremos. En la Tabla 3.1 se muestra un set de 25 números que fueron representados con tallos y hojas en la Tabla 3.2. Para darle un contexto de Investigación Educativa, esos 25 números serían el número de aciertos que tuvieron 25 estudiantes. En este caso, los tallos fueron las decenas y las hojas, las unidades. Estos tallos y hojas se pueden ajustar de acuerdo con las características del conjunto de datos: *e. g.*, los tallos pueden ser las centenas y las hojas, las decenas.

Tabla 3.1 Conjunto de datos

	A	B	C	D	E
1	13	15	17	18	19
2	14	16	17	18	19
3	14	16	17	18	20
4	15	16	17	18	20
5	15	16	17	19	21

Tabla 3.2 Tallos y hojas

	A	B
1	Tallos	Hojas
2	1	3
3	1	4, 4
4	1	5, 5, 5
5	1	6, 6, 6, 6
6	1	7, 7, 7, 7, 7
7	1	8, 8, 8, 8
8	1	9, 9, 9
9	2	0, 0
10	2	1

Por otro lado, Excel 2016 no contiene una función automática para crear una presentación de tallos y hojas. El resultado de la representación de estos datos, en particular, indica que los valores se concentraron en el centro de la distribución.

c. Distribución de frecuencias

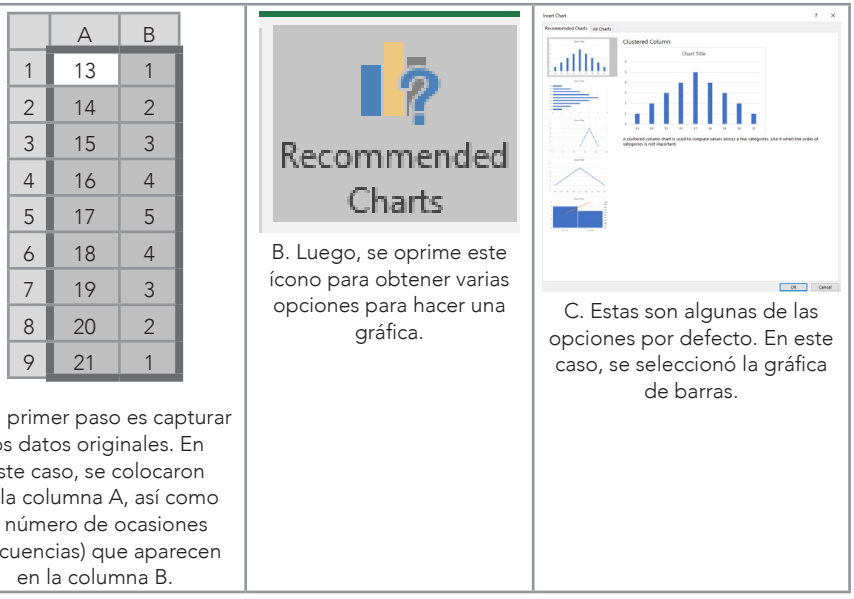
Una tabla de distribución de frecuencias es una tabulación que indica el número de veces que aparece un valor. Se puede utilizar con variables que tienen escalas nominales, ordinales, de intervalo y de razón. Se puede simplemente ordenar una serie de números, como los de la Tabla 3.1, de la siguiente manera (Tabla 3.3):

Tabla 3.3 Tabla de frecuencias

	A	B
1	Número de aciertos	Frecuencia (*f*)
2	13	1
3	14	2
4	15	3
5	16	4
6	17	5
7	18	4
8	19	3
9	20	2
10	21	1

La Tabla 3.3 fue transformada en gráficas de barras para observar la distribución.

Figura 3.2 Tablas de frecuencia y sus correspondientes gráficas de barras

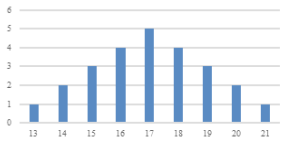


A. El primer paso es capturar los datos originales. En este caso, se colocaron en la columna A, así como el número de ocasiones (frecuencias) que aparecen en la columna B.

B. Luego, se oprime este ícono para obtener varias opciones para hacer una gráfica.

C. Estas son algunas de las opciones por defecto. En este caso, se seleccionó la gráfica de barras.

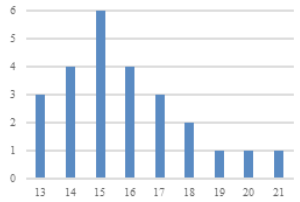
Continúa...



D. Una vez seleccionada una gráfica, se le puede colocar un título (Barras). Inclusive, se pueden cambiar los colores, el tipo de letra, la escala de los límites del plano, entre otros.

	A	B
1	13	3
2	14	4
3	15	6
4	16	4
5	17	3
6	18	2
7	19	1
8	20	1
9	21	1

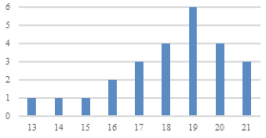
E. Un segundo conjunto de datos con sus respectivas frecuencias muestra una distribución de datos diferente.



F. Distribución concentrada hacia el lado izquierdo.

	A	B
1	13	1
2	14	1
3	15	1
4	16	2
5	17	3
6	18	4
7	19	6
8	20	4
9	21	3

G. Un tercer conjunto de datos con sus respectivas frecuencias muestra una distribución de datos diferente a las dos anteriores.



H. Distribución concentrada hacia el lado derecho.

Nota: Las distribuciones de frecuencias F y G son tratadas en el Capítulo 4 y en el Apéndice E bajo el tema de Asimetría/Sesgo.

d. Intervalos de clase

En la Tabla 3.3 se muestran los datos en varios grupos o clases como puntajes: *i. e.*, muestra el número en una columna y su frecuencia en otra. Por otro lado, hay maneras de resumir los datos y una de estas

es mediante intervalos de clase. Esto se hace al reducir el número de clases al agrupar varios puntajes dentro de cierto intervalo. Hinkle *et al.* (2003) explicaron cómo crear intervalos de clase bajo dos reglas:

1ª. Para el conjunto de datos grandes (100 o más valores) con un rango amplio de puntajes, se recomienda tener de 10 a 20 intervalos. Para sets de datos menores (menores a 100 valores), se sugiere de 6 a 12 intervalos. Por ejemplo, si se tienen 60 valores (del 1 al 60), se pueden hacer 12 intervalos de un rango de 5: *i. e.*, 60 / 12 = 5; del 1 al 5; del 6 al 10;…; del 56 al 60.

2ª. Cuando sea posible, la amplitud[1] del intervalo de clase (*i. e.*, diferencia entre el valor mayor y menor: *e. g.*, 5 – 1 = 4) debe ser un número par, para que exista un número entero en medio del intervalo: *e. g.*, en un intervalo del 1 al 5, el 3 es el punto medio que está dos unidades por encima del 1 y dos por debajo del 5. Un punto medio ayudaría a obtener otras maneras de describir los datos, como los percentiles, cuando se calculan manualmente (véase: Capítulo 6).

Tabla 3.4 Intervalos de clase

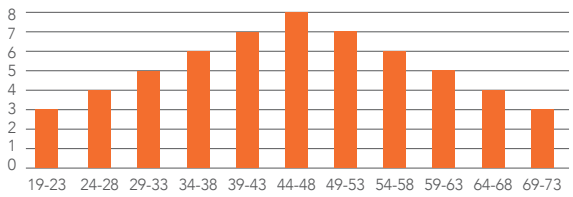
	A	B
1	Intervalo de clase	Frecuencia
2	19-23	3
3	24-28	4
4	29-33	5
5	34-38	6
6	39-43	7
7	44-48	8
8	49-53	7
9	54-58	6
10	59-63	5
11	64-68	4
12	69-73	3
13	Total de valores	58

1 Esta amplitud se conoce como un rango no-inclusivo.

En la Figura 3.3 se muestra cómo se capturaron los datos de la Tabla 3.4 en Excel para crear una gráfica de barras (véase: Figura 3.2 para los pasos). En este caso, los valores se concentran en el centro de la distribución, que aparenta ser una distribución normal.

Figura 3.3 Intervalos de clase

	A	B
1	Intervalo de clase	Frecuencia
2	19-23	3
3	24-28	4
4	29-33	5
5	34-38	6
6	39-43	7
7	44-48	8
8	49-53	7
9	54-58	6
10	59-63	5
11	64-68	4
12	69-73	3



e. Graficando datos

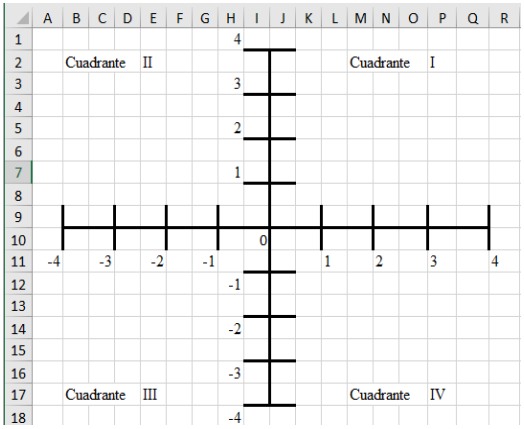
Se han mostrado varias gráficas de barras en las Figuras 3.2 y 3.3, donde se puede apreciar la distribución de los datos: centrados en el medio de la distribución o cargados hacia la izquierda o la derecha. Este tema de la distribución de los datos, se trata en el Capítulo 4, ya que se necesitan los conceptos de medidas de tendencia central, dispersión de datos, curtosis y asimetría para comprender este fenómeno.

El plano cartesiano

El plano cartesiano es muy útil para representar los datos de dos variables (Figura 3.1). Se le llama plano cartesiano por el filósofo y matemático de origen francés René Descartes, a quien se le atribuye el descubrimiento de la geometría analítica, así como el uso de este plano, que consta de dos ejes que forman un sistema de coordenadas

al ser perpendiculares entre sí (ángulos de 90°). Se intersectan en el origen (punto 0). La línea horizontal es representada por la *x* (abscisa) y la vertical, por la *y* (ordenada). Tiene cuatro cuadrantes: en el primero (I), los valores de ambas variables son positivos (usualmente en estadística descriptiva solo se usa este cuadrante, pero cabe la posibilidad de que se utilicen los demás); en el II, los valores de *x* son negativos, pero los de *y* son positivos; en el III, los valores de ambas variables son negativos; y en el IV, los valores de *x* son positivos, pero los de *y* son negativos.

Figura 3.4 Plano cartesiano



Nota: Este plano cartesiano fue creado manualmente en Excel para mostrar que también se puede dibujar, hasta cierto punto, con este *software*.

Uno de los usos del plano es ubicar un par de datos emparejados: *e. g.*, (*x*, *y*), donde la *x* va primero y la *y* después. Usualmente la *x* representa una variable independiente y la *y*, la variable dependiente, pero no necesariamente tienen esta denominación de independiente y dependiente. Los datos emparejados en una función serían: $y = x$. Una función es una operación matemática, donde *y* va adquiriendo ciertos valores conforme *x* cambia: *e. g.*, bajo la función de $y = x$, si $x = 1$, entonces $y = 1$. En este caso, los datos emparejados serían: (1, 1). Otro ejemplo de función sería: $y = 2x$, si $x = 1$, entonces $y = 2$; la operación fue: $y = 2(1)$, donde $x = 1$. Por lo tanto, los datos emparejados serían: (1, 2). Asimismo, el plano sirve para representar y analizar

diferentes figuras geométricas, como triángulos, elipses, parábolas, hipérbolas y circunferencias, entre otros.

Un ejemplo para graficar es cuando se tienen seis estudiantes y se está analizando la posible relación entre las horas que pasan estudiando para un examen y el número de aciertos en el mismo (Tabla 3.5). En este caso, la variable independiente son las horas que se invierten en estudiar y la variable dependiente son los aciertos que se obtuvieron en el examen. Podría advertirse un patrón desde la Tabla 3.5, en el cual las horas van de una unidad a otra (*i. e.*, 0, 1, 2, 3, 4 y 5) y los aciertos van de dos en dos unidades (0, 2, 4, 6, 8 y 10). Esto es lo que se conoce en matemáticas como una función[2] y tomaría la siguiente forma: *y* es igual a 2 por *x*; se puede escribir: $y = 2x$ o $f(x) = 2x$.

Tabla 3.5 Datos para el plano cartesiano: horas de estudio y aciertos

	A	B	C	D	E
1	Estudiante	Horas de estudio (variable independiente)	Numero de aciertos en un examen (variable dependiente)	Parejas de datos para el plano	Función $y = 2x$
2	José	0	0	(0, 0)	0 = 2 x 0
3	María	1	2	(1, 2)	2 = 2 x 1
4	Francisco	2	4	(2, 4)	4 = 2 x 2
5	Fernando	3	6	(3, 6)	6 = 2 x 3
6	Guadalupe	4	8	(4, 8)	8 = 2 x 4
7	Karla	5	10	(5, 10)	10 = 2 x 5

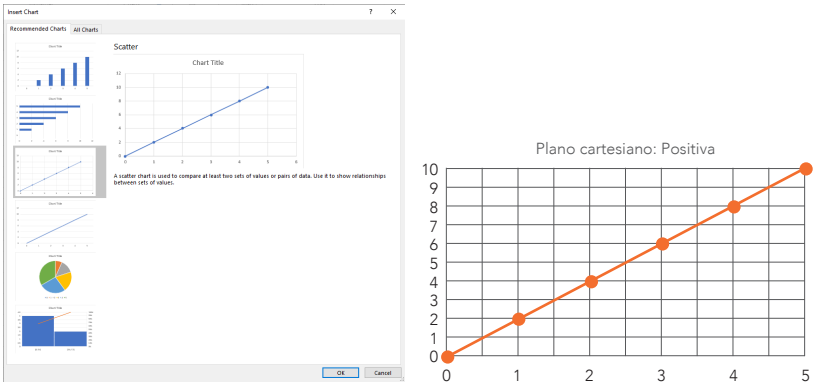
Fuente: Ponce-Renova (2019).

El patrón es evidente cuando se coloca en un plano cartesiano (Figura 3.5). En este caso, se puede observar que los datos emparejados se alinean perfectamente. Aunque en la realidad esto no suele suceder, sí sirve para ilustrar una situación ideal, en la que se dice que existe una relación positiva y perfecta entre las variables. La relación positiva se refiere a que si las horas de estudio aumentan también lo hacen los aciertos. Para el uso de crear esta gráfica, se siguen los pasos de la Figura 3.2 y de los ejemplos que se muestran; después de

2 Es una expresión, regla o ley que define una relación entre una variable (variable independiente) y otra variable (variable dependiente). Las funciones están en todas partes en las matemáticas y son esenciales para formular relaciones físicas en las ciencias (*Encyclopedia Britannica*, s.f.).

oprimir *Recommended charts*, se selecciona el plano cartesiano y se oprime OK.

Figura 3.5 Horas de estudio y aciertos



Nota: Este plano cartesiano es modificable al poder cambiar los colores de la línea, puntos y fondo, al igual que el tamaño y el tipo de letra, así como agregar letreros a las variables, entre muchas otras posibilidades. Se recomienda consultar la página de Microsoft (Apéndice A).

Siguiendo con el caso hipotético de la Tabla 3.5, las y los estudiantes ya obtuvieron el número de aciertos del primer examen. Ahora, de acuerdo con estos resultados, ellas y ellos hacen ajustes a las horas de estudio para el segundo examen acorde con lo obtenido en el primero (Tabla 3.6). La variable independiente son los aciertos en el primer examen, porque sucedieron primero y potencialmente pueden tener un efecto en las horas de estudio para el segundo (variable dependiente). Como se puede observar (Tabla 3.6), aquí también se da un patrón: i.e., cuando el número de aciertos incrementa, las horas de estudio disminuyen. Aunque no es tan obvia la función entre las variables, esta es: $y = 5 - 1/2x$.

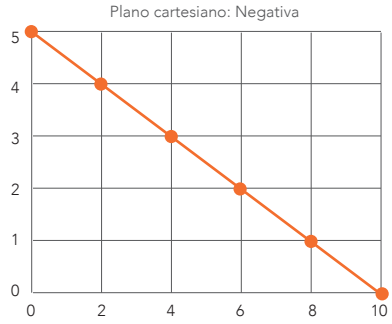
Tabla 3.6 Aciertos del primer examen y horas de estudio para el segundo examen

	A	B	C	D	E
1	Estudiante	Aciertos del primer examen (variable independiente)	Horas de estudio para el segundo examen (variable dependiente)	Parejas de datos para el plano	Función $y = 5 - 1/2x$
2	José	0	5	(0, 5)	5 = 5 – 1/2(0)
3	María	2	4	(2, 4)	4 = 5 – 1/2(2)
4	Francisco	4	3	(4, 3)	3 = 5 – 1/2(4)
5	Fernando	6	2	(6, 2)	2 = 5 – 1/2(6)
6	Guadalupe	8	1	(8, 1)	1 = 5 – 1/2(8)
7	Karla	10	0	(10, 0)	0 = 5 – 1/2(10)

Fuente: Ponce-Renova (2019).

Al igual que en la Figura 3.5, se siguen los pasos de la Figura 3.2 y se obtiene la Figura 3.6, en la que se muestra una relación negativa entre las variables. Esto es, cuando el número de aciertos del primer examen aumenta, las horas dedicadas al estudio disminuyen. Se podría teorizar que esto se debe a que las y los alumnos con los más altos puntajes confían en su habilidad para obtener buenos resultados en el segundo examen, sin tener que invertir tanto tiempo en el estudio. Por otro lado, las y los estudiantes que obtuvieron puntajes bajos en el primer examen, ahora están invirtiendo más tiempo para revertir sus resultados en el segundo examen.

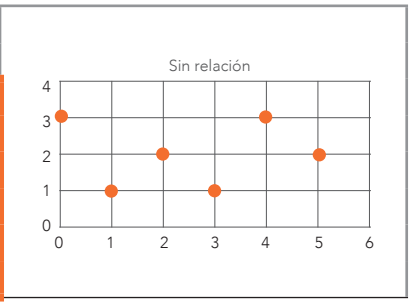
Figura 3.6 Aciertos del primer examen y horas de estudio para el segundo examen



Los siguientes cuatro casos muestran cuándo no existe una relación *lineal* entre la variable independiente, que son las horas de estudio para el primer examen (*x*), y la variable dependiente, que es el número de aciertos en el primer examen (*y*). En la Figura 3.7, se puede apreciar que no aparece un patrón lineal: ya sea que los datos formen una línea que se incremente o decremente. Algunos ejemplos de que no hay un patrón son: José, quien no estudió (cero horas), y Guadalupe, quien sí estudió (cuatro horas), obtuvieron el mismo número de aciertos (= 3); igualmente, María y Fernando, quienes obtuvieron el mismo número de aciertos, pero la primera estudió una hora y el segundo, tres horas; también, Francisco y Karla, quienes estudiaron dos y cinco horas, respectivamente, obtuvieron dos aciertos. En pocas palabras, no existe relación entre las horas de estudio y el número de aciertos obtenido. Se podría teorizar que algunas y algunos estudiantes saben más de la materia que otras y otros, y eso podría ser la razón. Otra posibilidad es que el examen no corresponde a lo que estudiaron. En fin, le correspondería a la investigadora o al investigador buscar más información al respecto para explicar lo sucedido.

Figura 3.7 Sin relación

	A	B	C
1	Estudiante	*x*	*y*
2	José	0	3
3	María	1	1
4	Francisco	2	2
5	Fernando	3	1
6	Guadalupe	4	3
7	Karla	5	2



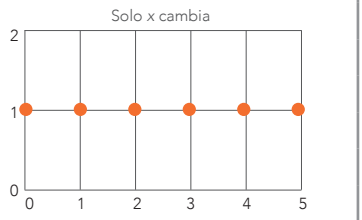
Nota: La correlación Producto-Momento de Pearson, mejor conocida como $r = 0$ (*i. e.*, está más allá de los alcances del presente texto el calcular *r*, pero para más información al respecto, véase: Hinkle *et al.*, 2003), donde: *x* = Horas de estudio; y *y* = Número de aciertos en un examen.

Otro escenario es cuando las y los alumnos estudiaron una diferente cantidad de horas, pero todas y todos obtuvieron el mismo resultado (Figura 3.8). En este caso, como en el anterior, no existe relación entre las variables, porque una sí cambia (*x*), pero la otra no (*y*), y para

que haya relación las dos deben de cambiar, no importando un patrón positivo o negativo.

Figura 3.8 *x* cambia, pero *y* es fija

	A	B	C
1	Estudiante	*x*	*y*
2	José	0	1
3	María	1	1
4	Francisco	2	1
5	Fernando	3	1
6	Guadalupe	4	1
7	Karla	5	1

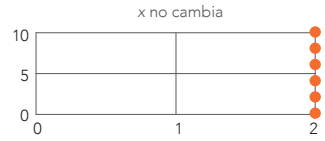


Nota: $r = 0$; x = Horas de estudio; y = Número de aciertos en un examen; $f(x) = a$.

Un escenario más es que todas y todos estudien la misma cantidad de tiempo (= 2 horas), pero que cada quien consiga un resultado de aciertos diferente. Tampoco, en este caso, existe relación, porque una de las variables está fija (x) y la otra cambia (y). En pocas palabras, no hay un patrón de relación. Una teoría podría ser que los resultados dependieron del conocimiento de las y los estudiantes, y no de las horas de estudio. Esto podría ser posible, pero en ese caso se tendrían que relacionar los aciertos de este examen con los de otros exámenes para ver si se forma algún patrón.

Figura 3.9 *x* es fija, pero *y* cambia

	A	B	C
1	Estudiante	*x*	*y*
2	José	2	0
3	María	2	2
4	Francisco	2	4
5	Fernando	2	6
6	Guadalupe	2	8
7	Karla	2	10

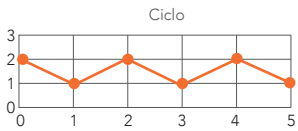


Nota: $r = 0$; x = Horas de estudio; y = Número de aciertos en un examen.

Finalmente, en la Figura 3.10 se muestra que, aunque sí existe un patrón, no es lineal. Explicando, las y los alumnos invirtieron diferentes cantidades de horas de estudio, pero sus aciertos suben y bajan como un ciclo, oscilando entre uno y dos. Se concluye que no existe un patrón lineal.

Figura 3.10 Ambas variables cambian, pero no existe relación lineal

	A	B	C
1	Estudiante	*x*	*y*
2	José	0	2
3	María	1	1
4	Francisco	2	2
5	Fernando	3	1
6	Guadalupe	4	2
7	Karla	5	1



Nota: Esta relación se podría modelar con una función trigonométrica (véase: Larson, & Edwards, 2010).

En el Apéndice C se muestran otras representaciones de datos a nivel *bivariado* (dos variables). Existen representaciones de datos a nivel multivariado donde se usan mapas (véase: Maciejewski, 2011).

Distribución de frecuencia acumulada y gráfica

Otra manera de describir un conjunto de datos es mediante la distribución de frecuencia acumulada, que se construye al ir sumando las frecuencias de los puntajes de los intervalos en cierto orden. Un ejemplo puede ser del intervalo de clase con el menor puntaje al intervalo de clase con el mayor puntaje. Antes de profundizar en el significado de la distribución de frecuencia acumulada, se usa Excel para calcular los valores (Tabla 3.7). La columna A contiene los intervalos de clase y la columna B, las frecuencias correspondientes. Lo primero que se calcula es la frecuencia acumulada (columna C): para el primer valor solo hay que escribir: =B2; para el resto de los valores, se comienza con escribir: =C2+B3, se selecciona esta celda y se desplaza hacia abajo, y se calcularán todos los valores (véase: Tabla 3.8 con los resultados).

Para calcular el porcentaje (columna D), se escribe la fórmula que corresponde a cada una de las frecuencias por intervalo de clase, se divide entre 58 (número total de frecuencias) y se multiplica por 100. Esto se hace con la fórmula: =B2/58*100 (la celda donde se escribió, se seleccionó y se desplazó hacia abajo para obtener todos los valores; véase: Tabla 3.8). Aquí también se puede cerciorar que tiene el total de las frecuencias al sumarlas con: =SUM(D2:D12).

Finalmente, el porcentaje acumulado (columna E) se obtiene al escribir primero: =D2 para el primer valor y luego con la fórmula: =D3+E2, seleccionarla y desplazarla hacia abajo para obtener todos los valores (véase: Tabla 3.8).

Tabla 3.7 Excel para distribución de frecuencia acumulada con intervalos de clase

	A	B	C	D	E
1	Intervalo de clase	Frecuencia	Frecuencia acumulada	Porcentaje	Porcentaje acumulado
2	19-23	3	=B2	=B2/58*100	=D2
3	24-28	4	=C2+B3	=B3/58*100	=D3+E2
4	29-33	5	=C3+B4	=B4/58*100	=D4+E3
5	34-38	6	=C4+B5	=B5/58*100	=D5+E4
6	39-43	7	=C5+B6	=B6/58*100	=D6+E5
7	44-48	8	=C6+B7	=B7/58*100	=D7+E6
8	49-53	7	=C7+B8	=B8/58*100	=D8+E7
9	54-58	6	=C8+B9	=B9/58*100	=D9+E8
10	59-63	5	=C9+B10	=B10/58*100	=D10+E9
11	64-68	4	=C10+B11	=B11/58*100	=D11+E10
12	69-73	3	=C11+B12	=B12/58*100	=D12+E11
13	Total de valores	58		=SUM(D2:D12)	

En corto, en la Tabla 3.8 se muestran los resultados de haber utilizado Excel. Un ejemplo hipotético de la interpretación de esta distribución de frecuencia acumulada, es que los intervalos de clase representan las edades de las y los alumnos que están tomando cursos de educación continua. Esto es, existen 3 estudiantes de entre 19-23 años en el curso que representan el 5.17 del alumnado y, por ser el primer grupo en la secuencia, representan asimismo el 5.17%. Ahora, una serie de preguntas que se podrían contestar es: ¿qué porcentaje de es-

tudiantes hay por debajo de cierto intervalo de clase? Ejemplificando, ¿qué porcentaje hay por debajo del intervalo de clase de 44-48 años (no incluyendo este intervalo)? La respuesta sería 43.10%.

Tabla 3.8 Resultados de la Tabla 3.7

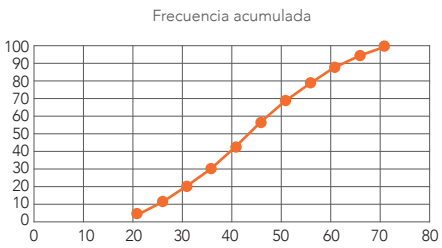
	A	B	C	D	E
1	Intervalo de clase	Frecuencia	Frecuencia acumulada	Porcentaje	Porcentaje acumulado
2	19-23	3	3	5.17	5.17
3	24-28	4	7	6.90	12.07
4	29-33	5	12	8.62	20.69
5	34-38	6	18	10.34	31.03
6	39-43	7	25	12.07	43.10
7	44-48	8	33	13.79	56.90
8	49-53	7	40	12.07	68.97
9	54-58	6	46	10.34	79.31
10	59-63	5	51	8.62	87.93
11	64-68	4	55	6.90	94.83
12	69-73	3	58	5.17	100
13	Total de valores	58			

Nota: Los porcentajes en este ejemplo son números aproximados.

Para este ejemplo, se les hace copiar y pegar las columnas A y E en un espacio diferente para crear la gráfica. Para graficar la distribución de frecuencia acumulada, se recomienda obtener el punto medio de los intervalos de clase para representarlos más fácilmente en Excel. Los intervalos de clase contienen el mismo rango de 4 unidades: 23 – 19 = 4; se divide el rango entre dos (4/2); se adicionan estas dos unidades al límite bajo del intervalo de clase (19 + 2 = 21) y se obtiene el punto medio. Para hacerlo en Excel, se tendría que manualmente capturar los límites bajos (LB) de los intervalos de clase y luego sumarles dos unidades (columna C): =B2+2. Una vez que se haya hecho esto, se seleccionan los datos y se crea la gráfica de frecuencia acumulada (Figura 3.11).

Figura 3.11 Obteniendo los puntos medios y con el porcentaje acumulado

	A	B	C	D
1	Intervalo de clase	LB	Punto medio	Porcentaje acumulado
2	19-23	19	21	5.17
3	24-28	24	26	12.07
4	29-33	29	31	20.69
5	34-38	34	36	31.03
6	39-43	39	41	43.10
7	44-48	44	46	56.90
8	49-53	49	51	68.97
9	54-58	54	56	79.31
10	59-63	59	61	87.93
11	64-68	64	66	94.83
12	69-73	69	71	100



Formas de las frecuencias de distribución

Se mencionan seis distribuciones en esta sección que pudieran ocurrir, pero en realidad los sets de datos pueden tomar una infinidad de formas cuando son graficados. Las diferentes formas dependen de cómo los puntajes se distribuyen en una escala de medición. En particular, la Tabla 3.9 contiene los datos para construir la Figura 3.12 (Distribución uniforme o rectangular). Un ejemplo de la Tabla 3.9 es cuando 14 estudiantes se distribuyen con la misma frecuencia (2) para cada una de las calificaciones que sucedieron. Ahora, de la Tabla 3.10 se obtuvo la Figura 3.12 (Distribución normal), que es el modelo teórico para varios tipos de análisis lineales (véase: Ponce-Renova, 2019; Ponce-Renova, 2020; Ponce-Renova, 2021, para una simple explicación acerca de la distribución normal). En esta distribución normal, los datos se concentran más en el centro (las más de las calificaciones con sus frecuencias) y después se distribuyen simétricamente hacia los extremos. Otra posibilidad está en la Tabla 3.11 relacionada con la Figura 3.12 (Positivamente asimétrica): las calificaciones se concentran en el lado izquierdo de la distribución. Tener una distribución asimétrica de los datos puede complicar análisis donde se comparen promedios (véase:

Ponce-Renova, 2021) o donde se correlacionen variables (véase: Ponce-Renova, 2020).

Tabla 3.9
Uniforme

	A	B
1	70	2
2	71	2
3	72	2
4	73	2
5	74	2
6	75	2
7	76	2
8	Total	14

Tabla 3.10
Normal

	A	B
1	70	1
2	71	2
3	72	3
4	73	4
5	74	3
6	75	2
7	76	1
8	Total	16

Tabla 3.11 Positivamente asimétrica

	A	B
1	69	3
2	70	4
3	71	5
4	72	4
5	73	3
6	74	2
7	75	1
8	Total	22

Nota: Para obtener el total, se aplica la fórmula de Excel: =SUM(B1:B7). A este total también se le conoce como tamaño de la muestra o *n*.

Siguiendo con las distribuciones, la contraparte de la Figura 3.12 (Positivamente asimétrica) es la Figura 3.12 (Negativamente asimétrica) [de la Tabla 3.12]. Los datos se concentran del lado derecho: en este caso, las calificaciones tienen más altas frecuencias del lado derecho. Asimismo, la Tabla 3.13 tiene la Figura 3.12 (Leptocúrtica), donde los datos se concentran altamente en el centro y no quedan muchos datos para los extremos. Esto sucedió en el ejemplo cuando la calificación del centro de la distribución fue la que más se repitió. La contraparte de esta distribución es la platicúrtica [de la Tabla 3.14 a la Figura 3.12 Platicúrtica], donde los datos forman una meseta y tienen una distribución muy uniforme. En este último caso, las frecuencias se distribuyeron más uniformemente que en el caso anterior.

Tabla 3.12
Negativamente

	A	B
1	70	1
2	71	2
3	72	3
4	73	4
5	74	5
6	75	4
7	76	3
8	Total	22

Tabla 3.13
Leptocúrtica

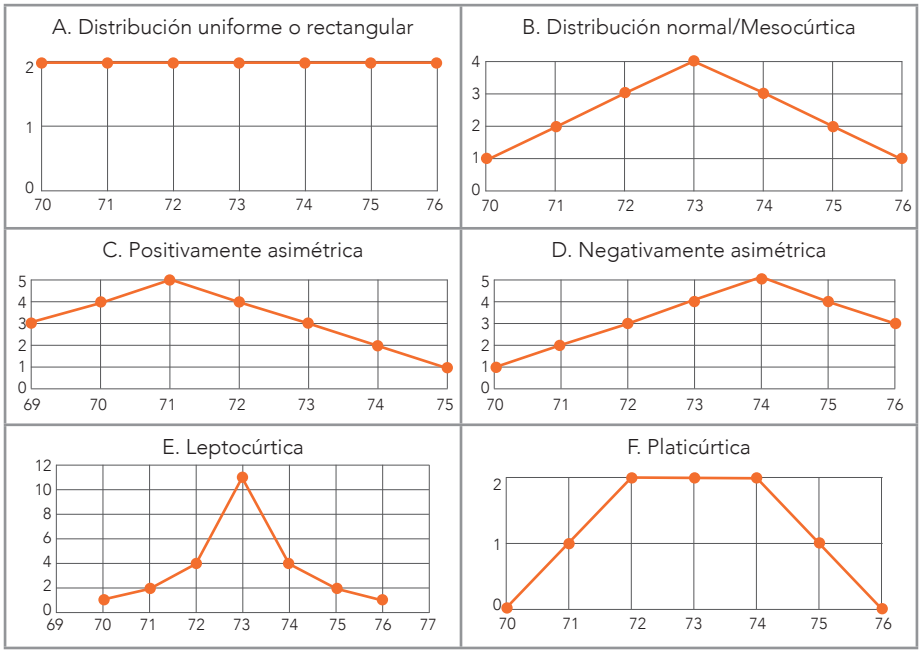
	A	B
1	70	1
2	71	2
3	72	4
4	73	11
5	74	4
6	75	2
7	76	1
8	Total	25

Tabla 3.14
Platicúrtica

	A	B
1	70	0
2	71	1
3	72	2
4	73	2
5	74	2
6	75	1
7	76	0
8	Total	8

Nota: Para obtener el total, se aplica la fórmula de Excel: =SUM(B1:B7).
A este total también se le conoce como el tamaño de la muestra o *n*.

Figura 3.12 Distribuciones



¿Cómo calcular las frecuencias de un conjunto de datos?

Hasta el momento las frecuencias han sido dadas para la serie de ejemplos de este capítulo. Por el contrario, cuando se comience a recolectar datos se encontrarán las frecuencias en los sets de datos. Para

contar desde un pequeño set hasta miles de datos, Excel tiene una función que cuantifica las frecuencias de los números.

Un ejemplo de conteo de frecuencias es el siguiente: el conjunto de datos de 16 x 2 (16 renglones por 2 columnas A y B).[3] El procedimiento:

1°. Para usar el procedimiento de conteo de frecuencias de Excel, el primer paso sería escribir los números de los cuales se desean obtener las frecuencias (columna D). En este caso, hay el interés de calcular las frecuencias de todos los números (71 al 86), pero se pueden seleccionar algunos números en particular si solo se desean ciertas frecuencias.

Tabla 3.15 Primer paso

	A	B	C	D	F
1	71	78		71	
2	72	78		72	
3	73	79		73	
4	73	79		74	
5	73	79		75	
6	73	79		76	
7	73	80		77	
8	75	80		78	
9	75	80		79	
10	75	80		80	
11	75	80		81	
12	76	81		82	
13	76	82		83	
14	76	86		84	
15	77	86		85	
16	78	86		86	
17				Total	

2°. En este caso, se oprime la celda F1 y se selecciona todo el espacio donde aparecerán las frecuencias (*i. e.*, columna F:

3 Para identificar una matriz por sus dimensiones, se coloca tradicionalmente, primero, el número de renglones y luego el número de columnas (véase: Levine *et al.*, 2021).

del F1 al F16; Tabla 3.16). Entonces, se escribe la siguiente fórmula en la función con la coordenada en la que comienza el set y donde termina (A1:B16), así como el intervalo de los números de los cuales se desean obtener las frecuencias (D1:D16): =FREQUENCY(A1:B16,D1:D16).

Tabla 3.16 Segundo paso

	A	B	C	D	F
1	71	78		71	=FREQUENCY(A1:B16,D1:D16)
2	72	78		72	
3	73	79		73	
4	73	79		74	
5	73	79		75	
6	73	79		76	
7	73	80		77	
8	75	80		78	
9	75	80		79	
10	75	80		80	
11	75	80		81	
12	76	81		82	
13	76	82		83	
14	76	86		84	
15	77	86		85	
16	78	86		86	
17				Total	

Nota: No se oprime *Enter* por el momento.

3°. Se oprime en medio de la fórmula para que el cursor quede en medio de la palabra: =FREQU|ENCY(A1:B16,D1:D16). El símbolo | representa al cursor que se encuentra entre la U y la E de la palabra FREQUENCY en la fórmula (el cursor debe de partir la palabra en algún lugar). Una vez que se está allí, se oprimen las siguientes tres teclas: *Control*, *Shift* y *Enter*. Entonces, las frecuencias de cada uno de los valores aparecen en la columna F (Tabla 3.17).

Tabla 3.17 Tercer paso

	A	B	C	D	F
1	71	78		71	1
2	72	78		72	1
3	73	79		73	5
4	73	79		74	0
5	73	79		75	4
6	73	79		76	3
7	73	80		77	1
8	75	80		78	3
9	75	80		79	4
10	75	80		80	5
11	75	80		81	1
12	76	81		82	1
13	76	82		83	0
14	76	86		84	0
15	77	86		85	0
16	78	86		86	3
17				Total	32

Nota: Aparecerán unas llaves: { }. El total es una simple suma de las frecuencias: =SUM(F1:F16).

Preguntas para resolver del Capítulo 3

- ¿Cuáles son algunos de los beneficios al codificar datos en escalas nominales?
- ¿Tiene alguna desventaja este tipo de codificación?
- ¿Qué puede indicar una representación gráfica de tallos y hojas?
- ¿Cuántos tipos de distribuciones y sus nombres se muestran en este capítulo?
- Si se tiene un conjunto de datos de 1,000 valores, ¿en cuántos intervalos sería conveniente dividirlo?
- ¿Cuál es, posiblemente, el cuadrante más usado en el plano cartesiano?
- Si se tiene la siguiente función: $y = 4x$ (con el rango de datos de x del 1 al 10), ¿cuáles serían los valores de y?
- ¿Qué tipos de relaciones pueden aparecer?

» ¿Para qué se pueden utilizar las frecuencias acumuladas?

Problemas para resolver

Problema. Este ejercicio implica pensar como una o un docente, donde se tiene un conjunto de datos y se desea analizarlo. En Excel, se crea un conjunto de datos en una matriz de 30 x 7: *i. e.*, 30 renglones y 7 variables: Nombre o identificación, Edad, Dirección postal, Género, Parcial 1, Parcial 2 y Examen final.

1. Asegúrese de que haya variedad entre los datos.
2. Codifique las variables de Dirección postal y Género: *e. g.*, para la Dirección postal: 0 = Área periférica y 1 = Área centro; para el Género: hombre = 0 y mujer = 1.
3. Cree representaciones gráficas de tallos y hojas (Tabla 3.2) para las variables de Parcial 1, Parcial 2 y Examen final.
4. De las variables Parcial 1, Parcial 2 y Examen final, logre que ninguno de los valores de estas variables se repita.

    a) Cree una relación positiva (Figura 3.5) entre el Parcial 1 y el Parcial 2, aunque no sea perfecta (Apéndice C; Figura C-5).
    b) Cree una relación negativa (Figura 3.6) entre el Parcial 2 y el Examen final, aunque no sea perfecta (Apéndice C; Figura C-6).
    c) Cambie los valores para que no exista relación (Figura 3.7) entre el Parcial 1 y el Parcial 2.

5. Ahora cambie el conjunto de datos para que algunos valores se repitan dentro del Parcial 1, Parcial 2 y el Examen final.

    a) Cree intervalos de clase y calcule la frecuencia por cada clase, el porcentaje y el porcentaje acumulado (Tabla 3.7).
    b) Calcule los puntos medios de los intervalos de clase y junto con el porcentaje acumulado realice una gráfica (Figura 3.11).

6. Para hacer diferentes formas de distribuciones, hay que cambiar los datos.

   a) Cree una distribución uniforme con el Parcial 1 y grafique (Tabla 3.9 y Figura 3.12).
   b) Lleve a cabo una distribución normal con el Parcial 2 y grafique (Tabla 3.10 y Figura 3.12).
   c) Elabore una distribución positivamente asimétrica con el Examen final y grafique (Tabla 3.11 y Figura 3.12).
   d) Efectúe una distribución negativamente asimétrica con el Parcial 1 y grafique (Tabla 3.12 y Figura 3.12).
   e) Realice una distribución leptocúrtica con el Parcial 2 y grafique (Tabla 3.13 y Figura 3.12).
   f) Haga una distribución platicúrtica con el Examen final y grafique (Tabla 3.14 y Figura 3.12).

7. Utilizando los datos que haya empleado en el número 6 (del ejercicio a al c, con el Parcial 1, Parcial 2 y Examen final), calcule las frecuencias del conjunto de datos usando las Tablas 3.15, 3.16 y 3.17.
8. ¿Qué ha aprendido de representar y manipular estos sets de datos?
9. ¿Le ve alguna utilidad en la vida académica?

Preguntas para reflexionar

- ¿Qué codificaciones se usan en la literatura de la Investigación Educativa?
- ¿Por qué es importante saber cuál es la distribución de los datos?
- Si uno encuentra los datos en intervalos de clase, ¿qué información se está perdiendo?
- Además de un plano cartesiano, ¿cómo se podría representar una relación entre variables?
- ¿Qué otras funciones, aparte de la lineal, existen en la literatura?

Opinión del Autor

Hay algunos psicólogos que han dicho algo así como que somos esclavos de nuestras propias experiencias. En otras palabras, vivimos en el mundo de la anécdota personal y en algo que alguien más nos contó. Sin embargo, aquí esta una cita textual que nos podría liberar de nuestra propia experiencia: "Es un impulso estadístico: El reconocimiento de lo inadecuado de nuestras experiencias personales y evidencia de anécdotas nos lleva a desear una base de toma de decisiones en una deliberada recolección de datos" (Wild, & Pfannkuch, 1999, p. 227). Concuerdo con estos autores y añadiría que el comenzar a trabajar con las representaciones gráficas, frecuencias, funciones y relaciones lineales nos abre la puerta para rascar la superficie del entendimiento de algún fenómeno y así comenzar a escapar de la anécdota personal. Por ello, hago énfasis en llevar a cabo los aspectos cubiertos en este capítulo.

CAPÍTULO 4

Medidas de tendencia central

Algunas personas piensan que para cantar son muy superiores al promedio y se presentan en un programa televisivo de concurso esperando ganar el premio al mejor artista. Pero se llevan un chasco cuando las y los jueces sin misericordia critican su mala actuación y, por lo tanto, destrozan ese sueño juvenil de ser una destacada o un destacado intérprete de las canciones populares. Tal vez el talento para cantar es como otras habilidades y características de los seres humanos que se distribuye normalmente. Esto es, si la habilidad para cantar se midiera con una regla de 100 centímetros, la mayoría de las personas estarían en el centro de la regla: como en el centímetro 50, aproximadamente; mientras que las personas con gran habilidad, como las que cantan ópera, se acercarían al centímetro 100, uno de los extremos. Por otro lado, las personas con muy escasa habilidad se acercarían al centímetro 0. Algo que se debe de tomar en cuenta para poder interpretar el promedio, es cómo se distribuyen los datos en estas reglas de medición que hemos creado para ser usadas en la Investigación Educativa, como calificaciones, puntajes en exámenes de admisión, coeficiente intelectual, entre muchas otras.

a. Medidas de tendencia central

Hablando de estadística descriptiva, Hopkins y Weeks (1990) dijeron que hay cuatro momentos para describir la distribución de un conjunto de datos:

- 1.° Las medidas de tendencia central;
- 2.° Las medidas de variabilidad;
- 3.° La asimetría; y
- 4.° La curtosis.

En este Capítulo 4, se tratan los 1.°, 3.° y 4.° momentos, y en el Capítulo 5, se trata el 2.° momento.

Un fenómeno de la naturaleza es que los valores de una escala, se concentren en medio de esta: esto se puede representar con una *distribución normal*[1] (véase: Ponce-Renova, 2019; y Ponce-Renova, 2020, para más detalles sobre esta distribución normal y la distribución normal estándar, que también se conoce como distribución normal estandarizada[2]). Por ejemplo, Frost (2021) declaró:

> La distribución normal es la más importante distribución de probabilidad[3] en estadísticas porque puede representar a muchos fenómenos naturales. Por ejemplo, la altura de las personas, su presión arterial, los errores en mediciones y los puntajes del coeficiente intelectual siguen una distribución normal. También, se le conoce como la Campana de Gauss. (Párr. 1)

En la Investigación Educativa, además del coeficiente intelectual, existen variables (constructos) con potencial para tener una distribución aproximadamente normal, como el aprendizaje y la motivación (véase: Dumont, Istance, & Benavides, 2010), entre otras. En el

1 Es una distribución teorética en la cual los valores se apilan en el centro, que es el promedio, y el resto se va hacia los dos extremos (VandenBos, 2015, p. 715).

2 Es una distribución normal cuyos valores han sufrido una transformación que resulta en un promedio de 0 y una desviación estándar de 1. De la misma manera, se le conoce distribución normal estándar (VandenBos, 2015, p. 1025).

3 Una distribución de probabilidad representa qué tan probable es un dato/evento: la frecuencia de un evento es dividida entre la frecuencia total. Por ejemplo, en la Figura 3.12 se muestra una serie de distribuciones (aunque está más allá de discutir la probabilidad en estos capítulos sí puede servir para darse una idea de que en la distribución uniforme todas las calificaciones tienen la misma frecuencia: (2) / frecuencia total (14) = 2/14 = 1/7; así que si un docente selecciona al azar alguna de las calificaciones, todas tendrían la misma probabilidad de ser seleccionadas: 1/7). Ahora, la distribución normal tiene diferentes frecuencias (véase: Figura 3.12). Si el docente selecciona al azar una calificación, la más probable sería 73, porque tiene la frecuencia más alta: 4. Para ser exactos, la probabilidad de 73 es 4/16 = ¼. Para más información acerca de la distribución normal, véase: Ponce-Renova (2019) y (2020).

Apéndice D se muestra cómo construir una distribución normal estándar en Excel, a partir de un conjunto de datos.

b. Promedio

También se le llama media y se suele representar por una $\bar{x}$ (equis barra) cuando se trata de una *muestra*[4] (*i. e.*, una estadística) y la letra griega μ (miu o mi) para una *población*[5] (un parámetro). Representa la suma de todos los valores (*e. g.*, calificaciones en una materia) dividida por el número de valores (tamaño de la muestra = *n* [también llamada frecuencia total]). La siguiente es la fórmula para el promedio de una muestra (Ecuación 4.1):

$$\bar{x} = \sum_{i=1}^{n} x_i / n \qquad \text{Ecuación 4.1}$$

Donde:

n = Número de valores o frecuencia total
Σ = Suma (véase: Apéndice B para más detalles acerca de este operador de suma)
x_i = Cada uno de los valores del set
i = Primer número en la suma

Desarrollando la Ecuación 4.1, se vuelve:

$$\bar{x} = \frac{(x_1 + x_2 + \ldots + x_n)}{n}$$

Donde:

$\bar{x}$ = Promedio de la muestra
$x_1, x_2, \ldots, x_n$ = Lista de valores del set
n = Tamaño de la muestra = Frecuencia total

4 Es una parte de una población.
5 Es el número total de individuos (humanos u otros organismos) en un área geográfica dada. En estadística, una definición teórica sería: grupo completo de objetos (gente, animales, instituciones, de los cuales una muestra con observaciones empíricas es obtenida para hacer generalizaciones). También, se le denomina universo (VandenBos, 2015, p. 808).

Fórmula para el promedio de una población (Ecuación 4.2):

$$\mu = \frac{(x_1 + x_2 + \ldots + x_N)}{N} \qquad \text{Ecuación 4.2}$$

Donde:

μ = Promedio de la población
$x_1, x_2, \ldots, x_n$ = Lista de valores de la población
N = Tamaño de la población

Para un ejemplo con la Tabla 4.1, se sigue la Ecuación 4.1:

$$\bar{x} = \frac{70 + 71 + 72 + 73 + 74 + 75 + 76}{7} = 73$$

Ahora, en Excel, se usa la fórmula con la Tabla 4.1: =AVERAGE(A1:A7)

Tabla 4.1 Promedio

	A
1	70
2	71
3	72
4	73
5	74
6	75
7	76
8	=AVERAGE(A1:A7)

Nota: El resultado de Excel es también 73.

Propiedades del promedio

Siendo probablemente la medición más empleada para describir la distribución de un conjunto de datos, se demuestran dos de sus propiedades (*cf.* Hinkle *et al.*, 2003), además de cómo se hacen los cálculos en Excel (Tabla 4.2):

1.ª La suma de las desviaciones del promedio dan cero. Esto es, se calcula una diferencia entre los valores del set y su promedio para luego sumarlos. La fórmula de la diferencia es:

$$x_{i \text{ de la diferencia}} = (x_i - \bar{x}) \qquad \text{Ecuación 4.3}$$

Donde:

x_i = Cada uno de los valores del set

$\bar{x}$ = Promedio del set

Tabla 4.2 Fórmulas en Excel

	A	B	C	D	E	F
1		x_i	Frecuencia	$x_{i \text{ de la diferencia}} = (x_i - \bar{x})$	$x_{i \text{ de la diferencia}}{}^2 = (x_i - \bar{x})^2$	$(x_i - 70)^2$
2		70	{=FREQUENCY (B2:B8,B2:B80}	=B1-C$11	=(B1-C$11)^2	=(B1-70)^2
3		71		=B2-C$11	=(B2-C$11)^2	=(B2-70)^2
4		72		=B3-C$11	=(B3-C$11)^2	=(B3-70)^2
5		73		=B4-C$11	=(B4-C$11)^2	=(B4-70)^2
6		74		=B5-C$11	=(B5-C$11)^2	=(B5-70)^2
7		75		=B6-C$11	=(B6-C$11)^2	=(B6-70)^2
8		76		=B7-C$11	=(B7-C$11)^2	=(B7-70)^2
9	**Σ**	=SUM(B2:B8)		=SUM(D2:D8)	=SUM(E2:BE)	=SUM(F2:F8)
10	*n*		=SUM(C2:C8)			
11	$\bar{x}$	=AVERAGE(B2:B8)				

Nota: Para calcular las frecuencias hay que seguir las instrucciones de las Tablas 3.15 a 3.17 (columna C). Para la columna D, se fija el promedio para que no cambie con el símbolo $ de esta manera: C$11. Asimismo, el símbolo ^ sirve para elevar a alguna potencia un número base o alguna operación: en este caso es al cuadrado.

Al final, se suman todas estas diferencias y resultan en cero (Tabla 4.3 [columna D en negrillas]). Entonces, se demuestra esta primera propiedad al sumar las diferencias entre cada uno de los valores del set y el promedio: *i. e.*, (-3) + (-2) + (-1) + 0 + 1 +2 + 3 = 0.

Tabla 4.3 Resultados de las fórmulas en Excel

	A	B	C	D	E	F
1		x_i	Frecuencia	$x_{i \text{ de la diferencia}} = (x_i - \bar{x})$	$x^2_{i \text{ de la diferencia}} = (x_i - \bar{x})^2$	$(x_i - 70)^2$
2		70	1	-3	9	0
3		71	1	-2	4	1
4		72	1	-1	1	4
5		73	1	0	0	9
6		74	1	1	1	16
7		75	1	2	4	25
8		76	1	3	9	36
9	Σ	511		**0**	**28**	**91**
10	*n*		7			
11	$\bar{x}$	73				

2.ª La suma de las desviaciones del promedio al cuadrado[6] es el número mínimo que se puede obtener al compararlo con cualquier número del set en lugar del promedio. Primero, se calculan las desviaciones del promedio al cuadrado mediante la fórmula:

$$x^2_{i \text{ de la diferencia}} = (x_i - \bar{x})^2 \qquad \text{Ecuación 4.4}$$

Solo difiere de la Ecuación 4.3 en que la Ecuación 4.4 está elevada al cuadrado. Luego, se suman estas diferencias al cuadrado (Tabla 4.3 [columna E en negrillas]). Se pudo haber seleccionado cualquier número del set, excepto el promedio, pero se seleccionó el 70 para ilustrar esta propiedad. Se sustituye el 70 por el promedio de 73 y se usa en la Ecuación 4.4 (Tabla 4.3 [columna F]). Para la comparación, 28 < 91, así que esta evidencia apoya la 2.ª Propiedad.

Para probar esta 2.ª Propiedad, se calcularon las sumas de los cuadrados de la diferencia de todos los valores del set

6 Se le conoce también como suma de cuadrados de las desviaciones del promedio y se usa para el cálculo de la varianza y de la desviación estándar; asimismo, es un coeficiente empleado en análisis más avanzados de estadística, como el análisis de la varianza (ANOVA) y regresión.

utilizando la Ecuación 4.4. Esto es, se sustituye el promedio por cada uno de los valores del set (Tabla 4.4). Las operaciones en Excel serían para completar la Tabla 4.4: =(B2-71)^2; =(B2-72)^2; =(B2-74)^2; =(B2-75)^2; =(B2-76)^2.

El resultado es que: *la suma de las desviaciones de un número del set al cuadrado* es mayor que *la suma de las desviaciones del promedio al cuadrado*. Por lo tanto, se prueba esta 2.ª Propiedad.

Tabla 4.4 Probando la Propiedad 2.ª

	A	B	C	D
1	Números del set	Suma de las desviaciones de un número del set al cuadrado	Comparación	Suma de las desviaciones del promedio al cuadrado
2	70	91	>	28
3	71	56	>	28
4	72	35	>	28
5	73	28	=	28
6	74	35	>	28
7	75	56	>	28
8	76	91	>	28

Usando frecuencias para los promedios

En ocasiones, uno no recibe los datos crudos completos y tiene que trabajar con algunos ya consolidados. Por ejemplo, el promedio del Grupo A es 75.5 (n = 28) y el promedio del Grupo B es 78.2 (n = 20). Pero no sería correcto hacer lo siguiente para obtener un promedio de los promedios: (75.5 + 78.2) / 2 = 76.85, porque las frecuencias (n) de las y los alumnos no son iguales. Por esta razón, hay que tomar las frecuencias en consideración y, con ello, darle el paso correspondiente a cada uno de los promedios. En la Tabla 4.5, se introdujeron los promedios y sus frecuencias para realizar las operaciones en Excel y así obtener el promedio de la combinación de ambos grupos:

Tabla 4.5 Operaciones con diferentes promedios y frecuencias

	A	B	C
1	Promedios	Frecuencia (número de estudiantes)	
2	75.5	28	=A2*B2
3	78.2	20	=A3*B3
4		Σ	=C2+C3
5		Frecuencia	=B2+B3
6		Promedio de los dos grupos	=C4/C5

En la Tabla 4.6 se muestran los resultados de las operaciones y el promedio de ambos grupos: 76.625, que, aunque está cercano al previamente calculado: 76.85, no es igual.

Tabla 4.6 Resultados de operaciones con diferentes promedios y frecuencias

	A	B	C
1	Promedios	Frecuencia (número de estudiantes)	
2	75.5	28	2114
3	78.2	20	1564
4		Σ	3678
5		Frecuencia	48
6		Promedio de los dos grupos	76.625

En cuestión de fórmulas quedaría así: Ecuación 4.5 (véase: Apéndice B con el Operador de suma para más detalles sobre Σ):

$$\frac{\sum_{A=1}^{28} x_A + \sum_{B=1}^{20} x_B}{n_A + n_B} = \bar{x}_{A\,y\,B} = \frac{(\bar{x}_A)\, n_A + (\bar{x}_B)\, n_B}{n_A + n_B} \qquad \text{Ecuación 4.5}$$

Donde:

$\sum_{A=1}^{28} x_A =$ Suma de los valores dentro del Grupo A

$\sum_{B=1}^{20} x_B =$ Suma de los valores dentro del Grupo B

$n_A =$ Frecuencia del Grupo A
$n_B =$ Frecuencia del Grupo B
$\bar{x}_A =$ Promedio del Grupo A
$\bar{x}_B =$ Promedio del Grupo B

Con la Ecuación 4.5, se llegaría también al resultado de 76.625. En el ejemplo anterior solo se usaron dos grupos por simplicidad, pero el número de grupos puede acercarse al infinito. Para calcular el promedio de estos grupos, se puede emplear la Ecuación 4.6:

$$\bar{x}_{grupos} = \frac{\sum_{i=1}^{N} n_i \bar{x}_i}{N} \qquad \text{Ecuación 4.6}$$

Donde:
$\bar{x}_i =$ Promedio de cada uno de los grupos
$n_i =$ Frecuencia de cada uno de los grupos
$N =$ Suma total de las frecuencias de todos los grupos

En palabras: se multiplica el promedio de cada grupo por su frecuencia y da un producto. Luego, se suman todos estos productos y se divide esta suma por la frecuencia total (suma de las frecuencias de cada grupo). Esto se puede llevar a cabo como en la Tabla 4.5.

Gráficas de promedios

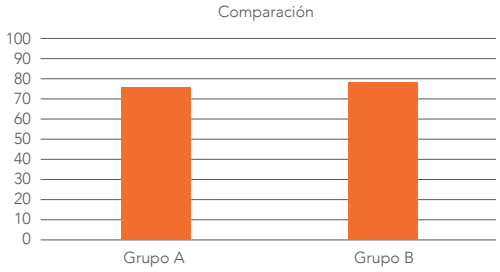
Para utilizar una gráfica de barras en Excel, se pueden emplear las instrucciones de la Figura 3.2. Los datos para construirla vienen de la Tabla 4.7.

Tabla 4.7 Promedios de dos grupos

	A	B
1	Grupos	Promedios
2	A	75.5
3	B	78.2

En la Figura 4.1 se muestra cómo la gráfica de barras evidencia una diferencia entre los promedios. Las diferencias entre los promedios se pueden deber al mero azar (véase: Hinkle *et al.*, 2003, para más información al respecto sobre significancia estadística. Esto se llama poner a prueba una hipótesis: *Null Hipothesis Statistical Significance Testing* o Prueba de Significancia Estadística de la Hipótesis Nula, para estimar si la diferencia fue al azar o si hay algo más como un efecto).

Figura 4.1 Promedios de dos grupos



Ejemplo de un estudio con promedios

Se desea llevar a cabo un estudio para observar el posible efecto de algún tratamiento; *e. g.*, el efecto de recibir un desayuno antes de comenzar un día académico en el aprendizaje de las matemáticas. El aprendizaje de las matemáticas, se podría medir con el número de aciertos en alguna prueba. Este estudio podría tener algún diseño experimental con un grupo que no recibiera el desayuno (grupo control) y un grupo de tratamiento que sí recibiera el desayuno (para más detalles sobre análisis y diseños experimentales, véase: Maxwell, Delaney, & Kelly, 2018; Schneider *et al.*, 2007). De cada uno de estos grupos, se tomaría el promedio de aciertos para ser comparado. Se podría tener una pregunta de investigación como: ¿cuál es el efecto de tomar un desayuno en el aprendizaje de las y los estudiantes? Esta interrogante se puede volver operacional de esta manera: ¿existe una diferencia entre el promedio de aciertos del grupo control y el grupo de trata-

miento? Derivando un par de hipótesis[7] de esta última pregunta, estas quedarían:

- Hipótesis nula (H_0): no existe diferencia entre el promedio del grupo control y el grupo de tratamiento.
- Hipótesis alternativa (H_A): existe diferencia entre el promedio del grupo control y el grupo de tratamiento.

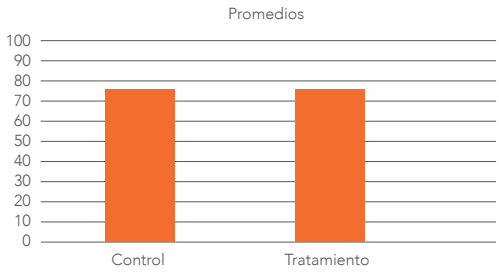
Antes de comenzar con el tratamiento, se medirían los conocimientos de ambos grupos por medio de una prueba para saber si se parte del mismo punto de aprendizaje de las matemáticas. En pocas palabras, lo ideal es que sean grupos equivalentes (*i. e.*, los promedios son iguales o casi iguales). En la Tabla 4.8 se muestra esta deseada equivalencia.

Tabla 4.8 Grupos equivalentes

	A	B
1	Grupos	Promedios
2	Control	75.5
3	De tratamiento	75.5

Asimismo, en la Figura 4.2 se muestran estos grupos equivalentes (véase: Figura 3.2 para la elaboración de figuras en Excel).

Figura 4.2 Grupos equivalentes



7 Una hipótesis es una afirmación que puede ser puesta a prueba para ver si los resultados de algún análisis estadístico la apoyan. La hipótesis que se pone a prueba es la nula que indica: *e. g.*, no hay diferencia entre los promedios de los grupos. La hipótesis nula es rechazada cuando los resultados indican que hay diferencia entre los promedios y no es rechazada cuando los resultados indican que no hay diferencia (véase: Hinkle *et al.*, 2003).

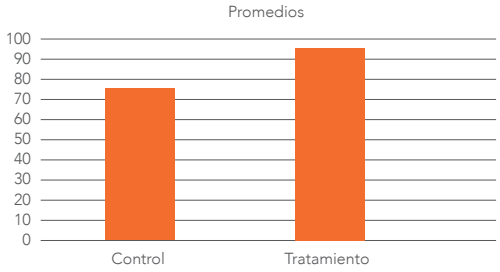
Una vez establecida esta equivalencia, se procedería a aplicar los desayunos al grupo de tratamiento durante cierto tiempo para ver el posible efecto en el aprendizaje de las matemáticas. Después, se les aplicaría la prueba de matemáticas para observar el posible efecto. En la Tabla 4.9 se muestran los resultados de esta prueba. El grupo control se mantuvo en 75.5 de promedio, que sería algo esperado porque no recibió el desayuno. Por otro lado, el promedio en el grupo de tratamiento fue de 75.5 a 95.6, ya que se esperaría algún tipo de mejora en el promedio.

Tabla 4.9 Después del tratamiento

	A	B
1	Grupos	Promedios
2	Control	75.5
3	De tratamiento	95.6

En la Figura 4.3 se muestran los resultados de la Tabla 4.9 elaborados en Excel 2016 (véase: Figura 3.2 para la elaboración de figuras).

Figura 4.3 Después del tratamiento



Conclusiones del ejemplo

En este tipo de casos, la estadística descriptiva (Tabla 4.8 y Tabla 4.9; Figura 4.2 y Figura 4.3) solo sirve para advertir un *posible* efecto, tanto en forma de un set de valores como visualmente, pero no sirve para contestar la pregunta de investigación antes escrita ni tampoco para rechazar o no una hipótesis nula. Para contestar y rechazar la hipótesis nula, se necesitan estadísticas inferenciales (véase: Hinkle *et al.*, 2003).

Sin embargo, esta visualización es complementaria a las estadísticas inferenciales.

c. Mediana

Es el valor que divide un conjunto de datos en dos partes iguales: el 50% de los valores será menor a la mediana y el otro 50% será mayor a esta. Por ello, se encuentra en el percentil 50 (se explican los percentiles más adelante: Capítulo 6). La mediana no es tan susceptible a valores extremos, como el promedio (los valores extremos pueden volver asimétrica una distribución). Esto implica que no se sesga por estos valores extremos. Para un ejemplo práctico de cómo usar la mediana en lugar del promedio cuando se tienen distribuciones asimétricas, se recomienda consultar a Ponce-Renova (2016). La mediana se obtiene al aplicar la siguiente fórmula a un conjunto de datos ordenados e impares con números enteros para encontrar la posición de esta (Ecuación 4.5):

$$\text{Mediana} = (n + 1) / 2 \qquad \text{Ecuación 4.5}$$

Se tiene el set ordenado: 8, 9 y 10; $n = 3$ valores; por lo tanto: = (3 + 1) / 2 = 2 (*i. e.*, segundo lugar). El lugar de la mediana sería el segundo lugar del set ordenado de números enteros e impares: 8, 9 y 10; por lo tanto, la mediana es 9. Para calcular la mediana en Excel, los datos no necesariamente tienen que estar en orden. Solo hay que seleccionar la celda donde se desea el resultado y aplicar la fórmula: =MEDIAN(A2:A3), que aparece en la Tabla 4.10 (celda A4).

Tabla 4.10 Mediana

	A
1	8
2	9
3	10
4	=MEDIAN(A2:A3)

Para cálculos manuales, cuando el set de números ordenados y enteros es par, se utiliza el siguiente procedimiento: se suman los dos números que quedaron en el centro y se dividen entre dos. Por ejemplo:

- Se ordenan: 7 + 8 + 9 + 10
- Se obtienen los dos valores de en medio: 8 y 9
- Se aplica la fórmula: (8 + 9) / 2
- El resultado es: 8.5

En Excel solo hay que aplicar la fórmula, tanto para sets pares como impares: =MEDIAN(Celda$_1$:Celda$_n$).

d. Moda

Es el valor que ocurre más frecuentemente en un conjunto de datos. Cuando se hace manualmente, no hay otra fórmula que simplemente contar el número de veces que aparece un valor. Una vez hecho esto, si un solo valor apareció más que los demás, se tiene una sola moda: una distribución *unimodal*. Caso contrario, si dos valores aparecen más que los demás, pero estos aparecen con la misma frecuencia el uno del otro sería una distribución *bimodal*. Si fueran tres sería una distribución *trimodal* y así sucesivamente. De igual forma, puede pasar que aparezcan múltiples modas: distribución *multimodal*. La moda es considerada la medida de tendencia central más inexacta.

Del ejemplo anterior: 8 + 9 + 9 + 10, la moda es el: 9.

En la Tabla 4.11 se muestra la fórmula (celda A9) para calcular la moda en una distribución unimodal. Asimismo, Excel cuenta con la capacidad de calcular más de una moda en una distribución usando la siguiente fórmula: =MODE.MULT(Celda$_1$:Celda$_n$). Asimismo, para calcular más de una moda se puede emplear el mismo procedimiento de la Tabla 3.15 a la Tabla 3.17, donde se calcularon varias frecuencias. En resumen: suponiendo que la Tabla 4.11 tuviera más de una moda, se selecciona un grupo de celdas (en este caso, se esperan tres mo-

das, así que se seleccionan tres: B2:B4); se escribe la fórmula: =MODE.MULT(B2:B4), sin que se pierda la selección de las celdas; se oprime *Control*, *Shift* y *Enter*, y aparecen las modas y la fórmula con llaves: {=MODE.MULT(B2:B4)}.

Tabla 4.11 Frecuencias de calificaciones para obtener la moda

	A	B
1	Calificación	
2	4	{=MODE.MULT(A2:A8)}
3	5	
4	6	
5	7	
6	7	
7	8	
8	9	
9	=MODE(A2:A8)	

e. Comparaciones entre el promedio, mediana y moda, y su relación con la curtosis y asimetría

En la Tabla 2.1 se muestra cómo algunas escalas (nominales, ordinales, de intervalo y de razón) sirven para calcular algunas estadísticas de tendencia central. Por ejemplo, la moda se puede usar para cualquier escala; la mediana, para las ordinales, de intervalo y de razón; pero el promedio solo se puede utilizar para las escalas de intervalo y de razón.

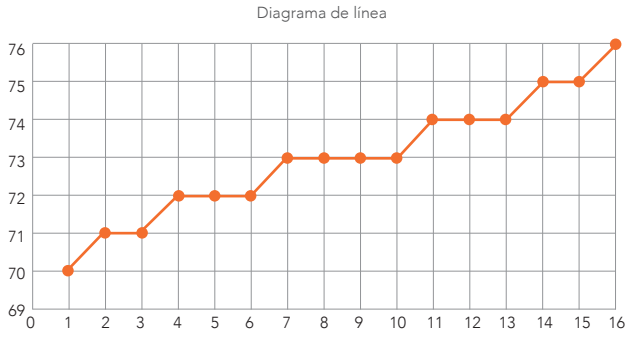
Un interés para la estadística descriptiva es qué significa cuando estas tres medidas de tendencia central coinciden en el mismo valor. Cuando coinciden las tres, se puede tener una distribución normal. Para ilustrar una distribución de la infinidad posible, en la Figura 4.4 se muestra cómo calcular las medidas de tendencia central, así como la forma de distribución de curtosis[8] (*kurtosis*) y asimetría[9] (*skewness*). El resultado fue un promedio = 73, una mediana = 73 y una moda =

8 Es el grado de qué tan puntiaguda o plana es una distribución (Salkind, 2017).
9 Es el grado en el cual la mayoría de los puntajes en una frecuencia de distribución, se localizan en uno de los extremos de una escala de medición con progresivamente menos puntajes hacia el lado opuesto de la escala (Hinkle *et al.*, 2003, p. 739).

73. Por lo tanto, las medidas de tendencia central fueron iguales (promedio = mediana = moda); la curtosis = -0.46 y la asimetría (también conocida como sesgo) = 0. Existen investigaciones que han dicho que mientras la curtosis como la asimetría, se encuentren dentro del rango: -3 a +3 (*i. e.*, | 3 |), no hay ningún problema con la normalidad de la distribución. Otros dicen que el rango es de | 2 | (véase: Apéndice D para las fórmulas y otros escenarios sobre la curtosis y la asimetría).

Figura 4.4 Histograma de una distribución normal

	A
1	70
2	71
3	71
4	72
5	72
6	72
7	73
8	73
9	73
10	73
11	74
12	74
13	74
14	75
15	75
16	76
17	=AVERAGE(A1:A16)
18	=MEDIAN(A1:A16)
19	=MODE(A1:A16)
20	=KURT(A1:A16)
21	=SKEW(A1:A16)
22	



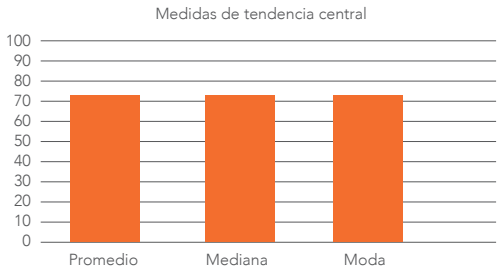
Nota: La curtosis es *kurtosis*: =KURT(Celda$_1$,Celda$_n$); y la asimetría es *skewness*: =SKEW(Celda$_1$,Celda$_n$).

En la Figura 4.4 se muestra cómo esta distribución de datos se vería reflejada en este tipo de gráfica llamada diagrama de línea,

donde la *x* representa la frecuencia de cada uno de los valores del set y la *y*, cada valor del set.

Una gráfica de barras, como la de la Figura 4.5 (véase: Figura 3.2 para la elaboración de figuras de barras), puede ayudar a ilustrar una distribución normal: *i. e.*, cuando las barras se asemejen en su altura, esto puede indicar una distribución normal; otras cosas siendo iguales (*i. e.*, sin considerar otros factores que podrían tener algún tipo de incidencia).

Figura 4.5 Medidas de tendencia central

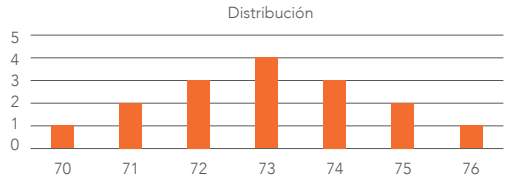


Nota: promedio (73) = mediana (73) = moda (73).

Por último, en la Figura 4.6 se muestra cómo organizar los datos en Excel para representar una posible distribución normal (véase: Figura 3.2 para la elaboración de figuras con barras).

Figura 4.6 Valores y frecuencias

	A	B
1	70	1
2	71	2
3	72	3
4	73	4
5	74	3
6	75	2
7	76	1



Preguntas para resolver del Capítulo 4

- ¿Cómo se podría describir una distribución normal?
- ¿Cuáles variables podrían tener una distribución normal respecto a la Investigación Educativa?
- ¿Por qué el promedio se puede representar con $\bar{x}$ o μ?
- ¿Qué quiere decir la Primera Propiedad del Promedio que expresa que la suma de las desviaciones del promedio genera una suma de cero?
- ¿Cómo se puede probar esta propiedad al usar un conjunto de datos que no venga en este libro?
- ¿Qué quiere decir la Segunda Propiedad del Promedio?
- ¿Cómo se puede probar esta propiedad?
- ¿Por qué es deseable en un diseño experimental el tener grupos equivalentes?
- ¿En qué porcentajes divide la mediana a un conjunto de datos?
- ¿Qué tan susceptible es la mediana en valores extremos?
- ¿Qué escalas se pueden utilizar para calcular la moda de un set?
- ¿Qué significa el término multimodal?
- ¿Qué pueden indicar el promedio, la mediana y la moda cuando coinciden?
- ¿Puede la curtosis afectar la distribución de los datos?
- ¿Puede la asimetría afectar la distribución de los datos?
- ¿Qué medida de tendencia central puede afectar más la asimetría?

Problemas para resolver

Problema 1. Se tienen los puntajes en un examen de admisión de estudiantes del Grupo A (n = 200; Tabla P-4.1). Para describir los datos, se desean obtener las medidas de tendencia central, así como graficar estas con barras. Las preguntas son:

a) ¿Cuál es el promedio, la mediana y la moda del set?
b) ¿Se podría inferir de los datos y la gráfica de barras que se tiene una distribución normal?

Tabla P-4.1 Set de puntajes de admisión

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	33	39	21	39	93	73	51	40	29	64	48	71	63	78	67	7	16	32	14	6
2	24	3	71	57	93	67	67	44	46	62	24	34	31	58	18	73	51	61	10	31
3	9	81	22	55	45	89	36	3	65	13	11	88	91	22	51	22	41	0	90	65
4	44	24	46	56	25	36	24	81	3	34	25	9	9	79	72	17	9	80	43	17
5	15	25	60	92	12	99	45	56	90	80	25	65	25	29	78	34	35	84	7	94
6	84	56	16	18	68	67	13	51	11	41	59	20	27	65	98	55	67	12	53	99
7	50	24	65	4	96	3	78	41	70	99	65	54	88	63	69	43	64	48	16	42
8	25	1	90	39	21	83	4	63	33	45	14	72	10	97	40	31	42	98	26	66
9	26	64	2	45	88	72	84	7	38	60	63	8	12	45	4	6	1	97	54	60
10	68	2	31	56	15	79	18	90	33	61	43	15	25	47	16	99	23	81	73	6

Problema 2. De igual manera, se tiene otro set de puntajes del examen de admisión de estudiantes del Grupo B (n = 100; Tabla P-4.2).

a) ¿Cuál es el promedio, la mediana y la moda del set?
b) ¿Se podría inferir de los datos y la gráfica de barras que se tiene una distribución normal?

Tabla P-4.2 Conjunto de datos del Grupo 1

	A	B	C	D	E	F	G	H	I	J
1	33	39	21	39	93	73	51	40	29	64
2	24	3	71	57	93	67	67	44	46	62
3	9	81	22	55	45	89	36	3	65	13
4	44	24	46	56	25	36	24	81	3	34
5	15	25	60	92	12	99	45	56	90	80
6	84	56	16	18	68	67	13	51	11	41
7	50	24	65	4	96	3	78	41	70	99
8	25	1	90	39	21	83	4	63	33	45
9	26	64	2	45	88	72	84	7	38	60
10	68	2	31	56	15	79	18	90	33	61

Problema 3. Del Grupo C de estudiantes, se desea obtener el promedio, la mediana y la moda (n = 100; Tabla P-4.3).

a) ¿Cuál es el promedio, la mediana y la moda del set?
b) ¿Se podría inferir de los datos y la gráfica de barras que se tiene una distribución normal?

Tabla P-4.3 Conjunto de datos del Grupo 2

	A	B	C	D	E	F	G	H	I	J
1	48	71	63	78	67	7	16	32	14	6
2	24	34	31	58	18	73	51	61	10	31
3	11	88	91	22	51	22	41	0	90	65
4	25	9	9	79	72	17	9	80	43	17
5	25	65	25	29	78	34	35	84	7	94
6	59	20	27	65	98	55	67	12	53	99
7	65	54	88	63	69	43	64	48	16	42
8	14	72	10	97	40	31	42	98	26	66
9	63	8	12	45	4	6	1	97	54	60
10	43	15	25	47	16	99	23	81	73	6

Problema 4. Habría que comparar los promedios de los Grupos A, B y C en forma gráfica y con los valores para advertir diferencias.

a) ¿Existe alguna diferencia entre los promedios de los tres grupos?
b) Si hay diferencias, ¿a qué podrían deberse?

Problema 5. Si se tienen dos promedios: $\bar{x}_1 = 95$ ($n = 15$) y $\bar{x}_2 = 80$ ($n = 30$), ¿cuál sería el promedio de acuerdo con el peso del tamaño de la muestra?

Preguntas para reflexionar

- Aparte de la distribución normal, ¿qué otras distribuciones se dan en la Investigación Educativa?
- ¿Cómo se usan esas distribuciones para describir o analizar los datos?
- ¿Cómo se puede definir una población?
- ¿Cómo evidenciar que una muestra corresponde a cierta población?
- ¿Por qué las medidas de tendencia central son importantes para la estadística descriptiva?
- ¿Qué se puede obtener de un diseño experimental?
- ¿Qué amenazas existen en los diseños experimentales?
- ¿Qué implicaciones tiene el no contar con una distribución normal?
- ¿Qué hacer para remediar problemas con la curtosis y la asimetría?

Opinión del Autor

Como dice un dicho: "Todos tenemos derecho a una opinión, pero no a fabricar hechos". En otras palabras, no podemos cambiar los datos *per se*, aunque se puedan hacer inferencias que pudieran ser contradictorias entre sí. Por ejemplo, el Examen Internacional de Evaluación de las y los Estudiantes (*i. e.*, conocido como Examen Pisa) tiene como objetivo evaluar los conocimientos y habilidades de las y los alumnos para participar en la sociedad del saber (Organización para la Cooperación y el Desarrollo Económicos [OCDE], 2021). En el año 2000, en el examen de matemáticas, el promedio del puntaje fue de 492 puntos con un error estándar del promedio[10] de 0.7 (n = 41 países), donde México obtuvo un promedio de 387 puntos (error estándar [*ES*] = 3.4);

10 El error estándar del promedio es una estadística que se obtiene al dividir la desviación estándar por la raíz cuadrada del tamaño de una muestra (n). Entre más grande sea el error estándar del promedio, más variación existe. El error estándar del promedio también es la desviación estándar de la distribución de muestras (*sampling distribution*; véase: Ponce-Renova, 2019). La distribución de muestras se obtiene cuando se toma un gran número de muestras de una población y se genera una distribución que se aproxima a una distribución normal (véase: Ponce-Renova, 2019, con el Teorema de Tendencia Central). Asimismo, el error estándar del promedio se utiliza para obtener márgenes de error (véase el Capítulo 1 del presente libro) y para obtener intervalos de confianza (véase: Cumming, 2013).

Estados Unidos, un promedio de 493 (*ES* = 7.6); y Canadá, un promedio de 533 (*ES* = 1.4); la fuente fue *Pisa Data Explorer* en la página de la OCDE (https://www.oecd.org/pisa/data/). Estos promedios son los hechos, aunque cuentan con cierto error (*i. e.*, *ES*). Una opinión sería que México está muy mal respecto a los países de América del Norte. Otra opinión podría ser que México *no* tiene los recursos para la educación como sí los tienen estos dos países. En fin, se podrían tener muchas opiniones al respecto, pero estas no deben de cambiar los datos.

Mi recomendación es que se enfoquen en obtener y graficar las medidas de tendencia central con la idea de que, posiblemente, estas van a representar a poblaciones. También, hay que estar atentas y atentos a las distribuciones de los datos en cuestión de curtosis y asimetría, porque estas estadísticas pueden afectar posibles análisis inferenciales, si es que se hacen. Asimismo, cuando se quieren combinar promedios de diferentes muestras, hay que tomar en cuenta el tamaño de estas para darles el peso indicado.

CAPÍTULO 5

Medidas de dispersión

Solemos pensar que el promedio de algo es el mejor descriptor de un fenómeno. Pero esto no siempre resulta así. Suponiendo que una docente tiene dos ofertas de trabajo en dos ciudades diferentes que no conoce. Ella considera que uno de los criterios para tomar una decisión de cuál oferta aceptar, es la temperatura por cuestiones de salud y para hacer ejercicio en el exterior. La Ciudad A tiene un promedio anual de 20 °C y la Ciudad B también tiene 20 °C. Por lo que llega a la conclusión de que la temperatura de ambas ciudades en cuestión no marcaría una diferencia para su elección. Algo que su proceso de decisión no está considerando —y es un factor crucial—, es la variación de la temperatura. En este caso, la variación de la temperatura durante todo el año. En detalle, y en forma hipotética, la Ciudad A tiene un promedio de 20 °C, pero medio año está a 40° y la otra mitad está a 0°, lo que da como promedio: 20°. En síntesis, tiene dos temperaturas que son extremas. Mientras que la Ciudad B tiene 20 °C todo el año, lo que hace que no tenga variación. Este es un caso hipotético para considerar el gran peso que tiene la variación para poder interpretar un promedio. De considerar la variación, la docente hubiera optado por la Ciudad B.

a. Generalidades

A diferencia de las medidas de tendencia central que solo implican una estimación de un punto (un solo coeficiente como promedio = 7.8), las medidas de variabilidad son espacios de intervalos que indican cómo los puntajes se es-

parcen/distribuyen en una distribución[1] (Hinkle *et al.*, 2003). Posiblemente, las medidas más recurrentes de variabilidad son: el rango, la desviación del promedio, la varianza y la desviación estándar, así como el *error estándar del promedio*[2] (la última está más allá de la cobertura del presente libro).

b. Rango

Es definido como el número de unidades en una escala de medición, que incluye al más alto y al más bajo de los valores (Hinkle *et al.*, 2003, p. 61). Hay dos maneras de calcular el rango, dependiendo de si se desea obtener un rango *inclusivo* o *exclusivo* (si se tiene un conjunto de datos: 5, 6, 8 y 10):

- Inclusivo: es el valor más alto del set (Límite Alto) menos el más pequeño (Límite Bajo) más 1. Esto es: (Límite Alto – Límite Bajo) + 1. Del ejemplo anterior, se tiene (10 – 5) + 1 = 6. Este rango se interpreta diciendo que este set va del valor 5 al 10.
- Exclusivo: es el valor más alto del set (Límite Alto) menos el más pequeño (Límite Bajo). Esto es: (Límite Alto – Límite Bajo). Del ejemplo anterior, se tiene: (10 – 5) = 5, que es un rango de 5.

En la Tabla 5.1 con funciones de Excel, se muestra cómo calcular un rango inclusivo y uno exclusivo con un conjunto de datos. Asimismo, en la Tabla 5.2 se muestran los resultados. Para ambos rangos, se obtiene el valor máximo: = 76, =MAX(B1:B16); y el valor mínimo: = 70, =MIN(B1:B16). Para el rango inclusivo, se obtiene la diferencia entre el valor máximo y el valor mínimo, y se agrega 1:

1 Además de la distribución normal, existen otras distribuciones de datos comúnmente encontradas: Bernoulli; uniforme; binomial; Poisson; y exponencial (véase: Analytics Vidhya, 2017).
2 Conocido en inglés como *the standard error of the mean* (SEM) es una estadística que indica qué tanto el promedio de una muestra en particular, es probable que difiera del promedio de la población de donde fue obtenido (VandenBos, 2015, p. 1025). Esta estadística es fundamental para calcular intervalos de confianza y márgenes de error.

i. e., 76 – 70 + 1 = 7, =B17-B18+1. Para el rango exclusivo, se deja de agregar 1 y resulta: 76 – 70 = 6, =B17-B18.

Tabla 5.1 Fórmulas de rangos

	A	B
1		70
2		71
3		71
4		72
5		72
6		72
7		73
8		73
9		73
10		73
11		74
12		74
13		74
14		75
15		75
16		76
17	Máximo	=MAX(B1:B16)
18	Mínimo	=MIN(B1:B16)
19	Rango inclusivo	=B17-B18+1
20	Rango exclusivo	=B17-B18

Tabla 5.2 Resultados de los rangos

	A	B
1		70
2		71
3		71
4		72
5		72
6		72
7		73
8		73
9		73
10		73
11		74
12		74
13		74
14		75
15		75
16		76
17	Máximo	76
18	Mínimo	70
19	Rango inclusivo	7
20	Rango exclusivo	6

El rango es fácil de calcular y de interpretar, pero puede ser distorsionado fácilmente con observaciones atípicas (véase: Apéndice E). No se puede comparar el rango de un conjunto de datos con el de otro set diferente, sin tener que estandarizar los valores de ambos sets (véase: Apéndice E).

c. Media de desviación del promedio

Es otra medida de dispersión que se observa parcialmente en las Tablas 4.2 y 4.3 del capítulo anterior con la fórmula: $x_{i \text{ de la diferencia}} = (x_i - \bar{x})$. Sin embargo, a esta ecuación le faltan otros elementos para completarla. Dada la 1.ª Propiedad (La suma de las desviaciones del promedio dan cero), se tienen que adoptar valores absolutos (véase: Apéndice B): *i. e.*, | valor absoluto |.

$$\text{MDP} = \frac{\sum_{i=1}^{n} | x_i - \bar{x} |}{n} = \frac{\sum_{i=1}^{n} | x_{i \text{ de la diferencia}} |}{n} \qquad \text{Ecuación 5.1}$$

Donde:

MDP = Media de desviación del promedio

$\sum_{i=1}^{n}$ = Suma de las diferencias: desde la primera del set ($i = 1$) hasta la última (n)

x_i = Cada uno de los valores del set

$\bar{x}$ = Promedio del set

n = Tamaño de la muestra/Frecuencia total

$x_{i \text{ de la diferencia}}$ = Cada una de las diferencias entre los valores del set y su promedio

Se puede definir la media de desviación del promedio (MDP) como la suma de las diferencias absolutas entre cada uno de los valores del set y el promedio dividida por la frecuencia total (n): véase: Ecuación 5.1. La MDP se puede usar para comparar las variaciones de diferentes sets de datos cuando estén medidos en la misma escala. Los sets de datos con escalas más amplias tendrán promedios y valores más altos, y harán difícil la comparación con otros sets. Para ello, se emplea la desviación estándar, pero la MDP se considera un antecedente tanto de la desviación estándar como de la varianza. En la Tabla 5.3 se muestran

las fórmulas en Excel para calcular la MDP y en la Tabla 5.4, se muestran los resultados de esta primera.

Tabla 5.3 Fórmulas de la MDP

	A	B	C
1		x_i	$\lvert x_i - \bar{x} \rvert$
2		70	=ABS(B2-B$18)
3		71	=ABS(B3-B$18)
4		71	=ABS(B4-B$18)
5		72	=ABS(B5-B$18)
6		72	=ABS(B6-B$18)
7		72	=ABS(B7-B$18)
8		73	=ABS(B8-B$18)
9		73	=ABS(B9-B$18)
10		73	=ABS(B10-B$18)
11		73	=ABS(B11-B$18)
12		74	=ABS(B12-B$18)
13		74	=ABS(B13-B$18)
14		74	=ABS(B14-B$18)
15		75	=ABS(B15-B$18)
16		75	=ABS(B16-B$18)
17		76	=ABS(B17-B$18)
18	$\bar{x}$	=AVERAGE(B2:B17)	
19		MDP	=AVERAGE(C2:C17)
20	s^2	=VAR.S(B2:B17)	
21	*SD*	=STDEV.S(B2:B17)	

Tabla 5.4 Resultados de la MDP

	A	B	C
1		x_i	$\lvert x_i - \bar{x} \rvert$
2		70	3
3		71	2
4		71	2
5		72	1
6		72	1
7		72	1
8		73	0
9		73	0
10		73	0
11		73	0
12		74	1
13		74	1
14		74	1
15		75	2
16		75	2
17		76	1
18	$\bar{x}$	73	
19		MDP	1.25
20	s^2	2.67	
21	*SD*	1.63	

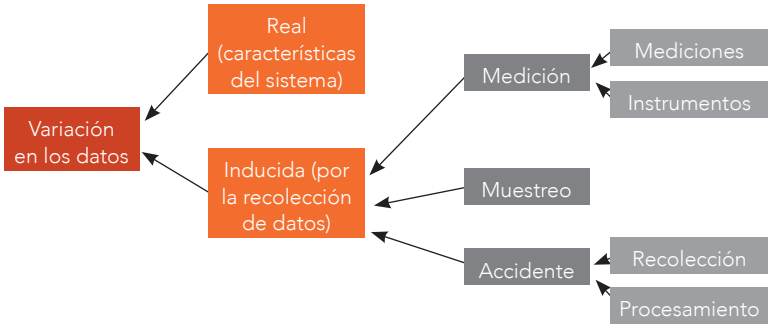
Nota: $\bar{x}$ = Promedio; s^2 = Varianza; *SD* = Desviación estándar.

d. Fuentes de variación

Si algo se va a encontrar en el proceso de enseñanza-aprendizaje (entre otros fenómenos en la Investigación Educativa) es la *variación*, que es omnipresente. Un ejemplo de esta variación es el aprendizaje de las y los estudiantes, que se da en un *continuo* con una distribución normal, aproximadamente (véase: Crocker y Algina, 2008, para aspectos

de la distribución y psicometría;[3] véase: curva normal estándar[4] en el Apéndice D [Figura D]). En una distribución normal, algunas y algunos estudiantes estarán en el inicio de aprender algo; otras y otros, en medio; y algunas y algunos más al final de este. Según Wild y Pfannkuch (1999), la variación en un conjunto de datos se puede deber a dos tipos de fuentes: las *reales* y las *inducidas* (véase: Figura 5.1). Usando de nuevo el aprendizaje como ejemplo, una variación *real* sería que una población de estudiantes expuestas y expuestos a una lección escolar aprendiera a diferentes niveles: *e. g.*, niveles bajos, medianos y altos. Esta sería una variación *entre* los puntajes de las y los alumnos (variación entre los puntajes de algunas y algunos estudiantes y los de otras y otros). La otra fuente de variación es la *inducida*. Esto quiere decir que aparece por las diferentes maneras de obtener/recolectar la información: *e. g.*, los niveles de aprendizaje.

Figura 5.1 Fuentes de variación en los datos



Fuente: Wild & Pfannkuch (1999, p. 235).

Un ejemplo de variación inducida sería una población de estudiantes que tomara un curso MOOC (*Massive Open Online Course*). Sus aprendizajes serían medidos (Mediciones; Figura 5.1) con una serie de pruebas a través del curso y habría variación *dentro* de los resultados de

3 *Grosso modo*, la psicometría es una rama de la psicología que concierne a la cuantificación y medición de atributos mentales, comportamientos, rendimiento en alguna área como el aprendizaje y otras similares, así como el diseño, análisis y mejoramiento de pruebas, cuestionarios y otros instrumentos usados en la medición (VandenBos, 2015, p. 860).
4 Una curva normal estándar (también llamada distribución normal estandarizada) se emplea para representar la distribución normal (véase: Ponce-Renova, 2019, para más detalles).

una misma o mismo estudiante (*e. g.*, primer parcial = 7.1; segundo parcial = 7.3; etcétera). Esta varianza se llama variación dentro de un grupo (cada variación dentro[5] de los puntajes de un mismo estudiante).

Además, otro aspecto de la medición son los instrumentos (Instrumentos; Figura 5.1), ya que, aunque se trata de medir lo mismo, el aprendizaje, cuando se usan diferentes pruebas como opción múltiple, respuesta corta o respuesta larga, los puntajes de una alumna o un alumno no siempre son los mismos, a pesar de estar midiendo los mismos niveles de aprendizaje en este constructo. Por ejemplo, una o un estudiante podría sacar un 90% de aciertos en una prueba de respuesta corta y un 50% en una prueba de respuesta larga cuando ambas pruebas están midiendo el mismo constructo. Esto se podría deber a que la alumna o el alumno en cuestión no sabe expresar apropiadamente en forma escrita sus conocimientos. Por lo tanto, esta es una variación inducida.

Otra fuente de variación inducida es el muestreo (Figura 5.1), que sucede porque se toma una parte de la población, que es una muestra, y de las estadísticas, que se obtienen de las muestras (porcentajes, promedio, desviación estándar, entre muchos otros más), se trata de inferir cuáles son los parámetros de la población. Esta diferencia o variación entre las estadísticas de la muestra y los parámetros de la población sería inducida por el muestreo.

Otra variación inducida es la que pasa por accidente (Figura 5.1), que ocurre cuando en la recolección de datos se comete un error en la captura, como podría suceder cuando se tiene un censo de población. Por ejemplo, se le podría preguntar la edad a una persona y esta contestar 36 años, pero al llenar la hoja de captura de datos se podría escribir un 39. Asimismo, se podría llevar esta hoja de captura de datos para su procesamiento, pero al colocar la edad de la persona u otro dato, se coloca otro diferente al que se había escrito. En pocas palabras, habría una variación inducida en forma accidental.

5 Tanto para la varianza dentro de los puntajes como para la variación entre los puntajes, se recomienda consultar a Maxwell, Delaney y Kelly (2018), quienes llevaron a cabo una serie de análisis de la varianza en el contexto de diseños experimentales.

e. Varianza

"Varianza es definida como el promedio de la suma de las desviaciones del promedio al cuadrado" (Hinkle *et al.*, 2003, p. 66). Como es una medición al cuadrado, no resulta fácil apreciar su dimensión cuando se compara con el conjunto de datos de donde se obtuvo. Para usar la misma métrica habría que obtener la *desviación estándar*, que es el siguiente tema. Asimismo, la varianza puede ser considerada como un fenómeno natural. Como Carlos Darwin la veía: hay una variación entre los miembros de alguna especie. Por ejemplo, al salir de un huevo un grupo de pichones con los mismos padres variarán en color, peso, longitud de sus extremidades, etcétera. En la educación, se puede apreciar esta variación en varios aspectos: calificaciones de las y los estudiantes, asistencia a la escuela, motivación, variables socioeconómicas e inteligencia de las y los alumnos, entre muchos otros más. Entre algunas de las propiedades de la varianza están las siguientes:

- La *varianza* se usa para medir la dispersión de los datos alrededor de alguna estadística, como el promedio de un conjunto de datos, entre otros. Es una medida al cuadrado (*i. e.*, se representa geométricamente con círculos y cuadrados) y no longitudinal, como la desviación estándar.
- La *varianza* nunca es negativa, porque está al cuadrado.
- La *varianza* es sensible a los valores atípicos (véase: Apéndice E para una explicación sobre estos valores). Esto es, un solo valor atípico puede incrementar la *varianza* y, por ello, distorsionar la dispersión de los datos.
- Para datos con el mismo promedio, aproximadamente, entre más grande sea la dispersión, mayor será la *varianza*.
- Si todos los datos del set son iguales, la *varianza* es igual a cero.

Las fórmulas de la varianza indican una suma de las diferencias entre cada uno de los valores de un set y el promedio de este último dividido por el número de valores del set. En esta sección, se ven fórmulas de la varianza para las poblaciones y para las muestras, porque

existen diferencias al estimar los parámetros y las estadísticas, respectivamente. La Ecuación 5.2 muestra *tres* diferentes maneras de expresar la misma fórmula para calcular la varianza de una población y para estimar la varianza de la población en Excel: =VAR.P(Celda$_1$:Celda$_n$):

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} = \sum_{i=1}^{N} \frac{(x_{i \text{ de la diferencia}})^2}{N} = \frac{SC}{N} \qquad \text{Ecuación 5.2}$$

Donde:

σ^2 = Varianza de la población

N

$\sum_{i=1}$ = Suma de las diferencias desde la primera del set ($i = 1$) hasta la última (N)

N = Tamaño de la población/Frecuencia total

x_i = Cada uno de los valores del set

μ = Promedio de la población

$x_{i \text{ de la diferencia}}$ = Cada una de las diferencias entre los valores del set y su promedio

SC = Suma de cuadrados

La suma de cuadrados (*SC*) es un concepto fundamental para el análisis de la varianza (ANOVA[6]), que sirve para comparar promedios de dos o más grupos.

Un ejemplo de varianza de una población es la Tabla 5.5, donde se muestra la función para obtenerla automáticamente: =VAR.P(Celda$_1$:Celda$_n$). Además, se muestra una serie de pasos para obtener la varianza de la población al calcular el promedio del conjunto de datos (Celda B8); obtener la diferencia entre cada valor y el promedio (Celda C2 a Celda C6); elevar al cuadrado la diferencia (Celda D2 a Celda D6); sumar las diferencias (Celda D7); y dividir la suma de las diferencias entre la frecuencia (*i. e.*, $n = 5$): Celda D9. En la Tabla 5.6 se muestra la ejecución de estos procedimientos como los resultados.

6 Debido a la importancia del ANOVA, se recomienda consultar a estos tres autores para lectores avanzados en la comparación de promedios de grupos: Maxwell *et al.* (2018); Paolella (2019); y Rutherford (2011).

Tabla 5.5
Varianza de una población con funciones

	A	B	C	D
1		x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2		120	=B2-B$8	=C2^2
3		130	=B3-B$8	=C3^2
4		110	=B4-B$8	=C4^2
5		160	=B5-B$8	=C5^2
6		150	=B6-B$8	=C6^2
7	Varianza*	=VAR.P(B2:B6)	Suma	=SUM(D2:D6)
8	Promedio	=AVERAGE(B2:B6)	*n*	5
9			Varianza**	=D7/D8

Tabla 5.6
Resultados

	A	B	C	D
1		x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2		120	-14	196
3		130	-4	16
4		110	-24	576
5		160	26	676
6		150	16	256
7	Varianza	344	Suma	1720
8	Promedio	134		5
9			Varianza	344

Nota: La varianza de la población es 344. La varianza* es obtenida automáticamente con la función de Excel. La varianza** es obtenida en forma manual en Excel.

La Ecuación 5.3 contiene cómo calcular la varianza de una muestra y en la Tabla 5.7, se muestra el cálculo en Excel para la varianza de una muestra: =VAR.S(B2:B6). Las fórmulas de la varianza de una población y de una muestra difieren, porque esta última implica restarle 1 al tamaño de la muestra (*n*) y, entonces, la varianza se hace más grande (véase: Ecuación 5.3) que la varianza de su respectiva población (comparar con la Ecuación 5.2).

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \sum_{i=1}^{n} \frac{(x_i \text{ de la diferencia})^2}{n-1} = \frac{SC}{n-1} \qquad \text{Ecuación 5.3}$$

Donde:

s^2 = Varianza de la muestra

$\sum_{i=1}^{n}$ = Suma de las diferencias desde la primera del set ($i = 1$) hasta la última (n)

n = Tamaño de la muestra/Frecuencia total

x_i = Cada uno de los valores del set

$\bar{x}$ = Promedio de la muestra

$x_{i \text{ de la diferencia}}$ = Cada una de las diferencias entre los valores del set y su promedio

SC = Suma de cuadrados

Al hacer la varianza de la muestra más grande, se considera el error de medición que existe al usar una muestra en lugar de la población. El error de medición es la diferencia que hay entre el promedio de la muestra y el promedio de la población ($\mu - \bar{x}$).

Un ejemplo de varianza de una muestra es la Tabla 5.7 con funciones y sus resultados en la Tabla 5.8. Se realizaron los mismos procedimientos que en las Tablas 5.5 y 5.6; lo único que fue diferente es que en este ejemplo, se usó $n - 1$ por ser para una muestra. Aunque no aparente que el $n - 1$ de la Ecuación 5.3 tenga un gran efecto en la varianza de una muestra en comparación con la varianza de una población, sí la puede tener. Esto es, los sets de datos de la población y la muestra son los mismos, pero en la primera de estas tablas se está calculando la varianza para la población (344) y en la segunda, para la muestra (430).

Tabla 5.7

Varianza de una muestra con funciones

	A	B	C	D
1		x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2		120	=B2-B$8	=C2^2
3		130	=B3-B$8	=C3^2
4		110	=B4-B$8	=C4^2
5		160	=B5-B$8	=C5^2
6		150	=B6-B$8	=C6^2
7	Varianza*	=VAR.S(B2:B6)	Suma	=SUM(D2:D6)
8	Promedio	=AVERAGE(B2:B6)	$n - 1$	=5-1
9			Varianza**	=D7/D8

Tabla 5.8

Resultados

	A	B	C	D
1		x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2		120	-14	196
3		130	-4	16
4		110	-24	576
5		160	26	676
6		150	16	256
7	Varianza	430	Suma	1720
8	Promedio	134	$n - 1$	4
9			Varianza	430

Nota: La varianza de la muestra es 430. La varianza* es obtenida automáticamente con la función de Excel. La varianza** es obtenida en forma manual en Excel.

f. Desviación estándar

La desviación estándar (*SD*) representa la cantidad promedio de variabilidad en un set de valores. La *SD* se obtiene al calcular la raíz

cuadrada de la varianza. La diferencia estriba en que la varianza está al cuadrado y la desviación estándar ya *no* está al cuadrado, así como que la *SD* está en las mismas unidades que el conjunto de datos de donde se obtuvo. Las propiedades de la *SD* son:

- La *SD* se usa para medir la dispersión de los datos alrededor de un promedio de un conjunto de datos. Es una medida longitudinal (*i. e.*, se representa con rectas como en un plano cartesiano).
- La *SD* nunca es negativa.
- La *SD* es sensible a los valores atípicos (véase: Apéndice E para una explicación sobre estos valores). Esto es, un solo valor atípico puede incrementar la *SD* y, por ello, distorsionar la dispersión de los datos.
- Para datos con el mismo promedio, aproximadamente, entre más grande sea la dispersión, mayor será la *SD*.
- Si todos los datos del set son iguales, la *SD* es igual a cero.

De igual manera que la varianza, la desviación estándar tiene una fórmula para una población (Ecuación 5.4) y otra para una muestra (Ecuación 5.5). Ambas ecuaciones (5.4 y 5.5) muestran *tres* diferentes maneras de expresar la misma fórmula para calcular la desviación estándar:

- I. Desviación estándar de una población:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}} = \sqrt{\frac{SC}{N}} \qquad \text{Ecuación 5.4}$$

Donde:

σ = Desviación estándar de la población

σ^2 = Varianza de la población

N

Σ = Suma de diferencias desde la primera del set (i = 1) hasta la última (N)

$i = 1$

N = Tamaño de la población/Frecuencia total

x_i = Cada uno de los valores del set
μ = Promedio de la población
$x_{i \text{ de la diferencia}}$ = Cada diferencia entre los valores del set y su promedio
SC = Suma de cuadrados

En la Tabla 5.9 se muestra cómo obtener la desviación estándar de una población en Excel: =STDEV.P(A1:A5). Asimismo, la *SD* de la población y de la muestra se pueden obtener manualmente como se hizo con las varianzas de la población y de la muestra de las Tablas 5.5 a la 5.8, pero hay que sacar la raíz cuadrada de la varianza que se obtenga al simplemente usar la función de Excel: =SQRT(Celda *x*).

Tabla 5.9 Desviación estándar de una población

	A
1	x_i
2	120
3	130
4	110
5	160
6	150
7	=STDEV.P(A1:A5)

Nota: La desviación estándar de esta población es 18.55. La fórmula de Excel para una población: =STDEV.P(Celda$_1$:Celda$_n$).

» II. Desviación estándar de una muestra:

$$SD = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}} = \sqrt{\frac{SC}{(n-1)}}$$ Ecuación 5.5

Donde:
SD = Desviación estándar de la muestra
s^2 = Varianza de la muestra

n
Σ = Suma de diferencias desde la primera del set (i = 1) hasta la última (n)
$i = 1$

n = Tamaño de la población/Frecuencia total

x_i = Cada uno de los valores del set

$\bar{x}$ = Promedio de la muestra

$x_{i \text{ de la diferencia}}$ = Cada diferencia entre los valores del set y su promedio

SC = Suma de cuadrados

En la Tabla 5.10 se muestra cómo obtener la desviación estándar de una muestra con una sola función de Excel: =STDEV.S(A1:A5). Manualmente, una *SD* de una muestra se puede obtener al sacar la raíz cuadrada de una varianza de una muestra: =SQRT(Celda *x*).

Tabla 5.10 Desviación estándar de una muestra

	A
1	x_i
2	120
3	130
4	110
5	160
6	150
7	=STDEV.S(A1:A5)

Nota: La desviación estándar de esta muestra es 20.74. Para complementar las ecuaciones de la desviación estándar, la fórmula de Excel para una población es: =STDEV.P(Celda$_1$:Celda$_n$).

Al igual que las ecuaciones para calcular la varianza, las ecuaciones para calcular la desviación estándar (5.4 y 5.5) difieren entre sí por esta parte de la fórmula: $n - 1$. Aunque sea el mismo conjunto de datos, la desviación estándar de la población es 18.55 y la de la muestra es 20.74; lo que expone que la desviación estándar de la muestra es más grande. Algunos investigadores han dicho que tanto la varianza como la desviación estándar de las muestras son más grandes que las de sus correspondientes poblaciones, porque se incluye un error. Además, una *desviación estándar* grande indica una gran variación y una desviación estándar pequeña indica una variación pequeña. La desviación estándar es una estadística que siempre debe aparecer, al igual que el promedio, en una Investigación Educativa.

Preguntas para resolver del Capítulo 5

- ¿Cuáles son las medidas de dispersión de este capítulo?
- ¿Cuál es la diferencia entre el rango inclusivo y el exclusivo?
- ¿Cómo se obtiene el valor máximo y mínimo de un conjunto de datos?
- ¿Cuáles son las dos fuentes de variación de este capítulo?
- ¿Qué es la varianza?
- ¿Cuáles son las propiedades de la varianza?
- ¿En qué se diferencia la fórmula de la varianza de una población de la de una muestra?
- ¿Cuál de estas dos varianzas suele ser más grande?
- ¿Qué es la desviación estándar?
- ¿Cuáles son las propiedades de la desviación estándar?
- ¿En qué se diferencia la fórmula de la desviación estándar de una población de la de una muestra?
- ¿Cuál de estas dos desviaciones estándar suele ser más grande?
- ¿Por qué es importante reportar la desviación estándar?

Problemas para resolver

Problema 1. Del conjunto de datos de la Tabla P-5.1 hay que contestar las siguientes preguntas:

a) ¿Cuál es el valor máximo y mínimo?; y **b)** ¿Cuál es el rango inclusivo y exclusivo?

Tabla P-5.1 Rangos, mínimo y máximo

	A	B
1		x_i
2		10
3		6
4		7
5		5
6		8
7		5
8		8
9		7

Continúa...

10		5
11		6
12		7
13		9
14		9
15		7
16		4
17		6
18	Máximo	
19	Mínimo	
20	Rango inclusivo	
21	Rango exclusivo	

Problema 2. De la Tabla P-5.2 y usando el procedimiento de la Tabla 5.3 (*i. e.*, obteniendo el promedio y el valor absoluto entre la diferencia de cada valor y el promedio ($| x_i - \bar{x} |$), hay que contestar la siguiente pregunta:

a) ¿Cuál es la MDP?

Tabla P-5.2 MDP

	A	B
1		x_i
2		1
3		5
4		9
5		6
6		7
7		8
8		10
9		8
10		9
11		5
12		4
13		3
14		2
15		7

Continúa...

16		2
17		1
18	$\bar{x}$	
19	MDP	
20	s^2	
21	*SD*	
22	Máximo	
23	Mínimo	
24	Rango inclusivo	
25	Rango exclusivo	

Asimismo, las otras preguntas serían:

b) ¿Cuál es la varianza?; **c)** ¿Cuál es la desviación estándar?; **d)** ¿Cuál es el valor máximo y mínimo?; **e)** ¿Cuál el rango inclusivo y exclusivo?; y **f)** ¿Cuál es el promedio?

Problema 3. De la Tabla P-5.3 resuelva las siguientes preguntas, tanto con funciones de Excel como manualmente:

a) ¿Cuáles son las medidas de tendencia central?; **b)** ¿Cuál es la varianza y desviación estándar cuando se asume que provienen de una muestra?; y **c)** ¿Cuál es el rango inclusivo y exclusivo del conjunto de datos?

Tabla P-5.3 Varias estadísticas

	A	B	C	D
1	Calificaciones	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	5			
3	5			
4	5			
5	6			
6	6			
7	6			
8	6			
9	6			

Continúa...

	A	B	C	D
10	7			
11	7			
12	7			
13	7			
14	7			
15	7			
16	7			
17	7			
18	8			
19	8			
20	8			
21	8			
22	8			
23	8			
24	8			
25	9			
26	9			
27	9			
28	9			
29	9			
30	10			
31	10			
32	10			
33	10			

Preguntas para reflexionar

- ¿Qué podría implicar para una estudiante o un estudiante si alcanza el máximo en alguna prueba?
- ¿Qué podría implicar para una alumna o un alumno el obtener cero aciertos en una prueba?
- Aparte de la normal, ¿qué otras distribuciones son comunes en la investigación y cómo se han usado?
- ¿Existen más de dos fuentes de variación?
- ¿Qué ejemplos existen en la naturaleza de variación?

- ¿Cómo se puede emplear la desviación estándar en una distribución normal?
- ¿Qué relación tiene una desviación estándar con un intervalo de confianza?

Opinión del Autor

Las medidas de dispersión no son reportadas en muchas de las Investigaciones Educativas que me han tocado leer y evaluar. El no incluirlas es un despropósito, porque son fundamentales para entender un fenómeno como el aprendizaje, entre otros muchos más. Una estadística fundamental, tanto en lo descriptivo como en lo inferencial, es la varianza, porque nos indica qué tanto varía algún fenómeno. Por otro lado, es difícil entender la varianza, porque es una estadística que está al cuadrado. En contraste, cuando a la varianza se le saca raíz cuadrada y se obtiene la desviación estándar, que es una medida longitudinal y en la misma escala que el conjunto de datos original, es más simple de entender. Hablando de estadísticas que reportar, la desviación estándar es una de ellas y, de hecho, la APA requiere que se le reporte. Está más allá de los propósitos de este libro, pero tanto la varianza como la desviación estándar tienen un papel fundamental en la estadística inferencial cuando se trata de medir errores, intervalos de confianza, varianza explicada, tamaños de efecto,[7] entre otros. Por ello, es una buena práctica estimarlas, entenderlas y reportarlas.

7 Los tamaños de efecto son varias medidas de magnitud y también explican el significado de lo encontrado cuando se trata de estudiar la relación entre dos variables. Por ejemplo, el *d* de Cohen muestra el número de desviaciones estándar entre dos promedios. Muy frecuentemente, los tamaños de efecto son interpretados como un indicativo del significado práctico de lo que se encontró en alguna investigación (VandenBos, 2015, p. 352). En términos más simples, un tamaño de efecto o efecto es el cambio que puede causar una variable sobre otra. Por ejemplo, el aprendizaje de las matemáticas puede ayudar a resolver problemas de la vida real donde las matemáticas estén involucradas; otras cosas siendo iguales.

CAPÍTULO 6

Percentiles, rango de percentil, rangos, cuartiles, deciles, diagramas de caja (*box plots*) y valores estándar

Dice por allí una canción popular: "No hay que llegar primero, sino hay que saber llegar". De esto surge una serie de preguntas: ¿qué tanta diferencia existe entre llegar en cierto lugar que en otro? Tal vez, llegar en cierto lugar podría hacer una gran diferencia. Por ejemplo, si se va a hacer un corte para admitir cierto número de estudiantes mediante un examen de admisión para un pregrado (*i. e.*, el 10% de los puntajes más altos) por semestre, se estaría compitiendo contra cierto número de aspirantes. Si uno es un estudiante promedio, tendría más oportunidad de quedar en ese 10% cuando menos aspirantes haya por semestre. Se tendrían que revisar los números de aspirantes para tomar la decisión de hacer el examen durante el año escolar cuando más posibilidades se tienen de admisión. Además, las y los hacedores de decisiones podrían ver cuándo existen más aspirantes con puntajes destacados, pero que no quedaron en ese 10%, para ser considerados para una próxima admisión, sin tener que hacer el examen de nueva cuenta. En resumen, el lugar en un examen de admisión u otra actividad donde se usen lugares para un *ranking*, es una comparación relativa: *i. e.*, el lugar de cada uno depende en parte de lo que las y los demás hagan.

a. Percentiles

Un percentil señala el porcentaje mediante un *puntaje* tomado de la escala original que está por debajo de cierto valor en un conjunto de datos ordenado de menor a mayor. Una definición más amplia de percentil fue dada por Salkind (2007), quien explicó:

> Un percentil es un punto a lo largo de un continuo de puntajes que dividen la distribución de los puntajes obtenidos por un grupo de referencia en dos partes entre las cuales se encuentra un porcentaje con puntajes del grupo que están por debajo de ese punto. Por ejemplo, si se tratara de una distribución normal de puntajes de coeficiente intelectual (con un promedio de 100 y una *SD* de 15), entonces el 75% de estos puntajes caería por debajo del 110.1. De este modo, el valor del puntaje (110.1) sería el percentil 75°. (p. 755)

Uno de los procedimientos es cuando se determina primero el percentil (*e. g.*, 25°, 50° o 75°) y basado en este, se calcula un puntaje crudo (*i. e.*, un valor en la escala original del conjunto de datos) que sea el límite alto[1] para determinar cuáles puntajes estarían por debajo del percentil. Así es como se hace en Excel este procedimiento (véase: Tabla 6.1): *e. g.*, se predeterminaron los percentiles del 25 al 100 en incrementos de 10 puntos percentiles para determinar estos puntajes crudos. Más acerca del cálculo, las fórmulas del percentil de Excel 2016 comienzan por establecer las funciones de los percentiles inclusivo[2] y exclusivo[3] (Tabla 6.1):

=PERCENTIL.INC(Celda$_1$:Celda$_n$,percentil); y =PERCENTIL.EXC(Celda$_1$:Celda$_n$,percentil).

Con las coordenadas (Celda$_1$:Celda$_n$) se especifica el conjunto de datos mediante coordenadas: *e. g.*, (A1:A20). Los números del set no tienen que estar necesariamente ordenados. Finalmente, se designa el percentil deseado en forma de decimal: *e. g.*, el percentil 25° se convierte en 0.25 (coordenada: C2). Para calcular cada uno de los percentiles, es necesario especificar cada percentil deseado manualmente: no solo el percentil es una opción cuando se integra a la función como una constan-

1 Este límite alto o percentil no necesariamente es parte del conjunto de datos original de donde se está calculando el percentil, pero sí indica qué números están por debajo del percentil.
2 Inclusivo: incluye el valor mínimo y máximo del set, y por lo tanto, se usa el percentil 0° y 100°: i. e., un percentil puede ser $0 \leq P \leq 1$.
3 Exclusivo: no incluye el valor mínimo ni máximo del set y, por lo tanto, no se usa el percentil 0° y 100°: *i. e.*, un percentil puede ser $0 < P < 1$.

te, sino que también puede ser la coordenada de un conjunto de datos o una operación dentro de una celda. En este caso, *no* se puede hacer automáticamente al seleccionar una celda y deslizarla hacia abajo, como en algunos otros casos expuestos en el presente libro: *i. e.*, habría que colocar cada función de percentil una por una.

Tabla 6.1
Fórmula de percentiles

	A	B	C	D
1	70	*P*	In.	Ex.
2	71	25°	=PERCENTILE.INC(A1:A20,0.25)	=PERCENTILE.EXC(A1:A20,0.25)
3	72	35°	=PERCENTILE.INC(A1:A20,0.35)	=PERCENTILE.EXC(A1:A20,0.35)
4	73	45°	=PERCENTILE.INC(A1:A20,0.35)	=PERCENTILE.EXC(A1:A20,0.45)
5	74	55°	=PERCENTILE.INC(A1:A20,0.55)	=PERCENTILE.EXC(A1:A20,0.55)
6	75	65°	=PERCENTILE.INC(A1:A20,0.65)	=PERCENTILE.EXC(A1:A20,0.65)
7	76	75°	=PERCENTILE.INC(A1:A20,0.75)	=PERCENTILE.EXC(A1:A20,0.75)
8	77	85°	=PERCENTILE.INC(A1:A20,0.85)	=PERCENTILE.EXC(A1:A20,0.85)
9	78	95°	=PERCENTILE.INC(A1:A20,0.95)	=PERCENTILE.EXC(A1:A20,0.95)
10	79	100°	=PERCENTILE.INC(A1:A20,1)	No aplica
11	80			
12	81			
13	82			
14	83			
15	84			
16	85			
17	86			
18	87			
19	88			
20	89			

Tabla 6.2
Resultados

	A	B	C	D
1	70	*P*	In.	Ex.
2	71	25°	74.75	74.25
3	72	35°	76.65	76.35
4	73	45°	78.55	78.45
5	74	55°	80.45	80.55
6	75	65°	82.35	82.65
7	76	75°	84.25	84.75
8	77	85°	86.15	86.85
9	78	95°	88.05	88.95
10	79	100°	89	Error
11	80			
12	81			
13	82			
14	83			
15	84			
16	85			
17	86			
18	87			
19	88			
20	89			

Nota: *P* = Percentil; In. = Inclusivo; y Ex. = Exclusivo. La columna A muestra un conjunto de datos que son los porcentajes de calificaciones en cierta materia. Desafortunadamente, Excel 16 no revela una fórmula para calcular el percentil. Excel 2016 define al percentil inclusivo: Devuelve un percentil de valores en un rango, donde el percentil está en el rango 0..1, inclusivo. Puede utilizar esta función para establecer un umbral de aceptación. Por ejemplo, se puede decidir examinar a los candidatos que puntúan por encima del percentil 90. Asimismo, Excel 2016 define al percentil exclusivo: Devuelve el rango de un valor en un conjunto de datos como un porcentaje (0..1, exclusivo) del conjunto de datos.

Más al respecto, Bluman (2018) expresó que los datos de un set se dividen en 100 grupos diferentes para los percentiles y se representan con la letra (*P*) y el número del percentil como subíndice (1,

2,..,100) que le corresponde: P_1, P_2, P_3,…, P_{100}. Dentro de estos 100 grupos,[4] cada valor del set tendrá una posición respecto al resto de los valores y se mide por un porcentaje por debajo del percentil. Por ejemplo, la Tabla 6.1 contiene un conjunto de datos ordenados (n = 20), donde el valor más pequeño = 70 (*i. e.*, mínimo) y el más grande = 89 (*i. e.*, máximo). También, en la Tabla 6.1 se indican las fórmulas para obtener percentiles inclusivo y exclusivo; y en la Tabla 6.2 se indican los resultados de ambos: el percentil inclusivo de 25° (P_{25}) significa que por debajo del valor 74.75, se encuentra el 25% de los valores del set. Asimismo, el percentil exclusivo (P_{25}) señala el porcentaje del 25% por debajo de 74.25. Otro ejemplo es el percentil inclusivo 100° (P_{100}), que indica que el 100% de los valores se encuentra por debajo del 89 (valor máximo de este set; Tabla 6.2). No existe P_{100} exclusivo para ser calculado en Excel.

Cálculos manuales de los percentiles

Para cálculos manuales de percentiles, Bluman (2018) explicó la Ecuación 6.1:

$$\text{Percentil} = \frac{(\text{Número de valores por debajo de } x) + 0.5}{\text{Número total de valores}} \times 100 \quad \text{Ecuación 6.1}$$

Donde:

x = Un número crudo del set

- Número de valores por debajo de x = *e. g.*, por debajo de 70 (Tabla 6.3) no hay otro número en el conjunto de datos. Sin embargo, al calcular sus percentiles no resulta cero con la Ecuación 6.1, porque resulta el percentil 2.5°, así que se puede generalizar que con esta Ecuación 6.1, siempre que sea el primer número del set de menor a mayor, tendrá un percentil diferente de cero.

4 Para efectos prácticos, los percentiles se expresan en números enteros, como lo mencionó Bluman (2018) con los 100 grupos, y así aparecen los percentiles en manuales para evaluar el coeficiente intelectual, entre muchos otros manuales y resultados de exámenes. En contraparte, los percentiles están medidos en una escala continua, donde siempre habrá un número entre otros dos: *e. g.*, entre 2.99 y 3 está 2.999 entre una infinidad de números más. De hecho, según la escala en la que se encuentre un conjunto de datos, pero Excel puede dar como posible resultado un percentil con decimales.

» Número total de valores = *n* o frecuencia total del conjunto de datos

Tabla 6.3 Cálculo manual de percentiles con la Ecuación 6.1

	A	B	C	D	E
1	Set	Frecuencias	Frecuencias acumuladas	Número de valores por debajo de *x*	Percentil
2	70	{=FREQUENCY(A2:A21, A2:A21)}	=B2	0	=(D2+0.5)/20*100
3	71		=B3+C2	=C2	=(D3+0.5)/20*100
4	72		=B4+C3	=C3	=(D4+0.5)/20*100
5	73		=B5+C4	=C4	=(D5+0.5)/20*100
6	74		=B6+C5	=C5	=(D6+0.5)/20*100
7	75		=B7+C6	=C6	=(D7+0.5)/20*100
8	76		=B8+C7	=C7	=(D8+0.5)/20*100
9	77		=B9+C8	=C8	=(D9+0.5)/20*100
10	78		=B10+C9	=C9	=(D10+0.5)/20*100
11	79		=B11+C10	=C10	=(D11+0.5)/20*100
12	80		=B12+C11	=C11	=(D12+0.5)/20*100
13	81		=B13+C12	=C12	=(D13+0.5)/20*100
14	82		=B14+C13	=C13	=(D14+0.5)/20*100
15	83		=B15+C14	=C14	=(D15+0.5)/20*100
16	84		=B16+C15	=C15	=(D16+0.5)/20*100
17	85		=B17+C16	=C16	=(D17+0.5)/20*100
18	86		=B18+C17	=C17	=(D18+0.5)/20*100
19	87		=B19+C18	=C18	=(D19+0.5)/20*100
20	88		=B20+C19	=C19	=(D20+0.5)/20*100
21	89		=B21+C20	=C20	=(D21+0.5)/20*100

Nota: La primera frecuencia acumulada (columna C: C2) pasa directamente de la columna B de esta manera: =B2. Para el número de valores por debajo de *x*, el primer valor es cero, porque van en un orden de menor a mayor y no hay ningún número por debajo de este mínimo. Para calcular todas las frecuencias acumuladas *x*, se hace la primera suma (=B3+C2) en la columna C (C3), se selecciona esa celda y se desliza hacia abajo, para que automáticamente se hagan todas las demás sumas. Para el número de valores por debajo de *x*, se pasan los resultados de la columna C a la D. Solo se tiene que pasar el primer valor: =C2. Luego se selecciona esta celda (D3) y se desliza hacia abajo, para que automáticamente pasen todos los valores restantes. Para el percentil (columna E) solo hay que escribir la fórmula para la primera celda (E2): =(D2+0.5) / 20*100 y deslizarla hacia abajo, para que se calcule el resto de los percentiles automáticamente.

En la Tabla 6.4 se muestran los resultados de las operaciones de la Tabla 6.3.

Tabla 6.4 Resultados del cálculo manual de percentiles

	A	B	C	D	E
1	Set	Frecuencias	Frecuencias acumuladas	Número de valores por debajo de *x*	*Porcentaje por debajo del percentil
2	70	1	1	0	2.5
3	71	1	2	1	7.5
4	72	1	3	2	12.5
5	73	1	4	3	17.5
6	74	1	5	4	22.5
7	75	1	6	5	27.5
8	76	1	7	6	32.5
9	77	1	8	7	37.5
10	78	1	9	8	42.5
11	79	1	10	9	47.5
12	80	1	11	10	52.5
13	81	1	12	11	57.5
14	82	1	13	12	62.5
15	83	1	14	13	67.5
16	84	1	15	14	72.5
17	85	1	16	15	77.5
18	86	1	17	16	82.5
19	87	1	18	17	87.5
20	88	1	19	18	92.5
21	89	1	20	19	97.5

Nota: Por ejemplo, y teoréticamente hablando, se puede asumir que el valor 70 significa cualquier valor entre 69.5 y 70.5 (*cf.* Bluman, 2018). *En la siguiente sección: al *Porcentaje por debajo del percentil* se le llama rango del percentil.

Para estimar el número de valores de un conjunto de datos por debajo de cierto percentil, hay que multiplicar el percentil deseado por la frecuencia total del set: *i. e.*, percentil expresado en decimal multiplicado por *n* = Número de valores por debajo del percentil. Por ejemplo, si se desea obtener el número de valores por debajo del P_{25} = 74.75 (Tabla 6.2) del conjunto de datos (Tabla 6.4; *n* = 20), se convierte P_{25} a decimal (0.25) y se multiplica por *n* (20): *i. e.*, 0.25 x 20 = 5 lugares por debajo (*i. e.*, 70, 71, 72, 73 y 74; Tabla 6.4).

b. Rango del percentil

Al respecto de este concepto de *rango del percentil*, Salkind (2007) escribió:

> Un rango de percentil describe el lugar, o posición, de un puntaje obtenido en comparación a un grupo de referencia. El rango de un percentil obtenido de un puntaje indica el porcentaje de puntajes en el grupo de referencia que están más abajo que el puntaje obtenido. De este modo, un estudiante cuyo puntaje en un examen le ha ganado un rango de percentil de 72 significaría que su puntaje está más alto que el 72 por ciento de aquellos en el grupo de referencia que tomaron el mismo examen. (p. 755)

La definición anterior es una contradicción hasta cierto punto de lo que expresaron Hinkle *et al.* (2003): "El rango de percentil de un puntaje es un punto en la escala percentil que da el porcentaje de puntajes que caen en el puntaje especificado o por debajo de él" (p. 50). Ambos grupos de autores son autoridades en estadística aplicada a las ciencias sociales, así que no habría manera de resolver quién está en lo cierto basándose en sus respectivos textos. Como una sugerencia, se podría decir que cuando se diga que el rango de un percentil está por encima de cierto porcentaje, se mencione a Salkind (2007; $\text{percentil}_x >$ porcentaje), quien parece estar hablando de un rango de percentil *exclusivo*, como el que se puede usar en Excel 2016, y cuando se diga que cierto percentil está hasta cierto porcentaje, se cite a Hinkle *et al.* (2003; $\text{percentil}_x \geq$ porcentaje), quien aparentemente explica un rango de percentil *inclusivo*, también usado en Excel 2016.

Más en detalle, Excel 2016 tiene dos opciones para calcular el rango del percentil: uno inclusivo y otro exclusivo. Sus respectivas fórmulas son (Referencia = Valor para obtener el rango del percentil):

$$=\text{PERCENTRANK.INC}(\text{Celda}_1{:}\text{Celda}_n,\text{referencia})$$
$$=\text{PERCENTRANK.EXC}(\text{Celda}_1{:}\text{Celda}_n,\text{referencia}).$$

En la Tabla 6.5 se muestra cómo calcular estos rangos del percentil; y en la Tabla 6.6 se muestran los resultados de estas funciones, también como el rango, que es explicado más adelante.

Tabla 6.5
Rangos del percentil

	A	B	C
1	Set	Inclusivo	Exclusivo
2	70	=PERCENTRANK.INC(A$2:A$21,A2)	=PERCENTRANK.EXC(A$2:A$21,A2)
3	71	=PERCENTRANK.INC(A$2:A$21,A3)	=PERCENTRANK.EXC(A$2:A$21,A3)
4	72	=PERCENTRANK.INC(A$2:A$21,A4)	=PERCENTRANK.EXC(A$2:A$21,A4)
5	73	=PERCENTRANK.INC(A$2:A$21,A5)	=PERCENTRANK.EXC(A$2:A$21,A5)
6	74	=PERCENTRANK.INC(A$2:A$21,A6)	=PERCENTRANK.EXC(A$2:A$21,A6)
7	75	=PERCENTRANK.INC(A$2:A$21,A7)	=PERCENTRANK.EXC(A$2:A$21,A7)
8	76	=PERCENTRANK.INC(A$2:A$21,A8)	=PERCENTRANK.EXC(A$2:A$21,A8)
9	77	=PERCENTRANK.INC(A$2:A$21,A9)	=PERCENTRANK.EXC(A$2:A$21,A9)
10	78	=PERCENTRANK.INC(A$2:A$21,A10)	=PERCENTRANK.EXC(A$2:A$21,A10)
11	79	=PERCENTRANK.INC(A$2:A$21,A11)	=PERCENTRANK.EXC(A$2:A$21,A11)
12	80	=PERCENTRANK.INC(A$2:A$21,A12)	=PERCENTRANK.EXC(A$2:A$21,A12)
13	81	=PERCENTRANK.INC(A$2:A$21,A13)	=PERCENTRANK.EXC(A$2:A$21,A13)
14	82	=PERCENTRANK.INC(A$2:A$21,A14)	=PERCENTRANK.EXC(A$2:A$21,A14)
15	83	=PERCENTRANK.INC(A$2:A$21,A15)	=PERCENTRANK.EXC(A$2:A$21,A15)
16	84	=PERCENTRANK.INC(A$2:A$21,A16)	=PERCENTRANK.EXC(A$2:A$21,A16)
17	85	=PERCENTRANK.INC(A$2:A$21,A17)	=PERCENTRANK.EXC(A$2:A$21,A17)
18	86	=PERCENTRANK.INC(A$2:A$21,A18)	=PERCENTRANK.EXC(A$2:A$21,A18)
19	87	=PERCENTRANK.INC(A$2:A$21,A19)	=PERCENTRANK.EXC(A$2:A$21,A19)
20	88	=PERCENTRANK.INC(A$2:A$21,A20)	=PERCENTRANK.EXC(A$2:A$21,A20)
21	89	=PERCENTRANK.INC(A$2:A$21,A21)	=PERCENTRANK.EXC(A$2:A$21,A21)

Tabla 6.6
Resultados

	A	B	C	D
1	Set	In.	Ex.	Ra.
2	70	0%	5%	20
3	71	5%	10%	19
4	72	11%	14%	18
5	73	16%	19%	17
6	74	21%	24%	16
7	75	26%	29%	15
8	76	32%	33%	14
9	77	37%	38%	13
10	78	42%	43%	12
11	79	47%	48%	11
12	80	53%	52%	10
13	81	**58%**	**57%**	9
14	82	63%	62%	8
15	83	68%	67%	7
16	84	74%	71%	6
17	85	79%	76%	5
18	86	84%	81%	4
19	87	89%	86%	3
20	88	95%	90%	2
21	89	100%	95%	1

Nota: Solo hay que fijar las coordenadas con el signo $, se desliza la celda seleccionada con la fórmula y se obtendrán tanto el rango del percentil inclusivo[5] como el exclusivo.[6] El porcentaje de los resultados de los rangos del percentil, se obtiene al seleccionar la celda y presionar el símbolo % en el menú de *Home*. Ra = Rango y se refiere al *ranking*; en este caso, el primer

5 Excel 2016 define al rango del percentil inclusivo como: Devuelve el rango de un valor en un conjunto de datos como un porcentaje (0..1, inclusivo) del conjunto de datos. Por otro lado, se puede observar (Tabla 6.6.; columna B) que se puede calcular el rango del percentil del valor mínimo (70) y máximo (89): *i. e.*, $0 \leq PR \leq 1$.

6 Asimismo, Excel 2016 define al rango del percentil exclusivo como: Devuelve el rango de un valor en un conjunto de datos como un porcentaje (0..1, exclusivo) del conjunto de datos. En contraparte, se puede observar (Tabla 6.6.; columna C) que, aunque se puede calcular el rango del percentil del valor mínimo (70) y máximo (89), no se obtienen los rangos del percentil 0° y 100° como en el inclusivo: por lo tanto, $0 < PR < 1$.

rango es el 89 por estar en el primer lugar (1) y así sucesivamente hasta llegar al último lugar (70), que lleva al rango 20. Los rangos se discuten más adelante en la sección C.

Una advertencia de Salkind (2007) fue que los términos percentil y rango del percentil (*percentil rank*) son comúnmente utilizados como sinónimos cuando conceptualmente *no* son lo mismo. La distinción es que el percentil es medido en la escala original e implica un porcentaje por debajo de este. En contraste, el rango del percentil es medido directamente en un porcentaje. Por otro lado, estas dos mediciones coinciden en que están medidas en una escala ordinal.

El rango del percentil de un puntaje es el punto en la escala de percentiles que da el porcentaje de puntajes que caen en ese puntaje designado o más abajo (Hinkle *et al*., 2003, p. 50). Por ejemplo, si se desea saber cuál fue el rango del percentil con un puntaje crudo de 81 (Tabla 6.6), se tendría que decir si va a ser un percentil inclusivo (*i. e*., que incluya los puntajes más abajo y a su nivel) o exclusivo (*i. e*., que solo muestre aquellos por debajo): para el inclusivo tendría el 58% de los puntajes, que estarían a su nivel o más abajo; y para el exclusivo, se tendría el 57% más abajo que el puntaje de 81. Para una discusión más a fondo acerca de percentiles y rango del percentil, véase el Apéndice F.

c. Rangos

Un rango de un conjunto de datos es el lugar que ocupa un valor respecto a los otros. En la Tabla 6.6 se muestran los rangos calculados en Excel 2016. Para establecer rangos, se pueden ordenar los datos jerárquicamente (en Excel 2016 no es necesario). En este caso, se hizo de menor a mayor, pero de mayor a menor también funciona. El número con el valor más alto (89) recibe el rango 1 (primer lugar); 88, el segundo y así sucesivamente hasta llegar al último número, que en este caso es el 70, que recibe el rango de vigésimo. Los rangos van en sentido inverso a los rangos del percentil. Por ejemplo, mientras el número de un conjunto de datos tiene el valor más alto con el rango del percentil inclusivo y el porcentaje más alto, tiene el valor más bajo

Nota: No hay diferencia entre las definiciones que da Excel 2016 y no se muestran las fórmulas de ambos rangos, pero los resultados de las operaciones dan diferentes rangos del percentil (véase: Tabla 6.6).

del rango, pero que significa el primer lugar. Al rango también se le conoce como *ranking*.

Excel 2016 contiene dos funciones para calcular los rangos y las llama: Rank EQ y Rank AVG. Ambas funciones sirven para encontrar los rangos, pero difieren cuando los valores de un set se repiten. Esto es, cuando la función Rank AVG obtenga un promedio de los números que se repiten del set. En contraparte, la función de Rank EQ le otorgará el mismo rango a los valores que se repiten. Como esta última función es más intuitiva, es la que se recomienda. Cuando los valores del set no se repiten, ambas funciones dan el mismo rango. En la Tabla 6.7 se muestra cómo obtener ambos rangos.

Tabla 6.7 Rangos

	A	B	C	"B"	"C"
1	Set	Rango	Rango	Rango	Rango
2	70	=RANK.EQ(A2,A$2:A$21)	=RANK.AVG(A2,A$2:A$21)	20	20
3	71	=RANK.EQ(A3,A$2:A$21)	=RANK.AVG(A3,A$2:A$21)	19	19
4	72	=RANK.EQ(A4,A$2:A$21)	=RANK.AVG(A4,A$2:A$21)	18	18
5	73	=RANK.EQ(A5,A$2:A$21)	=RANK.AVG(A5,A$2:A$21)	17	17
6	74	=RANK.EQ(A6,A$2:A$21)	=RANK.AVG(A6,A$2:A$21)	16	16
7	75	=RANK.EQ(A7,A$2:A$21)	=RANK.AVG(A7,A$2:A$21)	15	15
8	76	=RANK.EQ(A8,A$2:A$21)	=RANK.AVG(A8,A$2:A$21)	14	14
9	77	=RANK.EQ(A9,A$2:A$21)	=RANK.AVG(A9,A$2:A$21)	13	13
10	78	=RANK.EQ(A10,A$2:A$21)	=RANK.AVG(A10,A$2:A$21)	12	12
11	79	=RANK.EQ(A11,A$2:A$21)	=RANK.AVG(A11,A$2:A$21)	11	11
12	80	=RANK.EQ(A12,A$2:A$21)	=RANK.AVG(A12,A$2:A$21)	10	10
13	81	=RANK.EQ(A13,A$2:A$21)	=RANK.AVG(A13,A$2:A$21)	9	9
14	82	=RANK.EQ(A14,A$2:A$21)	=RANK.AVG(A14,A$2:A$21)	8	8
15	83	=RANK.EQ(A15,A$2:A$21)	=RANK.AVG(A15,A$2:A$21)	7	7
16	84	=RANK.EQ(A16,A$2:A$21)	=RANK.AVG(A16,A$2:A$21)	6	6
17	85	=RANK.EQ(A17,A$2:A$21)	=RANK.AVG(A17,A$2:A$21)	5	5
18	86	=RANK.EQ(A18,A$2:A$21)	=RANK.AVG(A18,A$2:A$21)	4	4
19	87	=RANK.EQ(A19,A$2:A$21)	=RANK.AVG(A19,A$2:A$21)	3	3
20	88	=RANK.EQ(A20,A$2:A$21)	=RANK.AVG(A20,A$2:A$21)	2	2
21	89	=RANK.EQ(A21,A$2:A$21)	=RANK.AVG(A21,A$2:A$21)	1	1

d. Cuartiles

Otra manera de ver el lugar que ocupan los datos ordenados de un set son los cuartiles, que dividen un conjunto de datos en *cuatro* grupos iguales y que pueden ser denotados con una C_x y un subíndice del número de cuarto correspondiente. Los cuartiles coinciden con los siguientes percentiles: $C_1 = P_{25}$, $C_2 = P_{50}$, $C_3 = P_{75}$ y $C_4 = P_{100}$. Más al respecto, Bluman (2018, p. 155) muestra una ingeniosa y simple forma de calcular los cuartiles:

- Primer paso: Organizar los datos de menor a mayor: el valor más pequeño es el mínimo de la jerarquía del set y el valor más grande es el máximo.
- Segundo paso: Encontrar la mediana de todo el conjunto de datos que representa en el C_2.
- Tercer paso: Localizar la mediana por debajo del C_2 (esta será el C_1).
- Cuarto paso: Identificar la mediana por encima del C_2 (esta será el C_3).

En las Tablas 6.8 y 6.9 se muestra cómo utilizar los pasos de Bluman (2018) para obtener los cuartiles del conjunto de datos pares manualmente en Excel: =(celda$_1$+celda$_n$) / 2 (*i. e.*, en este contexto, la *n* como subíndice significa hasta dónde se encuentra la última celda del conjunto de datos que se desea utilizar); y usando la función de la mediana: =MEDIAN(celda$_1$:celda$_n$). Para obtener el C_2 en forma manual, se ordenan los datos de menor a mayor. En este caso, es un conjunto de datos pares ($n = 20$), se identifican los dos valores de en medio (79 y 80) y se dividen por 2, lo cual da 79.5. Luego, se crea un subset por debajo de 79.5 y se vuelve a sacar la mediana para obtener el $C_1 = 74.50$. Asimismo, se crea otro subset por arriba de 79.5 y se repite la obtención de la mediana para calcular el $C_3 = 84.50$. Para el uso de la antes mencionada función de la mediana, los cuartiles se pueden obtener al sacar la mediana para el C_2 con todo el conjunto de datos: =MEDIAN(A2:A21). Para el C_1, se consideran solo los valores del set menores a 74.5 y se le saca la mediana: =MEDIAN(C2:C11).

Finalmente, para el C_3 se emplean los valores mayores a 74.5: =MEDIAN(E12:E21).

Tabla 6.8 Cuartiles calculados a mano y en Excel

	A	B	C	D	E	F
1	Set		Set			
2	70	Mínimo	70			
3	71		71			
4	72		72	Segundo paso:		
5	73		73	C_1 =74.50		
6	74		**74**	=(C2:C11)/2		
7	75		**75**	Una alternativa		
8	76		76	=MEDIAN(C2:C11)		
9	77		77			
10	78	Primer paso:	78			
11	**79**	C_2 = 79.50	79			
12	**80**	=(A11+A12)/2			80	
13	81	Una alternativa			81	
14	82	=MEDIAN(A2:A21)			82	Tercer paso:
15	83				83	C_3 = 84.50
16	84				**84**	=(E16+E17)/2
17	85				**85**	Una alternativa
18	86				86	=MEDIAN(E12:E21)
19	87				87	
20	88				88	
21	89	Máximo			89	

Nota: Los números en negrillas indican que la mediana se encuentra entre estos valores por ser sets de pares. Una aclaración es que los cuartiles/medianas del set y de los subsets aparecerían donde están las funciones.

El procedimiento de Bluman (2018) es fácil de utilizar cuando se tienen conjuntos de datos relativamente pequeños, pero para sets más grandes Excel tiene dos funciones para calcularlos automáticamente para un cuartil inclusivo[7] y un cuartil exclusivo,[8] respectivamente (Tabla 6.8):

7 Excel 2016 dice que la función del cuartil inclusivo da como resultado un cuartil de un conjunto de datos basado en los valores de los percentiles, que va del 0…1 y que es inclusivo.
8 Excel 2016 dice que la función del cuartil exclusivo da como resultado un cuartil de un conjunto de datos basado en los valores de los percentiles, que va del 0…1 y que es exclusivo.

=QUARTILE.INC(Celda$_1$:Celda$_n$,cuartil); y =QUARTILE.EXC(Celda$_1$:Celda$_n$,cuartil).

Para el cuartil inclusivo, los cuartiles pueden ir del 0 al 4 (*i. e.*, el 0 será el valor mínimo y el 4, el máximo de cuartiles del conjunto de datos), pero para el cuartil exclusivo solo pueden ir del 1 al 3. En la Tabla 6.8, con las columnas B y C, se muestra el cálculo de ambos cuartiles, así como el mínimo y máximo valores del set.

Tabla 6.9 Cuartiles calculados en Excel

	A	B	C
1	Set	Excel INC.	Excel EXC.
2	70	Mínimo = 70	
3	71	=QUARTILE.INC(A2:A21,0)	No aplica
4	72		
5	73		
6	74	C_1 = 74.75	C_1 = 74.25
7	75	=QUARTILE.INC(A2:A21,1)	=QUARTILE.EXC(A2:A21,1)
8	76		
9	77		
10	78	C_2 = 79.50	C_2 = 79.50
11	79	=QUARTILE.INC(A2:A21,2)	=QUARTILE.EXC(A2:A21,2)
12	80		
13	81		
14	82	C_3 = 84.25	C_3 = 84.75
15	83	=QUARTILE.INC(A2:A21,3)	=QUARTILE.EXC(A2:A21,3)
16	84		
17	85		
18	86		
19	87		
20	88	Máximo = 89	No aplica
21	89	=QUARTILE.INC(A2:A21,4)	

Por otro lado, no coinciden los resultados del método de Bluman (2018) y los de Excel (QUARTILE.INC y QUARTILE.EXC), como se puede ver a continuación, respectivamente (Tabla 6.10): C_1 (74.50 ≠ 74.75 ≠ 74.25) y C_3 (84.50 ≠ 84.25 ≠ 84.75). Donde sí coinciden los tres

procedimientos es con el C_2. A pesar de estas discrepancias, esto no afecta en cómo se distribuyen los datos en los cuartiles. Esto es, el C_1, tanto para el método de Bluman (2018) como para los de Excel, contiene cinco valores por debajo de este. Del mismo modo sucede para el resto de los cuartiles: cinco valores por debajo entre los cuartiles.

Tabla 6.10 Comparación entre Bluman (2018) y Excel

	A	B	C	D
1	Set	Bluman (2018)	Excel inclusivo	Excel exclusivo
2	70			
3	71			
4	72			
5	73			
6	74			
7		$C_1 = 74.50$	$C_1 = 74.75$	$C_1 = 74.25$
8	75			
9	76			
10	77			
11	78			
12	79			
13		$C_2 = 79.50$	$C_2 = 79.50$	$C_2 = 79.50$
14	80			
15	81			
16	82			
17	83			
18	84			
19		$C_3 = 84.50$	$C_3 = 84.25$	$C_3 = 84.75$
20	85			
21	86			
22	87			
23	88			
24	89			

Los cuartiles también se pueden emplear para medir la variación entre ellos. A esta medición se le llama rango intercuartil (*interquartile range*) y un ejemplo de este rango es: $C_3 - C_1$. De la Tabla 6.9, $84.50(C_3) - 74.50(C_1) = 10$, que indica la diferencia entre estos cuartiles.

Un comentario acerca de los cuartiles es que están medidos en una escala ordinal, al igual que los percentiles, así que tienen limitaciones a la hora de intentar hacer operaciones con ellos (véase: Tabla 2.1 y Apéndice F, para más detalles al respecto).

e. Deciles

Otra forma de dividir los datos para un análisis descriptivo es en deciles,[9] que ocupan los datos ordenados de un set. Los deciles dividen un set en 10 grupos iguales: *i. e.*, entre un decil y otro, se encuentra el 10% de los valores. Los deciles pueden ser denotados con una D_x y un subíndice del número de cuarto correspondiente. Los deciles coinciden con ciertos percentiles: $D_1 = P_{10}$, $D_2 = P_{20}$, $D_3 = P_{30}$,…,$D_9 = P_{90}$. Los percentiles, cuartiles y deciles coinciden en: $P_{50} = C_2 = D_5$. Una fórmula para encontrar los deciles es:

Lugar en la lista por abajo del decil = $(n + 1) \times (D_x / 10)$ Ecuación 6.2

Donde:
n = Tamaño de la muestra/Frecuencia total
D_x = Decil deseado: 1°, 2°, 3°,…,10°

Por ejemplo (Tabla 6.10), el decil 1° se obtiene al aplicar la Ecuación 6.2 (columna C): =(20+1)*(1/10); donde 20 es el tamaño de la muestra (n) y 1, el decil deseado. El resultado de esta ecuación aparecería en la misma columna C en Excel, pero para evitar confusión se colocó en la columna "C" para ilustrar el ejemplo (Tabla 6.11). Dado que Excel 2016 no tiene una función para calcular los deciles, se calcularon los percentiles equivalentes a estos (*i. e.*, $D_1 = P_{10}$,…,$D_{10} = P_{100}$; mostrados en el párrafo anterior; véase Tabla 6.1 para más detalles sobre cómo calcular percentiles).

La interpretación de los deciles de la Tabla 6.11 en la columna "C", es que el primer decil (D_1) se encuentra a 2.1 posiciones por encima del primer valor (70) de la secuencia ascendente. En otras pal-

9 Un decil es la posición de un valor en una distribución expresado en decenas (Bluman, 2018, p. 157).

abras, el 10% de los valores están por debajo del D_1 y cuando este es calculado en percentiles inclusivos, porque $D_1 = P_{10}$ resulta ser 71.9. De la misma manera, se puede interpretar el resto de los deciles hasta llegar al $D_{10} = 89$, que está a 21 posiciones por encima del primer valor del set. Lo anterior significa que todo el conjunto de datos estaría por debajo de este decil, porque $n = 20$.

Tabla 6.11 Fórmula de los deciles

	A	B	C	"C"	D
1	Set	Decil	Fórmula	Lugar en la lista	Deciles
2	70				
3	71	1°	=(20+1)*(1/10)	2.1	71.9
4	72				
5	73	2°	=(20+1)*(2/10)	4.2	73.8
6	74				
7	75	3°	=(20+1)*(3/10)	6.3	75.7
8	76				
9	77	4°	=(20+1)*(4/10)	8.4	77.6
10	78				
11	79	5°	=(20+1)*(5/10)	10.5	79.5
12	80				
13	81	6°	=(20+1)*(6/10)	12.6	81.4
14	82				
15	83	7°	=(20+1)*(7/10)	14.7	83.3
16	84				
17	85	8°	=(20+1)*(8/10)	16.8	85.2
18	86				
19	87	9°	=(20+1)*(9/10)	18.9	87.1
20	88				
21	89	10°	=(20+1)*(10/10)	21	89

Nota: Los deciles de la columna D fueron calculados con la fórmula de los percentiles: =PERCENTIL.INC(Celda$_1$:Celda$_n$,percentil).

Una última nota acerca de los deciles es que están medidos en una escala ordinal, al igual que los percentiles y cuartiles, así que tienen ciertas limitaciones (véase: Tabla 2.1 y Apéndice F, para más detalles al respecto).

f. Diagramas de caja

Se le da crédito a Tukey (1977) por la creación de los diagramas de caja (*box plots*), que sirven para poder observar la dispersión gráfica de los datos en un set. De igual manera, es una medida de tendencia central y ayuda a ver si algunos valores se alejan mucho de la mediana, a los que también se les conoce como valores atípicos (véase: Apéndice E para más información acerca de estos valores). Los valores atípicos pueden sesgar una distribución de datos y esto puede hacer cuestionable el uso del modelo de distribución normal para otros análisis, como comparaciones de grupos (*e. g.*, prueba *t* y análisis de la varianza; véase: Hinkle *et al.*, 2003) y relación entre variables (*r* de Pearson y regresión; véase: Tabachnick, & Fidell, 2019). *De nuevo*: la medida de tendencia central usada en los diagramas de caja es la mediana; y la medida de dispersión es el ya mencionado previamente rango intercuartil[10] (*i. e.*, llamado rango entre-cuartiles). Para crear un diagrama de caja, se necesitan cinco valores:

1. El valor máximo del conjunto de datos = Máx = C_4.
2. El tercer cuartil (equivalente al percentil 75°) = C_3.
3. El segundo cuartil (equivalente al percentil 50°) [Mediana] = Mdn = C_2.
4. El primer cuartil (equivalente al percentil 25°) = C_1.
5. El valor mínimo del conjunto de datos = Mín = C_0.

En la Tabla 6.12 se muestran las funciones de Excel (columna C) para calcular los valores que son incluidos en un diagrama de caja. Se usó la fórmula del cuartil inclusivo, porque con esta se pueden calcular los cinco valores para delimitar el diagrama de caja: =QUARTILE.INC(Celda$_1$:Celda$_n$,cuartil). En la columna "C" se muestran los resultados de los cuartiles.

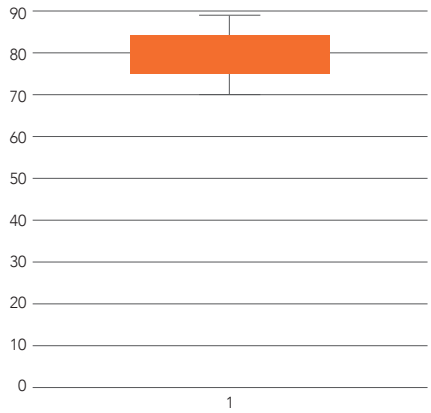
10 El rango intercuartil es la diferencia entre el $C_3 = P_{75}$ y el $C_1 = P_{25}$.

Tabla 6.12 Diagramas de caja

	A	B	C	"C"
1	Set		Fórmulas	
2	70	Mín/C_0	=QUARTILE.INC(A2:A21,0)	70
3	71	C_1	=QUARTILE.INC(A2:A21,1)	74.75
4	72	Mdn/C_1	=QUARTILE.INC(A2:A21,2)	79.50
5	73	C_3	=QUARTILE.INC(A2:A21,3)	84.25
6	74	Máx/C_4	=QUARTILE.INC(A2:A21,4)	89
7	75			
8	76			
9	77			
10	78			
11	79			
12	80			
13	81			
14	82			
15	83			
16	84			
17	85			
18	86			
19	87			
20	88			
21	89			

En la Figura 6.1 se muestran los cinco valores (Tabla 6.12, columna "C") que forman el diagrama de caja. Tres cuartiles forman la caja y el mínimo y el máximo, los bigotes. En detalle, la parte baja de la caja la conforma el cuartil 1° (C_1), la línea de en medio es la mediana o cuartil 2° (C_2) y la parte alta de la caja es el cuartil 3° (C_3). Asimismo, los bigotes son representados por el mínimo (C_0) y el máximo (C_4), respectivamente.

Figura 6.1 Diagrama de caja



Nota: Para la creación del diagrama de caja, se presiona: *Insert, Recomend Graphs, Box* y *Whisker*.

Un diagrama de caja puede ser utilizado para observar si existen observaciones atípicas[11] potencialmente. Para identificar estas posibles observaciones atípicas, hay que establecer una *frontera razonable inferior* (FRI) y una *frontera razonable superior* (FRS). El primer paso es obtener el rango intercuartil (RIC; Ecuación 6.3):

$$RIC = C_3 - C_1 \qquad \text{Ecuación 6.3}$$

Luego, se obtienen la FRI (Ecuación 6.4) y la FRS (Ecuación 6.5):

$$FRI = C_1 - 1.5\,(REC) \qquad \text{Ecuación 6.4}$$
$$FRS = C_3 + 1.5\,(REC) \qquad \text{Ecuación 6.5}$$

En la Tabla 6.13 se muestra el conjunto de datos de la Tabla 6.11 con dos valores adicionales para ponerlos a prueba (*i. e.*, posibles observaciones atípicas): 50 y 100, que serían los nuevos mínimo y máximo, respectivamente. Estos dos valores podrían ser considerados ex-

11 Una observación atípica (*outlier*) es definida como una puntuación inusual o extrema en un conjunto de datos y que requiere atención especial (Hinkle *et al.*, 2003, p. 63). Sería en parte el juicio de un o una investigadora el concluir que una observación es atípica y que debe de ser eliminada o modificada. Ya sea que se elimine o mantenga habría que argumentar la decisión.

tremos a simple vista, pero habría que ver si rebasan las fronteras: FRI y FRS. Los cuartiles han sido calculados (Tabla 6.11, columna C).

Sustituyendo en forma manual de la Ecuación 6.3 (en Excel: Tabla 6.13; celda C8):

$$\text{RIC} = 84.75 - 74.25 = 10.50$$

Sustituyendo con la Ecuación 6.4 y Ecuación 6.5, respectivamente (en Excel: Tabla 6.13; celdas C9 y C10):

$$\text{FRI} = 74.25 - 1.5\,(10.50) = 58.50$$
$$\text{FRS} = 84.75 + 1.5\,(10.50) = 100.50$$

Tabla 6.13 Diagramas de caja con observación atípica

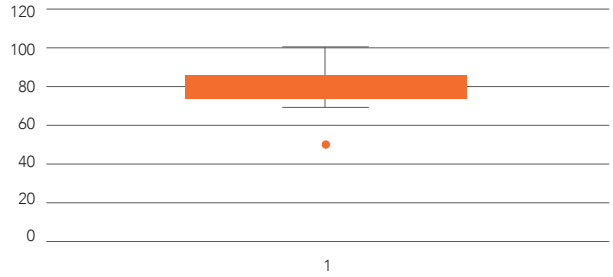
	A	B	C	"C"	D
1	Set		Fórmulas		Resultados de la Tabla 6.10
2	50	Mín/C_0	=QUARTILE.INC(A2:A23,0)	50	70
3	70	C_1	=QUARTILE.INC(A2:A23,1)	74.25	74.75
4	71	Mdn/C_1	=QUARTILE.INC(A2:A23,2)	79.50	79.50
5	72	C_3	=QUARTILE.INC(A2:A23,3)	84.75	84.25
6	73	Máx/C_4	=QUARTILE.INC(A2:A23,4)	100	89
7	74				
8	75	rec	=C5-C3	10.50	
9	76	fri	=C3-(C8*1.5)	58.50	
10	77	frs	=C5+(C8*1.5)	100.50	
11	78				
12	79	Reorganizando los datos			
13	80	Valor por debajo de FRI/ Observación atípica	50		
14	81	fri	58.50		
15	82	C_1	74.25		
16	83	Mdn/C_1	79.50		
17	84	C_3	84.75		
18	85	frs	100.50		

Continúa...

	A	B	C	"C"	D
19	86	Valor por encima de FRS/ Observación atípica			
20	87				
21	88				
22	89				
23	100				

En la Figura 6.2 se muestra cómo el valor 50 va más allá de donde termina la FRI = 58.50: *i. e.*, 58.50 > 50. Por lo tanto, 50 parece ser una observación atípica; otras cosas siendo iguales.[12] Por otro lado, el valor 100 no saltó la FRS, así que no se consideraría una observación atípica; otras cosas siendo iguales.

Figura 6.2 Diagrama de caja con observaciones atípicas



g. Valores estándar

Los valores estándar o puntajes *z* están medidos en un número de desviaciones estándar del promedio, porque la unidad de medición es la *SD*. En otras palabras, los puntajes *z* describen la posición relativa de un solo puntaje en toda la distribución en términos del promedio y la *SD*. Adicionalmente, se pueden hacer operaciones aritméticas con ellos y se pueden utilizar para análisis multivariados, como análisis múltiple de la varianza (MANOVA), regresiones y análisis de factores, entre muchos otros más (para estos análisis, véase: Hair *et al.*, 2019;

12 *Otras cosas siendo iguales* significa que otras variables que podrían tener algún efecto no fueron consideradas para obtener una conclusión.

Tabachnick, & Fidell, 2019). De hecho, es recomendable transformar los datos crudos de algún conjunto de datos a valores *z* para tener una misma métrica y así poder realizar análisis multivariados, como los antes descritos.

$$z_i = \frac{x_i - \bar{x}}{SD}$$ Ecuación 6.6

Donde:
z_i = Un valor estándar individual
x_i = Un valor individual del conjunto de datos
$\bar{x}$ = Promedio del conjunto de datos
SD = Desviación estándar del conjunto de datos

Los puntajes *z* tienen como promedio siempre el 0 y su desviación estándar (SD_z) = 1. La escala puede ir del infinito negativo (-∞) al infinito positivo (+∞). Los valores *z* con signo negativo están por debajo del promedio y aquellos con signo positivo están por encima de este. Los valores *z* se asocian con la distribución normal estándar (también llamada distribución normal estandarizada) que se usa para las probabilidades. Está más allá de los propósitos descriptivos del presente libro discutir esta distribución, pero se recomienda consultar a Ponce-Renova (2019) y Ponce-Renova (2020), para más información y ejemplos al respecto de esta distribución normal estándar.

Los valores estándar pueden ayudar a solucionar problemas que traen escalas, como la de valores crudos, la de percentiles y la de rangos del percentil cuando se trata de saber el rango/*ranking* de un valor y llevar a cabo operaciones aritméticas. Aunque se pueden calcular el promedio y la *SD*, el utilizar valores crudos puede dar problemas cuando se trata de comparar las calificaciones de una o un estudiante en tres diferentes materias, porque no se sabe el rango/*ranking* o la distancia que existe entre una calificación en particular del promedio. Por ejemplo, si la o el alumno está cursando matemáticas, donde obtiene un 75; ciencias naturales, con un 80; y en lectura, con un 85, no son comparables, porque son diferentes materias y no sabemos el lugar de estas calificaciones en estos tres sets de datos (más adelante, se da un ejemplo de estas comparaciones; véase: Tabla 6.14). Asimismo, los

percentiles y los rangos del percentil pueden indicar el lugar de un valor en el conjunto de datos, pero no se sabe cuál es el promedio en valores crudos ni la *SD*, así como tampoco se pueden hacer operaciones aritméticas con estos valores. Para solucionar esta situación de *ranking* y operaciones aritméticas, los valores estándar son una alternativa.

En la Tabla 6.14 se expone cómo convertir la Ecuación 6.6 a Excel, para obtener los puntajes *z* de forma manual. Un ejemplo es:

$=(\text{Celda}_{B2}-\text{Celda}_{C2})/\text{Celda}_{D2}$
$\text{Celda}_{B2} = x_i$
$\text{Celda}_{C2} = \bar{x}$
$\text{Celda}_{D2} = SD$

Una vez que se ejecutan las funciones y se les otorga un rango con una comparación entre los puntajes *z*, se puede notar que la calificación de matemáticas, a pesar de ser el puntaje crudo más bajo, es la que tiene el valor *z* más alto en comparación con las de ciencias naturales y lectura. Por lo tanto, la alumna o el alumno tiene un mejor rango en matemáticas relativo a los valores *z* de las otras materias. También, se puede apreciar que la asignatura de lectura, que tuvo la calificación cruda más alta relativa a las demás, obtuvo el valor más bajo de los valores *z*.

Tabla 6.14 Puntajes *z*

	A	B	C	D	E	"E"	F
1	Materia	Calificaciones	Promedio de la materia	*SD* de la materia	Fórmula/*z*	Fórmula/*z*	Comparación/ *Ranking* dentro de la misma/mismo estudiante
2	Matemáticas	75	65	10	=(B2-C2)/D2	1	Primero
3	Ciencias Naturales	80	80	5	=(B3-C3)/D3	0	Segundo
4	Lectura	84	90	6	=(B4-C4)/D4	-1	Tercero

Además de la forma manual, Excel tiene una función donde es necesario especificar el promedio y la *SD* del conjunto de datos (Tabla 6.15): $=\text{STANDARDIZE}(\text{Celda}_1:\text{Celda}_n,\text{promedio},SD)$. Tanto el promedio como la *SD*, se pueden introducir en forma de una constante o

de una celda. En el caso de la Tabla 6.15, el promedio y la *SD* fueron introducidos por medio de celdas y se fijaron con el signo numérico $, para que, al deslizar la fórmula en la columna C, no cambiaran: *e. g.*, =STANDARDIZE(B2,B$22,B$23).

Tabla 6.15 Conversión a valores estándar

	A	B	C	D	"C"	"D"
1		Set	Fórmula para los valores z	Forma manual	Fórmula para los valores z	Forma manual
2		70	=STANDARDIZE(B2,B$22,B$23)	=(B2-B$22)/B$23	-1.61	-1.61
3		71	=STANDARDIZE(B3,B$22,B$23)	=(B3-B$22)/B$23	-1.44	-1.44
4		72	=STANDARDIZE(B4,B$22,B$23)	=(B4-B$22)/B$23	-1.27	-1.27
5		73	=STANDARDIZE(B5,B$22,B$23)	=(B5-B$22)/B$23	-1.10	-1.10
6		74	=STANDARDIZE(B6,B$22,B$23)	=(B6-B$22)/B$23	-0.93	-0.93
7		75	=STANDARDIZE(B7,B$22,B$23)	=(B7-B$22)/B$23	-0.76	-0.76
8		76	=STANDARDIZE(B8,B$22,B$23)	=(B8-B$22)/B$23	-0.59	-0.59
9		77	=STANDARDIZE(B9,B$22,B$23)	=(B9-B$22)/B$23	-0.42	-0.42
10		78	=STANDARDIZE(B10,B$22,B$23)	=(B10-B$22)/B$23	-0.25	-0.25
11		79	=STANDARDIZE(B11,B$22,B$23)	=(B11-B$22)/B$23	-0.08	-0.08
12		80	=STANDARDIZE(B12,B$22,B$23)	=(B12-B$22)/B$23	0.08	0.08
13		81	=STANDARDIZE(B13,B$22,B$23)	=(B13-B$22)/B$23	0.25	0.25
14		82	=STANDARDIZE(B14,B$22,B$23)	=(B14-B$22)/B$23	0.42	0.42
15		83	=STANDARDIZE(B15,B$22,B$23)	=(B15-B$22)/B$23	0.59	0.59
16		84	=STANDARDIZE(B16,B$22,B$23)	=(B16-B$22)/B$23	0.76	0.76
17		85	=STANDARDIZE(B17,B$22,B$23)	=(B17-B$22)/B$23	0.93	0.93
18		86	=STANDARDIZE(B18,B$22,B$23)	=(B18-B$22)/B$23	1.10	1.10
19		87	=STANDARDIZE(B19,B$22,B$23)	=(B19-B$22)/B$23	1.27	1.27
20		88	=STANDARDIZE(B20,B$22,B$23)	=(B20-B$22)/B$23	1.44	1.44
21		89	=STANDARDIZE(B21,B$22,B$23)	=(B21-B$22)/B$23	1.61	1.61
22	$\bar{x}$	79.50	=AVERAGE(C2:C21)	=AVERAGE(D2:D21)	0	0
23	SD	5.92	=STDEV.S(C2:C21)	=STDEV.S(C2:C21)	1	1

Nota: El promedio ($\bar{x}$) y la *SD* de la columna B también fueron obtenidos por medio de las fórmulas: =AVERAGE(B2:B21) y =STDEV.S(B2:B21).

Entre algunas de las propiedades de los valores *z* están:

1. La distribución de los valores *z* preserva una forma similar a la de los valores crudos del conjunto de datos original.
2. El promedio de los valores *z siempre* va a ser 0, independientemente del promedio del conjunto de datos.
3. La varianza de la distribución de los valores *z* siempre es 1, porque la *SD* es la raíz cuadrada de la varianza que también es 1.

Preguntas para resolver del Capítulo 6

- ¿Qué es un percentil?
- ¿Qué significa el percentil 65°?
- ¿Cuál es la diferencia entre un percentil inclusivo vs. uno exclusivo?
- ¿Qué es un rango del percentil?
- ¿Cuál es la diferencia entre un rango del percentil inclusivo vs. uno exclusivo?
- ¿Cuál es la diferencia entre un percentil y un rango del percentil?
- ¿Qué es el rango desde la perspectiva del Capítulo 6?
- ¿Qué es un cuartil?
- ¿Cuáles cuartiles coinciden con los percentiles?
- ¿En qué consiste el método de Bluman (2018)?
- ¿Cuál es la diferencia entre el cuartil inclusivo y el cuartil exclusivo?
- ¿Cuáles son las diferencias en la estimación de cuartiles entre el proceso de Bluman (2018) y Excel?
- ¿Qué es el entre-cuartiles?
- ¿Qué es un decil?
- ¿Dónde coinciden los deciles con los cuartiles y los percentiles?
- ¿De qué sirve un diagrama de caja?
- ¿Cómo se calcula un diagrama de caja?
- ¿Cómo se pueden calcular observaciones atípicas mediante cajas de diagrama?
- ¿Qué es un valor estándar?
- ¿Cómo se calcula un valor estándar?
- ¿Para qué sirven los valores estándar?
- ¿Cómo se demuestra que el promedio de un conjunto de datos estándar siempre es 0 y su *SD* = 1?

Problemas para resolver

Problema. La primera instrucción es copiar el conjunto de datos (n = 39) en Excel y ordenarlos de menor a mayor. Lo siguiente es contestar las preguntas con funciones de Excel, así como manualmente cuando sea posible:

a) ¿Cuáles son los percentiles inclusivos y exclusivos del P_{25}, P_{50} y P_{75}, y qué significan?; b. ¿Cuáles son los rangos del percentil inclusivos y exclusivos de cada número del set y qué significan?; c. ¿Cuáles son los rangos de los valores en los que caen los números del set y qué significa esto?; d. ¿Cuáles son los cuartiles inclusivos y exclusivos, y qué significan?; e. ¿Cuál es el rango intercuartil y qué significa?; f. ¿Existen aparentemente observaciones atípicas según los diagramas de caja?; y g. ¿Cuáles son los valores estándar del set y qué significan?

Tabla P-6.1 Puntajes de aciertos

	A	B	C	D	E	F	G	H
1	46	10	100	26	27	9	96	63
2	17	49	55	83	10	49	75	91
3	38	65	73	64	99	65	42	65
4	13	4	60	88	12	90	16	58
5	81	52	84	86	60	33	43	

Preguntas para reflexionar

- ¿Cómo se podrían usar los percentiles, cuartiles y deciles en alguna investigación propia?
- Además de los percentiles, cuartiles y deciles, ¿qué otras formas existen de organizar los datos de una manera jerárquica?
- Se podría decir que siempre que un valor supera la frontera razonable inferior (FRI) y la frontera razonable superior (FRS) de un diagrama de caja, ¿es una observación atípica?
- En todos los casos, ¿no es una observación atípica si está dentro de estas fronteras?

» ¿Existen análisis que requieren que los datos crudos de las variables sean convertidos a valores estándar?

Opinión del Autor

Podría pasar —como me sucedió a mí ya hace algunas décadas atrás— que uno piense que el conocer la posición relativa del valor de un número no tenga gran importancia. ¡No es así! El saber la ubicación de cierto valor nos podría indicar el nivel que se tiene en algún rendimiento en alguna área del aprendizaje, atributo mental,[13] comportamiento, entre muchas otras posibles características de una persona o hasta de objetos (*i. e.*, escuelas, regiones, estados y países, entre otros) y de observaciones (*i. e.*, calificaciones, puntajes en exámenes de admisión, etcétera). Esta posición, en un espectro como el aprendizaje, nos puede indicar si alguna o algún estudiante con cierto nivel en un atributo está destacando, está en el promedio o está por debajo del promedio. Estas descripciones les dan un significado a los promedios. De hecho, cuando se aplican baterías de inteligencia, como las pruebas de Wechsler (Wechsler, 1997), estas contienen información de cómo convertir un puntaje crudo a percentiles, cuartiles, deciles y valores estándar, con el fin de poder ubicar este atributo. Esta ubicación permite observar si el atributo de una alumna o un alumno podría estar en la zona de recibir educación especial o de aptitudes sobresalientes.

Otros aspectos descriptivos a los que hay que poner atención son los diagramas de caja, que son relevantes cuando se están conociendo los datos y nos dan una idea visual de estos a nivel de una sola variable (*i. e.*, univariado). Estas cajas pueden ser un primer paso para

13 VandenBos (2015, p. 859) ejemplifica un atributo mental como una actitud (relativamente durable y general evaluación de un objeto, persona, grupo, tema o concepto, que va de lo negativo a lo positivo. Se asume que se origina de creencias en específico, emociones y comportamientos previos que se asocian a lo que se haya evaluado); funcionamiento emocional (felicidad, miedo, ira, sorpresa, disgusto, asco/desprecio y tristeza); inteligencia; habilidades cognitivas (razonamiento, comprensión, abstracción, etcétera), aptitudes (capacidad de adquirir competencia o habilidad: potencial); valores (un principio moral, social o estético aceptado por un individuo o sociedad acerca de lo que es bueno, deseable o importante acerca de algo); intereses (un actitud caracterizada por la necesidad de dar atención selectiva a algo que es significativo para el individuo, como una actividad, meta o un área de investigación) y características personales (edad, discapacidad, raza, género, educación y estatus socioeconómico).

la detección de observaciones atípicas. El tipificar una observación como atípica requeriría un argumento por parte de la investigadora o el investigador. Habría que buscar en la literatura qué se ha reconocido como una observación atípica o se podría elaborar un argumento propio para ello. Los valores estándar tienen aún más peso que los aspectos meramente descriptivos, como los percentiles, cuartiles y deciles con su *ranking* relativo, porque pueden ser usados en estadística inferencial y para análisis multivariados, como *análisis de componentes principales* y *análisis exploratorio de factores* (véase: Hair *et al.*, 2019; Tabachnick, & Fidell, 2019). Además, los valores estándar nos permiten saber la probabilidad de que suceda un valor (está más allá de los propósitos del presente libro ahondar en asuntos de estadística inferencial, así que se recomienda consultar a Hinkle *et al.*, 2003, para observar la asociación entre valores estándar y probabilidad).

Apéndice A. Recursos de Excel en una página de Microsoft

Tabla A-1 Temas, contenido y dirección de recursos de Excel

Tema	Contenido	Dirección
Comienzo rápido (*Quick start*)	Muestra cómo comenzar a usar una hoja de cálculo, las máquinas (computadoras, tabletas, teléfonos) que se pueden utilizar, así como gráficas y tablas.	https://support.microsoft.com/en-us/office/create-a-workbook-in-excel-94b00f50-5896-479c-b0c5-ff74603b35a3?wt.mc_id=otc_excel
Introducción a Excel (*Intro to Excel*)	Un libro de trabajo es un archivo que contiene una o más hojas de trabajo para ayudar a organizar los datos. Se puede crear un nuevo libro de trabajo, a partir de un libro en blanco o una plantilla.	https://support.microsoft.com/en-us/office/create-a-new-workbook-ae99f19b-cecb-4aa0-92c8-7126d6212a83?wt.mc_id=otc_excel
Renglones y columnas (*Rows & columns*)	Enseña cómo insertar y borrar información. Los límites de una hoja de cálculo son 16,384 columnas por 1'048,576 renglones.	https://support.microsoft.com/en-us/office/insert-or-delete-rows-and-columns-6f40e6e4-85af-45e0-b39d-65dd504a3246?wt.mc_id=otc_excel
Celdas (*Cells*)	Contiene cómo utilizar cortar, copiar y pegar para mover o copiar el contenido de la celda (o copiar contenidos o atributos específicos de las celdas). *Por ejemplo*: copiar el valor resultante de una fórmula sin copiar la fórmula, o copiar solo la fórmula. Cuando se mueve o se copia una celda, Excel mueve o copia la celda (incluidas las fórmulas y sus valores, formatos de celda y comentarios resultantes). Se pueden mover celdas en Excel arrastrando y soltando o usando los comandos cortar y pegar.	https://support.microsoft.com/en-us/office/move-or-copy-cells-and-cell-contents-803d65eb-6a3e-4534-8c6f-ff12d1c4139e?wt.mc_id=otc_excel
Formato (*Formatting*)	Ayuda a formatear números en celdas para cosas, como monedas, porcentajes, decimales, fechas, números de teléfono o números de identificación.	https://support.microsoft.com/en-us/office/available-number-formats-in-excel-0afe8f52-97db-41f1-b972-4b46e9f1e8d2?wt.mc_id=otc_excel
Fórmulas y funciones (*Formulas & functions*)	Empieza a crear fórmulas, a partir del signo de igual (=), para resolver problemas con fórmulas.	https://support.microsoft.com/en-us/office/overview-of-formulas-in-excel-ecfdc708-9162-49e8-b993-c311f47ca173?wt.mc_id=otc_excel https://support.microsoft.com/en-us/office/create-custom-functions-in-excel-2f06c10b-3622-40d6-a1b2-b6748ae8231f

Continúa...

Tema	Contenido	Dirección
Tablas (*Tables*)	Es para crear tablas para organizar y agrupar datos.	https://support.microsoft.com/en-us/office/create-and-format-tables-e81aa349-b006-4f8a-9806-5af9df0ac664?wt.mc_id=otc_excel
Gráficos (*Charts*)	Sirve para crear gráficas de diferentes tipos, como la de barras, que puede ayudar a comparar promedios de grupos.	https://support.microsoft.com/en-us/office/create-a-chart-from-start-to-finish-0baf399e-dd61-4e18-8a73-b3fd5d5680c2?wt.mc_id=otc_excel
Tablas dinámicas (*Pivot tables*)	Son una herramienta para calcular, resumir y analizar datos, que permiten ver comparaciones, patrones y tendencias en los datos.	https://support.microsoft.com/en-us/office/create-a-pivottable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576?wt.mc_id=otc_excel
Compartir y ser coautor (*Share & co-author*)	Esta es una característica para compartir con otras autoras y autores, quienes pueden editar y manipular el documento.	https://support.microsoft.com/en-us/office/share-your-excel-workbook-with-others-8d8a52bb-03c3-4933-ab6c-330aabf1e589?wt.mc_id=otc_excel
Tipos de datos vinculados (*Linked data types*)	Con los tipos de datos vinculados, se puede insertar y trabajar con datos de fuentes de datos en línea confiables. Microsoft se ha asociado con Wolfram para proporcionar datos y hechos sobre numerosos temas que se pueden usar y actualizar, todo sin salir de Excel.	https://support.microsoft.com/en-us/office/convert-text-to-a-linked-data-type-in-excel-preview-7530df24-3e3f-40a2-9cd0-3c31120831bb?wt.mc_id=otc_excel
Conociendo el Power Query (*Get to know Power Query*)	Es para importar datos de la web. Se puede comenzar con Power Query para importar algunos datos. Aunque los videos de esta capacitación se basan en Excel para Microsoft 365, se han agregado instrucciones, como etiquetas de video, si se usa Excel 2016.	https://support.microsoft.com/en-us/office/excel-video-training-9bc05390-e94c-46af-a5b3-d7c22f6990bb
Plantillas en Excel para mostrar ejemplos	Toma un tour (*Take a tour*); Tutorial de fórmulas (*Formula Tutorial*); Crea la primera tabla dinámica (*Make your first pivot table*); y obtén más de las tablas dinámicas (*Get more out of pivot tables*). Estas plantillas en Excel ofrecen la oportunidad de emplear una serie de ejemplos para usar las fórmulas y funciones, así como tablas dinámicas.	https://support.microsoft.com/en-us/office/excel-video-training-9bc05390-e94c-46af-a5b3-d7c22f6990bb

Nota: Asimismo, Microsoft tiene una serie de videos a la venta para aprender a usar Excel 2016: https://www.microsoft.com/en-us/p/learn-to-use-microsoft-excel-2016-guides/9nblggh42x9t?activetab=pivot:overviewtab

Tabla A-2 Algunos análisis de comparaciones de grupos

Análisis para diseños experimentales y no-experimentales: comparaciones de promedios de grupos	Variable independiente (con escala nominal)	Variable dependiente: Intervalo y Razón	Excel 2016	Videos de YouTube (por diversos autores)
Prueba *t* de un solo grupo	Una sola variable independiente.	Una sola variable dependiente.	Sí	https://youtu.be/bOCjW765vBk
Prueba *t* con dos grupos independientes	Una sola variable independiente.	Una sola variable dependiente.	Sí	https://youtu.be/kmww0Eewlp0
Prueba *t* con dos grupos dependientes	La variable independiente es un mismo grupo medido en dos ocasiones.	Una sola variable dependiente.	Sí	https://youtu.be/rhELjzK3ohM
Comparaciones de promedios entre dos o más promedios de grupos	Variable independiente: nominal (con dos o más niveles). Puede tener más de una variable independiente (también llamada factor)	Variable dependiente: Intervalo y Razón	Excel 2016	Videos de YouTube
Análisis de la varianza de un factor (ANOVA de un factor)	Una sola variable independiente.	Una sola variable dependiente.	Sí	https://youtu.be/jwAn7JTMknA
Análisis de la varianza factorial (dos o más factores; ANOVA factorial)	Dos o más variables independientes: *e. g.*, género (mujeres vs. hombres) y trabajo (empleados vs. desempleados).	Una sola variable dependiente.	Sí	https://youtu.be/rbWj6Sp3Jeg
Análisis de la varianza de medidas repetidas	La variable independiente es el mismo grupo medido en dos o más ocasiones. También, pueden ser dos grupos independientes medidos en dos o más ocasiones.	Una sola variable dependiente.	Sí	https://youtu.be/uXv7SDsc_04

Continúa...

Análisis de la covarianza de un factor (ANCOVA)	Una sola variable independiente cuando se usan una o más covariables (una variable medida en una escala continua, como la edad de los estudiantes).	Una sola variable dependiente.	No directamente, pero se puede usar.	https://youtu.be/tMRq1TaOv5s
Comparaciones de varios promedios entre dos o más promedios de grupos	Variable independiente: nominal (con dos o más niveles). Puede tener más de una variable independiente (también llamada factor)	Variable dependiente: Intervalo y Razón	Excel 2016	Videos de YouTube
Análisis múltiple de la varianza (MANOVA)	Una o más variables independientes.	Más de una variable dependiente.	No	https://youtu.be/hfRfOJwA988 https://youtu.be/CBXYxs9pLW8
Análisis múltiple de la covarianza (MANCOVA)	Una o más variables independientes con una o más covariables.	Más de una variable dependiente.	No	https://youtu.be/4ytDLOYq__s
Relaciones entre variables	Variable independiente	Variable dependiente	Excel 2016	Videos de YouTube
Correlación Producto-Momento de Pearson ®	Una variable independiente.	Una variable dependiente.	Sí	https://youtu.be/LI14ftR6EVs
Regresión lineal simple	Una variable independiente.	Una variable dependiente.	Sí	https://youtu.be/tMRq1TaOv5s
Regresión lineal múltiple	Varias variables independientes.	Una variable dependiente.	Sí	https://youtu.be/icipgz8T7dw
Correlación canónica	Más de una variable independiente. Las variables pueden estar medidas en cualquier escala.	Más de una variable dependiente. Las variables pueden estar medidas en cualquier escala.	No	https://youtu.be/p-jTBJcAKcc

Nota: Los videos pueden cambiar con el tiempo en *YouTube*, pero se pueden hacer búsquedas de estos análisis que aparecen tanto en español como en inglés. Como referencias generales a estos análisis de comparaciones y relaciones, se recomienda consultar a Hinkle *et al.* (2003); y para los multivariados, a Tabachnick y Fidell (2019).

Apéndice B. Matemáticas elementales para la estadística

La probabilidad y estadística involucran desde la simple aritmética hasta complejos algoritmos por series de computadoras, que pueden llevar días para analizar datos. Para hacer más accesibles las fórmulas de este libro, se presenta este Apéndice B con aritmética, álgebra y símbolos usados comúnmente en la estadística.

El operador de suma

Este *operador de suma* significa una adición de un grupo (set) de números y se representa con la letra griega sigma: Σ. Antes de comenzar con Σ, una serie de variables puede ser diferenciada al cambiar la *i* (esta indica donde se inicia la suma del set de números), que es un suscrito de *x* (*i. e.*, x_i), por algún número en el suscrito. Un ejemplo es cuando se tienen seis valores en un set, como el siguiente con x_i: *i. e.*, $x_1 = 6$; $x_2 = 5$; $x_3 = 1$; $x_4 = 3$; $x_5 = 4$; y $x_6 = 2$. Este set se puede abreviar de la siguiente manera:

$$\sum_{i=1}^{6} x_i \qquad \text{Ecuación B-1}$$

La Ecuación B-1 se puede leer como la suma de todos los números del set, los cuales comienzan con el número uno ($i = 1$) y terminan con el número seis. Asimismo, la Ecuación B-1 significa: $x_1 + x_2 + x_3 + x_4 + x_5 + x_6$. Da como resultado cuando se sustituyen las variables con números: 6 + 5 + 1 + 3 + 4 + 2 = 21. Por la tanto, se tiene la siguiente equivalencia:

$$\sum_{i=1}^{6} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 6 + 5 + 1 + 3 + 4 + 2 = 21.$$

En la Tabla B-1 se muestra cómo realizar la sumatoria en Excel para obtener el resultado de 21.

Tabla B-1 Suma

	A
1	6
2	5
3	1
4	3
5	4
6	2
7	=SUM(A1:A7)

De una manera general, la ecuación para generalizar se escribe de esta manera: Ecuación B-2, donde i^{o} (*i. e.*, número primero del set) indica que la suma comienza con este valor y termina con N^{o}, que es el último número del set.

$$\sum_{i=1}^{N} x_i$$ Ecuación B-2

Reglas del operador de suma

Estas son las reglas para el operador de suma, según Hinkle *et al.* (2003, p. 3):

> *Aplicando el operador de suma a la Regla 1.* Los productos resultantes por multiplicar un set de números por una constante (*C*; un valor que no cambia, como 1, 2, 3,…*n*, entre otros; *n* indica el valor más grande del set) son iguales a multiplicar la constante por la suma de los números. Para la Regla 1, las Ecuaciones B-3 y B-4 son equivalentes.

$$\sum_{i=1}^{N} C x_i = C \sum_{i=1}^{N} x_i$$ Ecuaciones B-3 y B-4

Ejemplificando y sustituyendo la Ecuación B-3 con el set de números anterior y con $C = 2$:

$$\sum_{i=1}^{6} (2)\, x_i = (2)1 + (2)2 + (2)3 + (2)4 + (2)5 + (2)6 = 42.$$

Asimismo, sustituyendo la Ecuación B-4 con el set de números anterior y con $C = 2$:

$$2 \sum_{i=1}^{N} x_i = 2\,(1 + 2 + 3 + 4 + 5 + 6) = 42.$$

Entonces, se demostró que la Ecuación B-3 es igual a la Ecuación B-4, porque se obtiene 42 como resultado en ambas ecuaciones.

Aplicando el operador de suma a la Regla 2. La suma de una serie de puntajes constantes es igual a multiplicar *N* veces por la constante.

$$\sum_{i=1}^{N} C = N\,C \qquad \text{Ecuaciones B-5 y B-6}$$

Ejemplificando y sustituyendo la Ecuación B-5 con este set de números: $x_1 = 2$; $x_2 = 2$; $x_3 = 2$; y $x_4 = 2$:

$$\sum_{i=1}^{4} C = 2 + 2 + 2 + 2 = 8 \qquad \text{Ecuación B-5}$$

Asimismo, sustituyendo B-6 con $N = 4$ y con $C = 2$:

$4(2) = 8$ Ecuación B-6

Por lo tanto, se demuestra que la Ecuación B-5 = Ecuación B-6.

Aplicando el operador de suma a la Regla 3. En la Tabla B-2 se muestra cómo se organizan tradicionalmente los datos: los

renglones representan individuos, observaciones u objetos y las columnas representan variables (*x*, *y*, *z*, entre otras).

Tabla B-2 Renglones y columnas

	A	B	C	D	E	F
1	Renglón/ Columna	Columna 1	Columna 2	Columna 3	Columna 4	
2	Renglón 1	3	9	7	5	=SUM(B2:E2)
3	Renglón 2	5	2	7	7	=SUM(B3:E3)
4	Renglón 3	8	6	1	2	=SUM(B4:E4)
5	Sumas	=SUM(B2:B4)	=SUM(C2:C4)	=SUM(D2:D4)	=SUM(E2:E4)	

Nota: A esta tabla se le considera una matriz de 3 x 4. El 3 es el número de renglones (participantes, observaciones u objetos) y 4 es el número de columnas (variables). La columna F es la suma de cada renglón y no se le considera para calcular el tamaño de la matriz. Asimismo, el renglón 5 es una suma por variable, pero tampoco se le considera para calcular el tamaño de la matriz.

Para sumar los resultados por variable, se especifican las coordenadas/celdas de la columna: *e. g.*, =SUM(B2:B4). Para las sumas de las personas, observaciones u objetos, se usan las coordenadas/celdas de los renglones: *e. g.*, =SUM(B2:E2). Lo mismo (*i. e.*, utilizar columnas, renglones o ambos) se podría hacer con otras estadísticas, como las de tendencia central o de dispersión, entre otras.

La Regla 3 dice: La suma de dos o más variables por cada renglón es igual a la suma de todas las columnas.

$$\sum_{i=1}^{N} (x_i + y_i) = \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i$$

Ecuaciones B-7 y B-8

Los datos muestran en la Tabla B-3 dónde aparecen dos calificaciones en las columnas (variables) y tres estudiantes en los renglones.

Tabla B-3 Datos para la Regla 3

	A	B	C	D
1	Estudiantes	Calificación 1	Calificación 2	Suma
2	José	7	8	15
3	Pablo	6	7	13
4	Jesús	9	10	19
5	Suma	22	25	47

Desarrollando las Ecuaciones B-7 y B-8:

$$\sum_{i=1}^{3} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + (x_3 + y_3) \qquad \text{Ecuación B-7}$$

Sustituyendo:

$= (7 + 8) + (6 + 7) + (9 + 10) = 15 + 13 + 19 = 47;$

$$\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i = (x_1 + x_2 + x_3) + (y_1 + y_2 + y_3) \qquad \text{Ecuación B-8}$$

Sustituyendo:

$= (7 + 6 + 9) + (8 + 7 + 10) = 22 + 25 = 47.$

Por lo tanto, se demuestra que las Ecuaciones B-7 y B-8 son equivalentes.

Valor absoluto

Es otro concepto habitual en estadística que indica el valor de un número (sin importar su valor algebraico) y se representa con dos líneas paralelas y perpendiculares a un renglón: $|x|$. Un ejemplo es el valor absoluto de –3 denotado por: $|-3|$, que resulta 3. En resumen, un valor absoluto es: $|-3| = |+3| = 3$.

Tabla B-4 Valor absoluto

	A	B
1	-3	=ABS(A1)
2	4	=ABS(A2)

Nota: Se obtienen 3 y 4, respectivamente.

Operaciones aritméticas e indicadores especiales

Existen dos operaciones aritméticas recurrentes en la estadística: el exponente y la raíz cuadrada. El exponente más común es el cuadrado: x^2. La *x* sería la base con algún número de un set y el 2 sería el exponente. Lo anterior significa: la multiplicación de *x* por *x*. Igualmente, se puede decir que la *x* fue elevada a la segunda potencia. Otro ejemplo con base numérica es: 3^3: 3 x 3 x 3 = 27. En la Tabla B-5 se muestra cómo usar los exponentes en Excel.

Tabla B-5 Exponente

	A	B
1	3	=A1^3

Nota: Se obtiene 27.

La otra operación aritmética de uso común en la estadística es la raíz cuadrada $= x^{1/2}$), cuya definición es un número real que, al ser multiplicado por sí mismo, dará el número adentro del símbolo de la raíz cuadrada. Por ejemplo, la raíz cuadrada de 9 es 3 y la demostración es elevar al cuadrado el 3: $= 3$ y 3 x 3 = 9 (véase: Tabla B-6).

Tabla B-6 Raíz cuadrada y exponente

	A	B	C	D
1	9	=SQRT(A1)	3	=C1^2

Nota: La raíz cuadrada de 9 es 3 y el exponente al cuadrado de 3 es 9.

El orden de las operaciones

La existencia de paréntesis indica qué operaciones se realizan primero: *e. g.*, (6 + 8) / 2 resulta en: 6 + 8 = 14 y luego: 14 / 2 = 7 (resultado final; véase: Tabla B-7). Por otro lado, si no existiera el paréntesis, ¿qué se hubiera hecho primero? Por esta razón, hay ciertas reglas para saber qué operaciones efectuar primero.

Tabla B-7 Orden de las operaciones: suma y división

	A	B
1	6	=(A1+A2)/A3
2	8	
3	2	

Nota: El resultado es 7.

1. **Primera regla:** Se deben de realizar primero las operaciones que aparecen entre paréntesis, llaves y corchetes, del interior hacia el exterior. Por ejemplo: {[(5 – 1) ÷ 2] + 1}x 3 = 9; esto es, primero (5 – 1) = 4; segundo, [4 ÷ 2] = 2; tercero, {2 + 1} = 3; y, finalmente, 3 x 3 = 9 (véase: Tabla B-8).

Tabla B-8 Suma, división y multiplicación

	A	B
1	5	=((A1+A2)/A3)*A4
2	1	
3	2	
4	3	

Nota: El resultado es 9.

2. **Segunda regla:** Se realizan las operaciones con exponentes y raíces cuadradas primero cuando no se encuentran fuera de los paréntesis, llaves o corchetes. Por ejemplo: $2^2 \div 2 = (2 \times 2) \div 2 = 2$ (véase: Tabla B-9).

Tabla B-9 Exponente y división

	A	B
1	2	=A1^2/A1

Nota: El resultado es 2.

3. **Tercera regla:** Las multiplicaciones y divisiones van primero que las sumas y restas cuando no hay paréntesis. Por ejemplo: 6 + 8 ÷ 2 = 6 + 4 = 10 (véase: Tabla B-10).

Tabla B-10 Suma y división

	A	B
1	6	=A1+A2/A3
2	8	
3	2	

Nota: El resultado es 10.

Redondeo de números

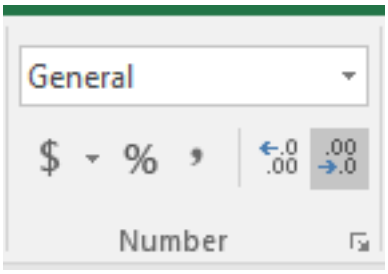
En muchas ocasiones, las operaciones de los análisis estadísticos dan resultados con varios decimales: *e. g.*, 7.59812567369712389468925 (véase: Tabla B-11).

Tabla B-11 Redondeo

	A	B
1	7.59812567369712389468925	7.60

Excel contiene una función que ayuda a redondear números con decimales, como el anterior (véase: Figura B-1). Solo hay que seleccionar la celda deseada y oprimir el ícono de la derecha (*i. e.*, comienza con .00).

Figura B-1 Redondeo



Hinkle *et al.* (2003) dijeron que después de cada operación aritmética que produzca una fracción, hay que llevar el decimal a tres lugares (*e. g.*, 2.312) y redondearlo a dos lugares más cuando haya decimales en los datos originales (*e. g.*, si el original fue 2.312 y después de realizar la operación queda en: 2.31241).

Apéndice C. Otras formas de los datos

Curvas

Puede pasar que los datos no tengan una *relación lineal*[1] y que no haya manera de transformarlos a modelos lineales. Lo anterior, se puede advertir cuando se grafican dos variables en un plano cartesiano y no se forma una línea. Para modelos no-lineales, como algunos de los que se describen en este Apéndice C, se recomienda consultar a Garson (2012) y Jones (2019). Asimismo, para una introducción breve a la transformación[2] de datos, se recomienda consultar a Osborne (2008). Otra recomendación para tener una base fuerte en matemáticas es el primer capítulo de Larson y Edwards (2010).

En la Figura C-1 se muestra que la variable independiente *x* tiene una relación con la variable dependiente *y*, porque los puntos de los *datos emparejados*[3] forman un patrón, pero no es lineal (*i. e.*, es una curva), y se representa con la fórmula/función: $f(x) =$ (en la cual los valores de *y* dependen de los de *x*). Más específicamente: la variable dependiente (*y*) es igual a la raíz cuadrada de cada valor de la variable independiente (*x*); de hecho, la relación es una curva. En Excel, la variable dependiente se creó a partir de sacar la raíz cuadrada a *x*, como se muestra en la fórmula que se aplicaría en Excel a una sola celda: solo habría que recorrer el cursor hacia abajo por la columna B para obtener todos los valores de *y*: = SQRT(Celda$_x$). Por otro lado, cabe la posibilidad de que los datos se encuentren en esta forma de manera natural y si no se pueden transformar en un modelo lineal, se podría usar alguno no-lineal para ver hasta dónde se relacionan las variables. Esta relación se podría ejemplificar al ser *x* el número de horas inver-

1 Una relación lineal sucede cuando los datos de dos variables forman un patrón, como los del Capítulo 3 del presente texto: Figuras 3.5 y 3.6.

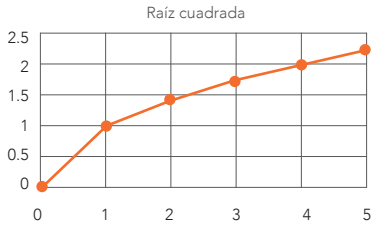
2 Una transformación es una modificación matemática de una variable para alcanzar una meta en particular (*e. g.*, normalidad para incrementar la posibilidad de interpretar). Existe una variedad casi infinita de posibles transformaciones de los datos: desde adicionar una constante hasta multiplicar, elevar al cuadrado, elevar a un poder; conversión a una escala logarítmica; invertir y reflejar; tomar la raíz cuadrada; o aplicar transformaciones trigonométricas, como la transformación del seno (Osborne, 2008, p. 197).

3 Los datos emparejados significan que por cada renglón que puede representar a una/un estudiante existe una calificación por cada una de las dos columnas.

tidas en hacer una tarea y *y*, el número de puntos que se obtendrían: *e. g.*, por cero horas invertidas, se obtienen cero puntos; por 5 horas invertidas, se obtienen 2.24 puntos.

Figura C-1 Datos no-lineales: raíz cuadrada

	A	B
1	*x*	*y*
2	0	0
3	1	1
4	2	1.41
5	3	1.73
6	4	2
7	5	2.24

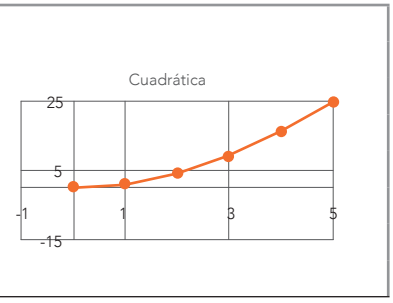


Nota: Esta curva es cóncava del origen al punto (5, 2.24) y se va incrementando.

En la Figura C-2 la relación cuadrática [$f(x) = x^2$] entre dos variables involucra la dependiente, que es el cuadrado de la independiente: *e. g.*, $5^2 = 5 \times 5 = 25$. En otras palabras, la fórmula de Excel: =A^2. Un ejemplo de esta relación podría ser el estudiar para un examen de admisión para ingeniería industrial por parte de cinco aspirantes: *x* = Número de horas de estudio y *y* = Número de aciertos en un examen. Esto se traduce en: cero horas de estudio es igual a cero aciertos y 5 horas de estudio se asocian con 25 aciertos.

Figura C-2 Datos no-lineales: cuadrática

	A	B
1	*x*	*y*
2	0	0
3	1	1
4	2	4
5	3	9
6	4	16
7	5	25

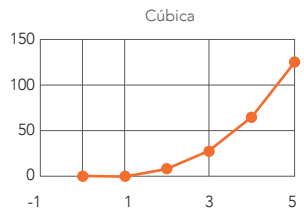


Una relación cúbica [$f(x) = x^3$] aparece en la Figura C-3, que representa una relación entre una variable independiente (horas de es-

tudio: *x*) y una dependiente (puntaje: *y*). Siguiendo el ejemplo de la Figura C-2 de los estudiantes de ingeniería, 5 horas de estudio se traducen en 125 puntos (véase: Figura C-3): *i. e.*, $5^3 = 5 \times 5 \times 5 = 125$. En Excel es: =A2^3.

Figura C-3 Datos no-lineales: cúbica

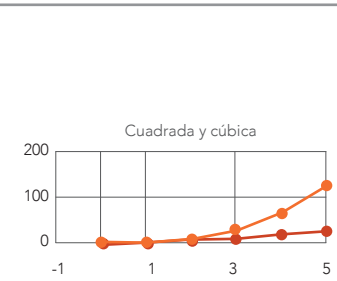
	A	B
1	*x*	*y*
2	0	0
3	1	1
4	2	8
5	3	27
6	4	64
7	5	125



Estas dos curvas (Figuras C-2 y C-3) pueden ser graficadas en Excel en la misma figura (Figura C-4). No solo dos curvas o líneas pueden ser representadas, sino una multitud de ellas. Probablemente, una de las limitantes a esta representación de múltiples datos sería el que fueran visibles e interpretables.

Figura C-4 Dos curvas a la vez

	A	B	C
1	*x*	*y*	*z*
2	0	0	0
3	1	1	1
4	2	4	8
5	3	9	27
6	4	16	64
7	5	25	125



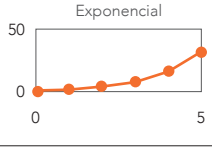
Nota: Estas dos curvas son convexas del origen al punto (5, 25) y (5, 125), respectivamente, y se van incrementando.

En la Figura C-5 se muestra una relación conocida como exponencial [$f(x) = a^x$; donde $a = 2$ para el siguiente ejemplo]. La base de esta relación es una constante (representada por la letra *a*): *e. g.*, 2, 3, 4,…, etcétera. Esta constante se llama base y el número al que se ele-

va la base, se llama exponente: a^x. En este caso, se tomó el número 2 como base y se elevó a la *x* (0, 1, 2, 3, 4, 5) mediante la fórmula: = 2^A2. Un número elevado a la potencia de 0 resulta en 1: $2^0 = 1$. Un ejemplo podría ser una profesora que está ofreciendo un curso de estadística en Coursera.org. La *x* representa el número de correos promocionales que reciben los usuarios de Coursera.org en su correo electrónico personal y la *y* es el número de usuarias o usuarios que se inscriben en el curso de estadística de la profesora. Es decir, cuando se envían cero correos, una usuaria o un usuario se inscribe y cuando se envía un solo correo, se inscriben dos usuarias o usuarios, y así sucesivamente.

Figura C-5 Una relación exponencial

	A	B
1	*x*	*y*
2	0	1
3	1	2
4	2	4
5	3	8
6	4	16
7	5	32



Nota: La curva exponencial es convexa del origen al punto (5, 32) y se va incrementando.

En la Figura C-6 se representa una relación logarítmica de base 10 entre la *x* y la *y*: *i. e.*, $f(x) = \text{Log}_b x$. Se puede explicar un logaritmo al explicar una potencia: *i. e.*, $\text{Log}_b a = c \Leftrightarrow b^c = a$; la flecha con dos puntas indica que uno se puede convertir en otro.

Para el logaritmo ($\text{Log}_b a$):

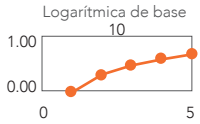
b = Base (puede ser cualquier número, pero en este ejemplo se usa el 10)
a = Argumento o antilogaritmo
c = Logaritmo
Para el exponencial ($b^c = a$)
b = Base (puede ser cualquier número)
c = Exponente
a = Potencia

Un ejemplo es cuando se tiene el siguiente exponencial: $10^2 = 100$ (diez es elevado al cuadrado y resulta en 100). Cuando lo anterior su vuelve un logaritmo: $\text{Log}_{10} 100 = 2$, se responde a la pregunta: ¿a qué potencia hay que elevar la base de 10 para que resulte 100? La respuesta es 2, que es el logaritmo de la base 10 para 100.

La Figura C-6 contiene las variables con una relación logarítmica [$f(x) = \text{Log}_b x$; donde a = 10] con la fórmula de Excel: = LOG10(A1).

Figura C-6 Una relación logarítmica

	A	B
1	*x*	*y*
2	0	No existe
3	1	0
4	2	0.30
5	3	0.48
6	4	0.60
7	5	0.70



Nota: La curva logarítmica es cóncava del punto (1, 0) al (5, .80) y se va incrementando.

Un ejemplo de relación logarítmica podría ser que una escuela va a dar tutorías a 6 estudiantes para que mejoren su idioma inglés, por lo que se les aplica un examen de diagnóstico para ver dónde se ubican. Cada alumna/alumno toma diferentes horas de tutorías (*x*; 0, 1, 2, 3, 4, 5). Una/un estudiante que no asistió no tiene puntajes. La variable dependiente es la *y*, que representa el puntaje de *diferencia* entre el puntaje de diagnóstico y la prueba después de haber tomado las tutorías (Tabla C-1).

Tabla C-1 Datos del ejemplo de mejora por tutorías

	A	*B*	C	D	E
1	Estudiante	*x* (Tutorías)	Prueba después de las tutorías	Examen diagnóstico	*y*
2	I	0	NA	NA	No existe
3	II	1	7.41	7.41	0
4	III	2	7.80	7.50	0.30
5	IV	3	7.48	7.00	0.48
6	V	4	7.80	7.20	0.60
7	VI	5	8.00	7.30	0.70

Nota: La columna E también se puede obtener de la diferencia entre las columnas C y D: $=C_x-D_x$.

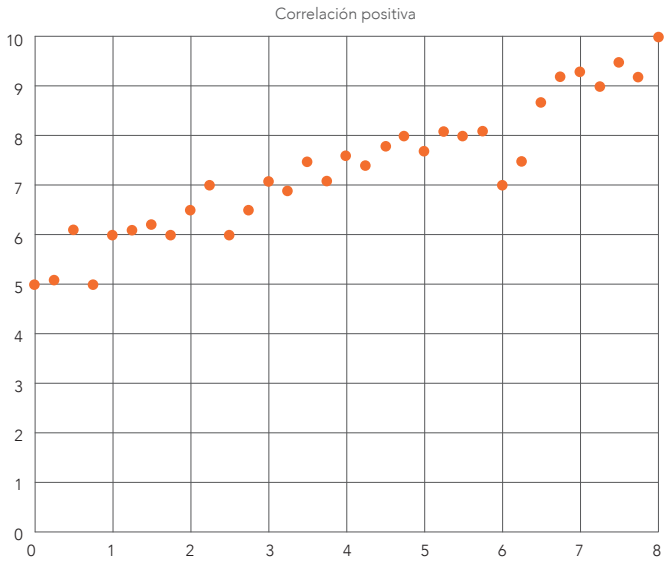
Esto es, la/el alumno (II) que fue a una hora de tutoría no mejoró (0), porque no hay diferencia entre su puntaje de 7.41 en el examen diagnóstico y el puntaje de 7.41 en la prueba después de las tutorías.

Relaciones lineales no-perfectas

Además de las curvas y las relaciones lineales en las que todos los puntos se alinean, existen las relaciones lineales no-perfectas donde no todos los puntos están formados sobre la misma línea; sin embargo, sí forman un patrón (Figura C-7). La *x* son las horas invertidas en estudiar y la *y* es la calificación en un examen. Por ejemplo, la/el estudiante que invirtió 0 horas en estudiar obtuvo una calificación de 5. Por otro lado, la/el alumno que invirtió 8 horas en estudiar obtuvo un 10, y las y los demás estudiantes quedaron en medio de estos dos extremos, tanto en las horas de estudio como en las calificaciones.

Figura C-7 Relaciones lineales no-perfectas: positiva

	A	B
1	*x*	*y*
2	0	5
3	0.25	5.1
4	0.5	6.1
5	0.75	5
6	1	6
7	1.25	6.1
8	1.5	6.2
9	1.75	6
10	2	6.5
11	2.25	7
12	2.5	6
13	2.75	6.5
14	3	7.1
15	3.25	6.9
16	3.5	7.5
17	3.75	7.1
18	4	7.6
19	4.25	7.4
20	4.5	7.8
21	4.75	8
22	5	7.7
23	5.25	8.1
24	5.5	8
25	5.75	8.1
26	6	7
27	6.25	7.5
28	6.5	8.7
29	6.75	9.2
30	7	9.3
31	7.25	9
32	7.5	9.5
33	7.75	9.2
34	8	10



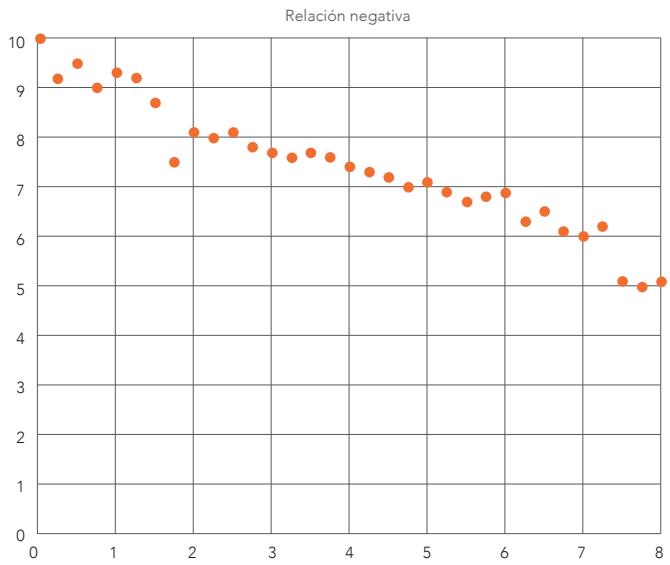
Los datos se pueden ingresar en Excel directamente, luego se seleccionan y se oprime el ícono de gráficas (como en la Figura 3.2). Para usar una función (*fx*), que refleje la relación entre estas variables, se tendría que emplear una regresión simple. Está más allá de los propósitos de este libro el abordar estadísticas inferenciales, pero se recomienda consultar a Tabachnick y Fidell (2019).

Otro ejemplo de una relación lineal no-perfecta es cuando se tienen los datos distribuidos (como en la Figura C-8). En este caso, es una relación negativa, porque conforme *x* se incrementa, *y* disminuye. Esto es en pares ordenados: comienza con (0, 10) y termina con (8, 5.1). Es decir, la *x* fue de 0 a 8 (aumentó) y, por otro lado, *y* fue de 10 a 5.1 (disminuyó).

Un ejemplo de lo anterior puede ser que la *x* represente las horas de ocio por día de un grupo de estudiantes universitarios y la *y*, la calificación final del semestre. En pocas palabras, la/el alumno que tuvo 0 horas diarias de ocio durante el semestre obtuvo un 10 de calificación. Pero, al contrario, aquella/aquel que tuvo 8 horas diarias de ocio consiguió un 5.1 de calificación. En este ejemplo, como en el anterior, la función se podría encontrar mediante una regresión simple.

Figura C-8 Relaciones lineales no-perfectas: positiva

	A	B
1	x	y
2	0	10
3	0.25	9.2
4	0.5	9.5
5	0.75	9
6	1	9.3
7	1.25	9.2
8	1.5	8.7
9	1.75	7.5
10	2	8.1
11	2.25	8
12	2.5	8.1
13	2.75	7.8
14	3	7.7
15	3.25	7.6
16	3.5	7.7
17	3.75	7.6
18	4	7.4
19	4.25	7.3
20	4.5	7.2
21	4.75	7
22	5	7.1
23	5.25	6.9
24	5.5	6.7
25	5.75	6.8
26	6	6.9
27	6.25	6.3
28	6.5	6.5
29	6.75	6.1
30	7	6
31	7.25	6.2
32	7.5	5.1
33	7.75	5
34	8	5.1



Apéndice D. Distribución normal estandarizada/estándar

En la Tabla D-1 se muestra cómo usar las fórmulas del promedio [=AVERAGE(Celda$_1$:Celda$_n$)] y de la desviación estándar [=STDEV.S(Celda$_1$:Celda$_n$)] para obtener de la columna A los valores estandarizados en la columna B mediante la fórmula matemática $(x_i - \bar{x}) / SD$. De los valores *z* de la columna B, se aplica la función de Excel: =NORM.S.DIST(Celda$_x$,FALSE), que devuelve la distribución normal estandarizada. En otras palabras: esta función muestra los valores de la variable dependiente *y* (columna C), que dependen de la variable *x* (columna B). Para calcular todos los valores *z* de la columna B y los de la variable dependiente (columna C) solo hay que utilizar primero la fórmula y la función de la Tabla D-1, y deslizar el cursor para obtener todos estos valores.

Tabla D-1 Fórmulas para la distribución normal estandarizada

	A	B	C
	Calcificación	Valor *z*	y
1	40	=(A1-A$62)/A$63	=NORM.S.DIST(B1,FALSE)
…	…	…	…
61	100	=(A61-A$62)/A$63	=NORM.S.DIST(B61,FALSE)
62	=AVERAGE(A1:A61)		
63	=STDEV.S(A1:A61)		

Nota: Véase la Tabla 6.14 del Capítulo 6, que muestra cómo calcular un valor *z* con una función de Excel: =STANDARDIZE(Celda$_1$:Celda$_n$,promedio,*SD*).

En la Tabla D-2 se muestran los resultados de la Tabla D-1: *i. e.*, los valores estandarizados en la columna B y los de la distribución normal en la columna C.

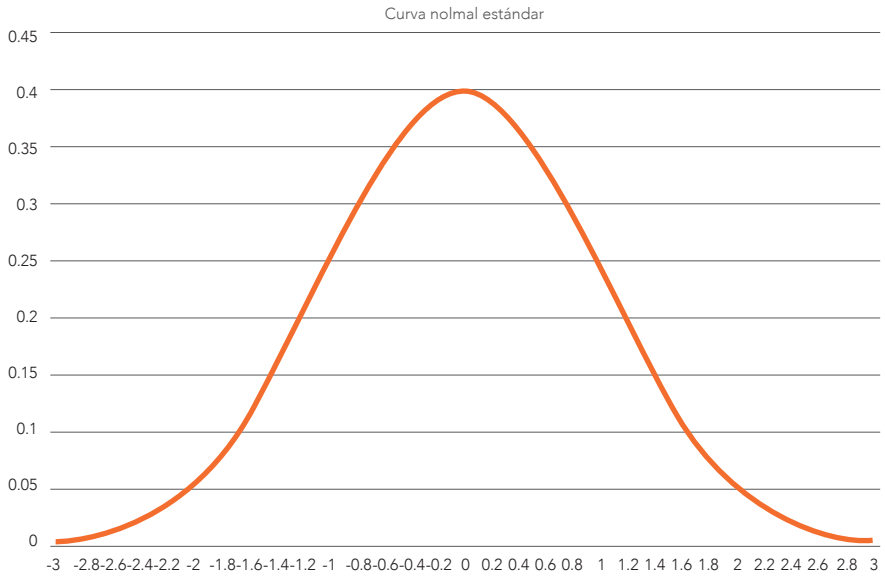
Tabla D-2 Resultados de los valores *z* y distribución normal estandarizada/ estándar

	A	B	C		A	B	C
1	40	-3	0.00443185	33	72	-0.2	0.39104269
2	41	-2.9	0.00595253	34	73	0.3	0.38138782
3	42	-2.8	0.00791545	35	74	0.4	0.36827014
4	43	-2.7	0.01042093	36	75	0.5	0.35206533
5	44	-2.6	0.01358297	37	76	0.6	0.3332246
6	45	-2.5	0.0175283	38	77	0.7	0.31225393
7	46	-2.4	0.02239453	39	78	0.8	0.28969155
8	47	-2.3	0.02832704	40	79	0.9	0.26608525
9	48	-2.2	0.03547459	41	80	1	0.24197072
10	49	-2.1	0.0439836	42	81	1.1	0.21785218
11	50	-2	0.05399097	43	82	1.2	0.19418605
12	51	-1.9	0.06561581	44	83	1.3	0.17136859
13	52	-1.8	0.07895016	45	84	1.4	0.14972747
14	53	-1.7	0.09404908	46	85	1.5	0.1295176
15	54	-1.6	0.11092083	47	86	1.6	0.11092083
16	55	-1.5	0.1295176	48	87	1.7	0.09404908
17	56	-1.4	0.14972747	49	88	1.8	0.07895016
18	57	-1.3	0.17136859	50	89	1.9	0.06561581
19	58	-1.2	0.19418605	51	90	2	0.05399097
20	59	-1.1	0.21785218	52	91	2.1	0.0439836
21	60	-1	0.24197072	53	92	2.2	0.03547459
22	61	-0.9	0.26608525	54	93	2.3	0.02832704
23	62	-0.8	0.28969155	55	94	2.4	0.02239453
24	63	-0.7	0.31225393	56	95	2.5	0.0175283
25	64	-0.6	0.3332246	57	96	2.6	0.01358297
26	65	-0.5	0.35206533	58	97	2.7	0.01042093
27	66	-0.4	0.36827014	59	98	2.8	0.00791545
28	67	-0.3	0.38138782	60	99	2.9	0.00595253
29	68	-0.2	0.39104269	61	100	3	0.00443185
30	69	-0.1	0.39695255	62	70	0	
31	70	0	0.39894228	63	10	1	
32	71	0.1	0.39695255				
			Continúa…				

Nota: El promedio de los datos originales fue 70 y la *SD* fue 10; para los datos transformados a valores *z*, el promedio = 0 y la *SD* = 1.

En la Figura D-1 se muestra la gráfica de la distribución normal estandarizada, a partir de los datos de la columna B en la horizontal y la columna C en la vertical. Para obtener la Figura D-1, se recomienda ver la Figura 3.2 del Capítulo 3.

Figura D-1 Distribución normal estándar de un conjunto de datos



Apéndice E. Curtosis y asimetría en relación con la curva normal

Los cuatro momentos

Para ampliar el panorama de la curtosis y la asimetría, Hopkins y Weeks (1990) los explicaron en el contexto de otras variables, como el promedio y la desviación estándar:

> Las características más importantes de las frecuencias de distribución son resumidas por índices estadísticos relacionados a los primeros cuatro momentos de una distribución. El primer momento es el promedio (el centro de gravedad – la medida más eficiente de las medidas de tendencia central); el segundo momento es la varianza (el momento de inercia), una medida de variabilidad; el tercer momento representa la asimetría -desviación de la simetría; el cuarto momento refleja la curtosis […] representa las desviaciones de la curva normal. (p. 717)

Curtosis

Fernández, Córdoba y Cordero (2002) declararon que la curtosis sirve para medir el grado en el que una distribución, se aleja de la distribución normal. *En corto*: cuando más se alejen los coeficientes de curtosis del cero, como consecuencia se alejarán los datos de una distribución normal. Por un lado, esto se puede advertir visualmente si la distribución tiende a ser plana (Coeficiente de Curtosis < 0: platicúrtica; véase: Tabla 3.14 y Figura 3.12 F): es cuando los valores no se concentran altamente en el centro de la distribución y están esparcidos hacia los extremos. Por otro lado, cuando los valores se concentran en la zona del centro y la distribución se vuelve picuda (Coeficiente de Curtosis > 0: leptocúrtica; véase: Tabla 3.13 y Figura 3.12 E). Cuando es una distribución normal se llama mesocúrtica (véase: Tabla 3.10 y Figura 3.12 B). En la Ecuación E-1 se muestra la forma en que Excel calcula la curtosis, aunque esta no es la única ecuación porque existen otras (véase: Barrantes-Aguilar, 2019, para más maneras):

$$\text{Curtosis} = \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)} X \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{SD}\right)^4\right] - \frac{3(n-1)^2}{(n-2)(n-3)} \qquad \text{Ecuación E-1}$$

Asimetría

La asimetría o sesgo describe el grado en que los datos no se distribuyen simétricamente (Hopkins, & Weeks, 1990). Entre más sesgados estén los datos, el Coeficiente de Asimetría se aleja mayormente del cero. Una distribución normal tiene un Coeficiente de Asimetría de 0: Promedio = Mediana = Moda. Cuando una distribución está positivamente sesgada (asimetría > 0; véase: Tabla 3.11 y Figura 3.12 C): Moda < Mediana < Promedio. Cuando una distribución está negativamente sesgada (asimetría < 0; véase: Tabla 3.12 y Figura 3.12 D): Promedio < Mediana < Moda (véase: Ponce-Renova, 2019, para más información al respecto). Las distribuciones se suelen sesgar por valores atípicos.[1]

Fórmula de la Asimetría/Sesgo

$$\text{Asimetría} = \left[\frac{n}{(n-1)(n-2)}\ X\ \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{SD}\right)^3\right] \qquad \text{Ecuación E-2}$$

Donde:
$\sum$ = Suma
n = Número de personas, observaciones u objetos/frecuencia
x_i = Cada uno de los valores del set
$\bar{x}$ = Promedio de la muestra
SD = Desviación estándar de la muestra

La falta de normalidad en las distribuciones de los datos raramente tiene una consecuencia práctica seria en la precisión de análisis inferenciales acerca de los promedios de las poblaciones (Hopkins, & Weeks, 1990). En contraparte, estos autores recomiendan reportar los coeficientes de curtosis y asimetría en las investigaciones, así como poner atención a los supuestos de distribución normal que tienen algunos análisis estadísticos.

1 Se les conoce como *outliers* o anomalías. Son observaciones con una combinación de características únicas identificables como distintivamente diferentes de lo que se considera normal (Hair *et al.*, 2019, p. 85). Usualmente se considera como un valor atípico cuando se tiene una distribución normal estándar (Apéndice D; Figura D-1) cuando una observación cobra un valor de más de | 3 |. Esto significa que la observación se separó del promedio de 0 en 3 *SD* o más. Un ejemplo es cuando una/un estudiante tiene un coeficiente intelectual de 145 puntos con una *SD* = 15 puntos cuando el promedio = 100. Convertido a valor *z*: (145-100) / 15 = 3, que es una observación atípica. Como conclusión, esta o este alumno tiene un atributo extraordinario acerca de su inteligencia.

Apéndice F. Percentil y rango de percentil

Otra manera de encontrar posiciones de percentiles

Universo de Fórmulas (2021) muestra una fórmula para calcular la posición de un valor, dado un percentil estipulado y el tamaño de la muestra:

$$\text{Posición del percentil} = X_{[(n+1)\,Xi]\,/\,100}$$

Donde:

X = Número de lugares comenzando con un conjunto de datos ordenados de manera ascendente

n = Tamaño de la muestra o frecuencia total

i = Percentil que está dentro de este rango: 0 a 100

En la columna C (Tabla F-1) se muestra cómo escribir la fórmula en Excel, donde n = 20 e i = Percentil deseado, con las constantes de 1 y 100.

Tabla F-1 Posición de percentil

	A	B	C	"C"
1	Set	Percentil	Fórmula	Fórmula
2	70	25	=(20+1)*25/100	5.25
3	71	50	=(20+1)*50/100	10.50
4	72	75	=(20+1)*75/100	15.75
5	73			
6	74			
7	75			
8	76			
9	77			
10	78			
11	79			
12	80			
13	81			
14	82			
15	83			
16	84			

Continúa...

	A	B	C	"C"
17	85			
18	86			
19	87			
20	88			
21	89			

La interpretación de la Tabla F-1 para el P_{25}, es que ocupa el lugar 5.25 en el set ordenado de menor a mayor: *i. e.*, existen 5 números por debajo de este *P*. De la misma manera, el P_{50} (10.50) y P_{75} (15.75) señalan los lugares que ocupan estos percentiles.

Usos de percentiles

Los percentiles son usados como resultados de las y los aspirantes a posgrado cuando se toman algunas pruebas estandarizadas, como el *Graduate Record Examinations* (GRE; véase: Tabla F-2 con la información de este examen en el periodo 2013-2016). Con estos percentiles, se le indica al aspirante el lugar respecto al resto. Por ejemplo, un puntaje de 151 coloca a una/un aspirante en el P_{52} y P_{43} para la parte verbal y cuantitativa, respectivamente. En otras palabras, por debajo del P_{52} hubo un 52% y por debajo del P_{43}, un 43% (*cf.* Hinkle *et al.*, 2003). Además, un mismo puntaje en estas dos áreas da como resultado un percentil diferente. Esto indica que el percentil es un lugar relativo a cierto grupo y, por ejemplo, cuando se cambia el grupo de la sección verbal a la cuantitativa, el percentil cambia. Esto no es necesariamente así (cambio de percentil), pero cabe la posibilidad, porque el percentil depende del grupo dentro del cual se hace la comparación.

Tabla F-2 Percentiles del GRE de 2013 a 2016

	A	B	C	D	E		"A"	"D"	"E"
1	Puntaje	*P* verbal	*P* cuantitativo	Diferencia entre *P* verbal	Diferencia entre *P* cuantitativo	1	Puntaje	Diferencia entre *P* verbal	Diferencia entre *P* cuantitativo
2	130	-	-	=B3-B2	=C3-C2	2	130	-	-
3	**131**	1	-	=B4-B3	=C4-C3	3	**131**	0	-
4	132	1	-	=B5-B4	=C5-C4	4	132	1	-
5	133	2	1	=B6-B5	=C6-C5	5	133	0	0
6	134	2	1	=B7-B6	=C7-C6	6	134	1	1
7	135	3	2	=B8-B7	=C8-C7	7	135	1	0
8	136	4	2	=B9-B8	=C9-C8	8	136	2	1
9	137	6	3	=B10-B9	=C10-C9	9	137	2	1
10	138	8	4	=B11-B10	=C11-C10	10	138	1	2
11	139	9	6	=B12-B11	=C12-C11	11	139	3	2
12	140	12	8	=B13-B12	=C13-C12	12	140	3	2
13	141	15	10	=B14-B13	=C14-C13	13	141	2	2
14	142	17	12	=B15-B14	=C15-C14	14	142	3	2
15	143	20	14	=B16-B15	=C16-C15	15	143	4	3
16	144	24	17	=B17-B16	=C17-C16	16	144	3	3
17	145	27	20	=B18-B17	=C18-C17	17	145	4	4
18	146	31	24	=B19-B18	=C19-C18	18	146	4	3
19	147	35	27	=B20-B19	=C20-C19	19	147	4	3
20	148	39	30	=B21-B20	=C21-C20	20	148	4	5
21	149	43	35	=B22-B21	=C22-C21	21	149	5	3
22	150	48	38	=B23-B22	=C23-C22	22	150	4	5
23	**151**	**52**	**43**	=B24-B23	=C24-C23	23	**151**	4	4
24	152	56	47	=B25-B24	=C25-C24	24	152	5	4
25	153	61	51	=B26-B25	=C26-C25	25	153	4	4
26	154	65	55	=B27-B26	=C27-C26	26	154	4	4
27	155	69	59	=B28-B27	=C28-C27	27	155	4	3
28	156	73	62	=B29-B28	=C29-C28	28	156	3	4
29	157	76	66	=B30-B29	=C30-C29	29	157	4	3
30	158	80	69	=B31-B30	=C31-C30	30	158	3	4
31	159	83	73	=B32-B31	=C32-C31	31	159	3	3
32	160	86	76	=B33-B32	=C33-C32	32	160	2	2
33	161	88	78	=B34-B33	=C34-C33	33	161	3	3

Continúa...

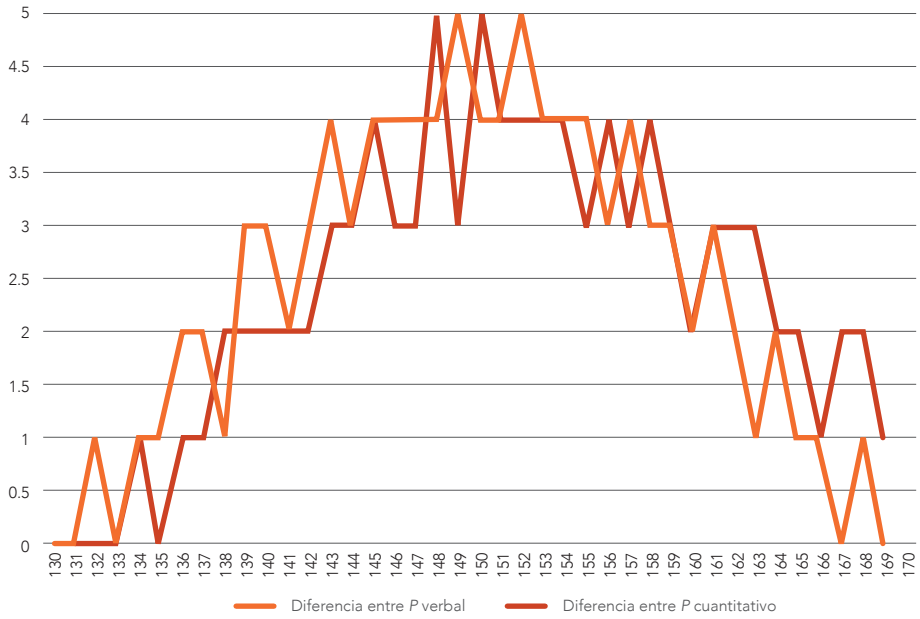
	A	B	C	D	E		"A"	"D"	"E"
34	162	91	81	=B35-B34	=C35-C34	34	162	2	3
35	163	93	84	=B36-B35	=C36-C35	35	163	1	3
36	164	94	87	=B37-B36	=C37-C36	36	164	2	2
37	165	96	89	=B38-B37	=C38-C37	37	165	1	2
38	166	97	91	=B39-B38	=C39-C38	38	166	1	1
39	167	98	92	=B40-B39	=C40-C39	39	167	0	2
40	168	98	94	=B41-B40	=C41-C40	40	168	1	2
41	169	99	96	=B42-B41	=C42-C41	41	169	0	1
42	170	99	97			42	170		

Fuente: McCammon (2016): Columnas de la A a la C.

En la Tabla F-2 se muestran las fórmulas de las diferencias entre percentiles de las secciones verbal y cuantitativa del GRE (columnas D y E). Dentro de cada sección de esta prueba, se restó el percentil más pequeño del más grande que le seguía. Por ejemplo, los puntajes 131 y 132 de la sección verbal se encuentran ambos con el percentil 1, así que al restarlos (=B4–B3) dan como resultado 0. Esto indica que la diferencia entre estos puntajes es nada en cuestión de percentil. Este mismo proceso se llevó a cabo con las dos secciones. En las columnas "D" y "E" se muestran los resultados de estas diferencias. Se puede apreciar en estas dos últimas columnas que las primeras y las últimas diferencias entre los percentiles son las más pequeñas (0, 1 y 2). Mientras que en medio de estas diferencias, se encuentran los valores más altos de estas restas (3, 4 y 5).

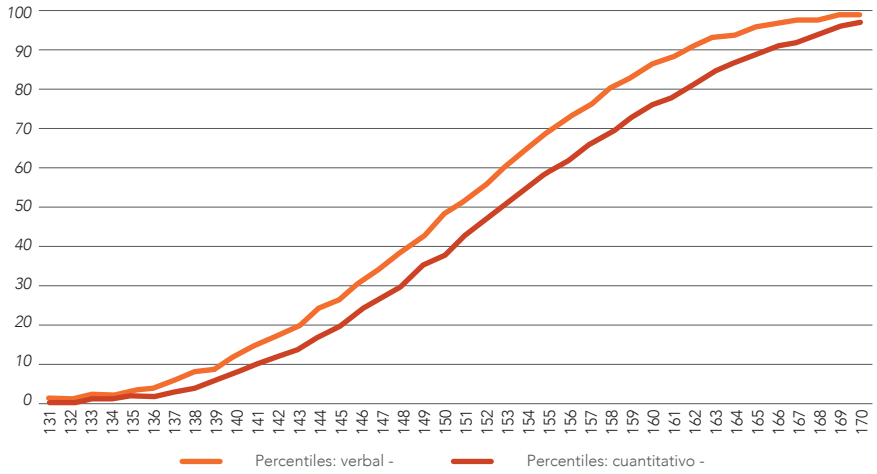
En la Figura F-1 se muestran estas diferencias de las columnas "D" y "E." Para hacer esta figura en Excel 2016 solo hay que seleccionar las columnas A, "D" y "E", y seguir los pasos de la Figura 3.2. En la Figura F-1 se muestra que las diferencias se incrementan entre los percentiles cuando se va a la mitad del conjunto de datos y decrecen en los extremos. De este fenómeno, Hinkle *et al.* (2003) advirtieron que las diferencias entre percentiles no son uniformes, ya que se exageran las diferencias entre los percentiles del centro y se minimizan las diferencias con los percentiles en los extremos. Esta es otra evidencia de que los percentiles están medidos en una escala ordinal; por ello, no se pueden hacer operaciones aritméticas con los percentiles (véase: Tabla 2.1 para las escalas).

Figura F-1 Diferencias entre percentiles de verbal y percentiles de cuantitativo



Asimismo, en la Figura F-2 se muestra visualmente algo que ya se había expuesto en párrafos anteriores y que se puede observar en la Tabla F-2: el mismo puntaje en la sección verbal y cuantitativa significa diferentes percentiles. Por esta diferencia en los percentiles, las curvas de estos se separan.

Figura F-2 Mismo puntaje y diferente percentil



Nota: Para crear esta figura hay que seleccionar las columnas A-C de la Tabla F-2 y seguir los pasos de la Figura 3.2.

En unas cuantas palabras, los percentiles son medidos en escalas ordinales y sirven para describir. Si se desea hacer operaciones aritméticas, se tendría que usar una escala diferente que no fuera la de los percentiles.

Rango de percentil

Con un propósito meramente descriptivo y para hacer *comparaciones* entre diferentes sets de datos, se pueden utilizar los rangos de percentiles. Para más detalles al respecto, se recomienda repasar el Capítulo 6 en la sección *b) Rango de percentil*. Una advertencia es que los rangos de percentil están medidos en una escala ordinal, así que no se deben hacer operaciones aritméticas con ellos, como sí se podría hacer en otras escalas (véase el Capítulo 2, donde en la Tabla 2.1 y la Tabla 2.2 se muestra qué operaciones se pueden hacer con escalas ordinales).

Apéndice G. Respuestas a algunos Problemas para resolver

Capítulo 1. Problema

1. Resultado: 185.
2. Se tiene un 95% de confianza de que el parámetro de la población, se encuentra a más o menos 2 puntos.
3. Aumenta a 191.
4. Baja a 132.

Capítulo 2. Problema

Tabla P-2.1 Solución de escala ordinal

	A	B
1	10	1
2	9.1	2
3	8.6	3
4	8.3	4
5	7.5	5
6	7.2	6

Tabla P-2.2 Solución de escala nominal (seis calificaciones aprobatorias)

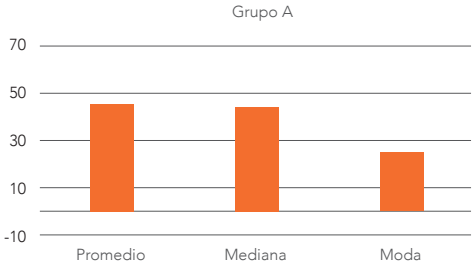
	A	B	C
1	10	Aprobatoria	1
2	9.1	Aprobatoria	1
3	8.6	Aprobatoria	1
4	8.3	Aprobatoria	1
5	7.5	Aprobatoria	1
6	7.2	Aprobatoria	1
7	6.9	No-aprobatoria	0
8	6.3	No-aprobatoria	0
9	5.2	No-aprobatoria	0

Capítulo 4

Problema 1.

- a. Promedio = 45.85; mediana = 44.50; y moda = 25.
- b. Como la moda no está al mismo nivel que el promedio y la mediana, tal vez no se sigue una distribución normal.

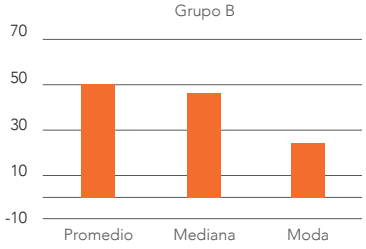
Figura S-4.1 Problema 1, pregunta b



Problema 2.

- a. Promedio = 49.79; mediana = 45; y moda = 24.
- b. Como la moda no está al mismo nivel que el promedio y la mediana, tal vez no se sigue una distribución normal.

Figura S-4.2 Problema 2, pregunta b

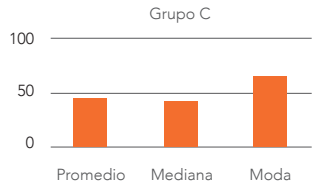


Problema 3.

- a. Promedio = 44.91; mediana = 43; y moda = 65.

» b. Como la moda no está al mismo nivel que el promedio y la mediana, tal vez no se sigue una distribución normal.

Figura S-4.3 Problema 3, pregunta b



Problema 5.

85.

Capítulo 5

Problema 1.

» a. Máximo: 10; mínimo: 4;
» b. Rango inclusivo = 7; rango exclusivo = 6.

Problema 2.

» a. MDP = 2.56.
» b. Varianza = 9.06.
» c. *SD* = 3.01.
» d. Máximo = 10; mínimo = 1.
» e. Rango inclusivo = 10; rango exclusivo = 9.
» f. Promedio = 5.44.

Problema 3.

» a. Promedio = 7.56; mediana = 7.50; y moda = 7.
» b. Varianza = 2.25; y *SD* = 1.50.
» c. Rango inclusivo = 6; rango exclusivo = 5.

Capítulo 6

Problema:

» a. En la Tabla S-6.1 se muestran los percentiles inclusivos y exclusivos calculados en Excel.

Tabla S-6.1 a. Percentiles

	A	B	C	D
1	Tipo de *P*	P_{25}	P_{50}	P_{75}
2	Inclusivo	30	58	78
3	Exclusivo	27	58	81

El significado del percentil inclusivo (P_{25}) es que el 25% de los valores sería igual o menor a 30 (≤ 30). De igual manera, el P_{50} inclusivo significa que el 50% de los datos sería ≤ 58; y el P_{75} inclusivo indica que el 75% de los valores sería ≤ 78. En cambio, el percentil exclusivo (P_{25}) denota que el 25% de los datos estaría por debajo de 27 (< 27). El P_{50} exclusivo coincide con el P_{50} inclusivo, ya que muestran el valor 58, pero cambia su interpretación, porque para el exclusivo el 50% estaría por debajo de 58 (< 27). Para el P_{75}, el 75% de los valores estaría por debajo de 81 (< 81).

» b. Las columnas B y C (Tabla S-6.2) contienen el Rango de Percentil Inclusivo (RPI) y el Rango de Percentil Exclusivo (RPE). Un rango de percentil *inclusivo* de un puntaje es un punto en la escala percentil, que da el porcentaje de puntajes que caen en el puntaje especificado o por debajo de él (*cf.* Hinkle *et al.*, 2003). Por ejemplo, el número 9 tiene un RPI = 3%, lo que significa que el 3% de los datos son iguales o menores a 9 (≤ 9). Similarmente, el significado de un RPI se aplicaría para cada uno de los valores del set: 5% (≤ 10), 11% (≤ 12),…, 100% (100). El RPE implica que el porcentaje asignado al valor en un conjunto de datos, es más alto que el resto en una secuencia de números de menor a mayor (*cf.* Salkind, 2007). Por ejemplo, el 9 tiene un RPE del 5%, lo que significa que el 5% de los valores son menores a

9 (< 9). De la misma manera, lo anterior aplica para el resto del set: 8% (< 10),…, 98% (< 100).

Tabla S-6.2 Rangos de percentil inclusivos y exclusivos, rangos y valores estándar

	A	B	C	D	E
1	x_i	RPI	RPE	Rango	*z*
2	4	0%	3%	34	-1.72
3	9	3%	5%	33	-1.55
4	10	5%	8%	32	-1.51
5	10	5%	8%	32	-1.51
6	12	11%	13%	31	-1.44
7	13	13%	15%	30	-1.41
8	16	16%	18%	29	-1.30
9	17	18%	20%	28	-1.27
10	26	21%	23%	27	-0.96
11	27	24%	25%	26	-0.92
12	33	26%	28%	25	-0.72
13	38	29%	30%	24	-0.54
14	42	32%	33%	23	-0.41
15	43	34%	35%	22	-0.37
16	46	37%	38%	21	-0.27
17	49	39%	40%	20	-0.16
18	49	39%	40%	20	-0.16
19	52	45%	45%	19	-0.06
20	55	47%	48%	18	0.04
21	58	50%	50%	17	0.15
22	60	53%	53%	16	0.22
23	60	53%	53%	16	0.22
24	63	58%	58%	15	0.32
25	64	61%	60%	14	0.35
26	65	63%	63%	13	0.39
27	65	63%	63%	13	0.39
28	65	63%	63%	13	0.39
29	73	71%	70%	12	0.66
30	75	74%	73%	11	0.73
31	81	76%	75%	10	0.94

Continúa...

	A	B	C	D	E
32	83	79%	78%	9	1.01
33	84	82%	80%	8	1.04
34	86	84%	83%	7	1.11
35	88	87%	85%	6	1.18
36	90	89%	88%	5	1.25
37	91	92%	90%	4	1.29
38	96	95%	93%	3	1.46
39	99	97%	95%	2	1.56
40	100	100%	98%	1	1.60
41	Promedio	53.77			
42	*SD*	28.97			

» c. Los rangos se encuentran en la columna D (Tabla S-6.2) y muestran la jerarquía de los valores del set. Por ejemplo, el número 100 tiene el primer lugar en el set, porque tiene el valor más alto del mismo; el 99 tiene el segundo lugar y así sucesivamente hasta llegar al 4, que tiene el último lugar (34). En este caso, los rangos (34) no corresponden al tamaño de la muestra (n = 39), porque hay números que se repiten, como el 65, y cada uno de ellos tiene el mismo rango (decimotercero = 13).

» d. Los cuartiles se muestran en la Tabla S-6-3. Los cuartiles inclusivos señalan qué porcentaje es igual o menor a cierto valor del set. El significado es que el C_1 inclusivo indica que el 25% de los datos son menores o iguales al valor 30 (≤ 30). De la misma manera, lo anterior aplica a los inclusivos: C_2 (50% ≤ 58), C_3(75% ≤ 78) y C_4 (78 ≤ 100). De una manera similar, el cuartil exclusivo C_1 indica que el 25% es menor a 27, así como que: C_2(50% ≤ 58) y C_3(75% ≤ 81). El cuartil exclusivo C_4 no existe.

Tabla S-6.3 Cuartiles e intercuartil

	C_1	C_2	C_3	C_4
Inclusivos	30	58	78	100
Exclusivos	27	58	81	NA

- e. El rango intercuartil para el cuartil inclusivo, se obtiene de la diferencia entre el C_3 y el C_1. En este caso, 78 – 30 = 48 (véase: Tabla S-6.3; Ecuación 6.3). Un intercuartil significa la variación entre los cuartiles. Esto cobraría un significado más amplio si se le compara con otros sets de datos para ver si esta variación es pequeña, mediana o grande.
- f. La Frontera Razonable Inferior (FRI) = -42 y la Frontera Razonable Superior (FRS) = 150. Como no son superadas, aparentemente no existen observaciones atípicas. Para más detalles, véanse las Ecuaciones 6.3-6.5.
- g. Los valores estándar aparecen en la Tabla S-6.2. Los valores estándar significan la distancia medida en desviaciones estándar que tiene cada uno de ellos con el promedio, que es cero. Por ejemplo, el valor 4, que al volverlo estándar (diferencia entre 4 menos el promedio [53.77] y dividido por la desviación estándar del set [28.397]) da -1.72, lo que significa que el 4 está por debajo del promedio en 1.72 desviaciones estándar. En resumen, los valores estándar negativos están por debajo del promedio y los positivos, por encima. Lo anterior no solo aplicaría a los valores del set del ejemplo, sino también a todos los valores estándar.

Referencias

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational Psychological Testing*. APA.

American Psychological Association (APA). (2019). *Publication Manual of the American Psychological Association: The Official Guide to APA Style* (7th Ed.). APA.

Analytics Vidhya. (September 18th, 2017). 6 Common Probability Distributions Every Data Science Professional Should Know. Retrieved July 8th, 2021, from https://www.analyticsvidhya.com/blog/2017/09/6-probability-distributions-data-science/

Bandalos, D. L., & Finney, S. J. (2010). Exploratory and Confirmatory Factor Analysis. In G. R. Hancock & R. O. Mueller (Eds.), *Quantitative Methods in the Social and Behavioral Sciences: A Guide for Researchers and Reviewers* (pp. 93-114). Routledge.

Barrantes-Aguilar, L. (2019). Diferencias en la estimación del Coeficiente de Curtosis en diferentes *softwares* estadísticos. *Revista E-Agronegocios, 5*(2), 1-7. https://doi.org/10.18845/rea.v5i2.4456

Batanero, C., Burrill, G., & Reading, C. (Eds.). (2011). *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education*. Springer.

Ben-Zvi, D. & Garfield, J. B. (2005). *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. Springer.

Bennett, J., Briggs, W. L., & Triola, M. F. (2014). *Statistical Reasoning for Everyday Life* (4th Ed.). Pearson.

Berenson, M. L., Levine, D. M., Szabat, K. A., & Stephan, D. F. (2019). *Basic Business Statistics: Concepts and Aplications* (14th Ed.). Pearson.

Blair, E. & Blair, J. (2015). *Applied Survey Sampling*. Sage.

Bluman, A. G. (2018). *Elementary Statistics: A Step by Step Approach* (10th Ed.). McGraw-Hill.

Byrne, B. M. (2016). *Modeling with Amos: Basic Concepts, Applications, and Programming* (3rd Ed.). Routledge.

Creative Research Systems. (2012). Sample Size Calculator. [En línea] Survey System. Retrieved May 4th, 2021, from https://www.surveysystem.com/sscalc.htm#one

Crocker, L. & Algina, J. (2008). *Introduction to Classical & Modern Test Theory*. Cengage.

Cumming, G. (2013). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. Routledge.

Dattalo, P. (2008). *Determining Sample Size: Balancing Power, Precision, and Practicality*. Oxford University Press.

Dumont, H., Istance, D., & Benavides, F. (Eds.). (2010). *The Nature of Learning: Using Research to Inspire Practice*. Centre for Educational Research and Innovation.

Encyclopedia Britannica (s.f.). Function Mathematics. Retrieved August 14th, 2021, from https://www.britannica.com/science/function-mathematics

Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press.

Fernández, S., Córdoba, A., & Cordero, J. M. (2002). *Estadística descriptiva*. ESIC Editorial.

Frost, J. (2021). Normal Distribution in Statistics – Statistics by Jim. [En línea] *Statistics by Jim*. Retrieved April 6th, 2021, from https://statisticsbyjim.com/basics/normal-distribution/

Garfield, J. B. & Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice*. Springer.

Garson, G. D. (2012). *Curve Fitting & Nonlinear Regression* (Blue Book Series). Statistical Publishing Associates.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis* (8th Ed.). Cengage Learning.

Harpe, S. E. (2015). How to Analyze Likert and Other Rating Scale Data. *Currents in Pharmacy Teaching and Learning, 7*, 836-850. https://doi.org/10.1016/j.cptl.2015.08.001

Hernández-Sampieri, R. & Mendoza, C. P. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta*. McGraw-Hill Education.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied Statistics for the Behavioral Sciences* (5th Ed.). Houghton Mifflin.

Hopkins, K. D. & Weeks, D. L. (1990). Tests for Normality and Measures of Skewness and Kurtosis: Their Place in Research Reporting. *Educational and Psychological Measurement, 50*(4), 717-729. https://doi.org/10.1177/0013164490504001

Hurley, P. & Watson, L. (2018). *A Concise Introduction to Logic* (13th Ed.). Cengage Learning.

ICEX España Exportación e Inversiones. (2019). Educación Superior en México. [En línea] Recuperado el 4 de mayo de 2021, de https://www.icex.es/icex/es/navegacion-principal/todos-nuestros-servicios/programas-y-servicios-de-apoyo/centros-de-negocio/4292242.html

Illeris, K. (Ed.). (2018). *Contemporary Theories of Learning: Learning Theorists… In their Own Words* (2nd Ed.). Taylor & Francis.

Jones, A. R. (2019). *Best Fit Lines and Curves, and Some Mathe-Magical Transformations* (Vol. III). Routledge.

Lane, D. M., Scott, D., Hebl, M., Guerra, R., Osherson, D., & Zimer, H. (2014). *Introduction to Statistics*. David Lane/Rice University.

Larson, R. & Edwards, B. H. (2010). *Cálculo 1 de una variable* (9.ª ed.). McGraw-Hill.

Levine, D. M., Stephan, D. F., & Szabat, K. A. (2021). *Statistics for Managers: Using Microsoft® Excel®* (9th Ed.). Pearson.

Libman, Z. (2010). Integrating Real-life Data Analysis in Teaching Descriptive Statistics: A Constructivist Approach. *Journal of Statistics Education, 18*(1), 1-23.

Maciejewski, R. (2011). *Data Representations, Transformations, and Statistics for Visual Reasoning*. Morgan & Calypool.

Maxwell, S. E., Delaney, H. D., & Kelly, K. (2018). *Designing Experiments and Analyzing Data: A Model Comparison Perspective* (3rd Ed.). Routledge.

McCammon, E. (2016, 17 de octubre). (Updated). GRE Score Percentiles: What They Mean for You [Blog Post]. https://www.prepscholar.com/gre/blog/gre-score-percentiles/

Morris, D. (2016). *Bayes' Theorem: A Visual Introduction for Beginners*. Blue Windmill Media.

Muñiz, J. (2018). *Introducción a la Psicometría: Teoría clásica y TRI*. Pirámide.

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2021, 6 de julio). *Programa Internacional de Evaluación de los Alumnos* (PISA). https://www.oecd.org/centrodemexico/medios/programainternacionaldeevaluaciondelosalumnospisa.htm

Osborne, J. W. (2008). Best Practices in Data Transformation. En J. W. Osborne (Ed.), *Best Practices in Quantitative Methods* (pp. 197-204). Sage.

Paolella, M. S. (2019). *Linear Models and Time-series Analysis: Regression, ANOVA, ARMA and Garch*. Wiley.

Ponce-Renova, H. F. (2016). Evaluación de los índices de reprobación de la universidad usando intervalos de confianza. *Revista Electrónica del Congreso de Investigación Educativa, 3*(5), 47-57.

----- (2019). *Conceptos básicos de estadísticas inferenciales aplicadas a la Investigación Educativa*. Universidad Autónoma de Ciudad Juárez.

----- (2020). *Estadística elemental para la Investigación Educativa: Probabilidad, distribuciones y correlación*. Universidad Autónoma de Ciudad Juárez.

----- (2021). *Estadística para comparaciones básicas de grupos: Con uso de SPSS y calculadoras en línea*. Universidad Autónoma de Ciudad Juárez.

Rajaraman, V. (2014). John McCarthy — Father of Artificial Intelligence. *Resonance*, 19, 198-207. https://doi.org/10.1007/s12045-014-0027-9

Ravid, R. (2020). *Practical Statistics for Educators* (6th Ed.). Rowman & Littlefield.

Ross, S. M. (1997). *Introduction to Probability Models* (6th Ed.). Academic Press.

Rossi, R. (2018). *Mathematical Statistics: An Introduction to Likelihood Bases Inference*. Wiley.

Russo, R. (2021). *Statistics for the Behavioral Sciences: An Introduction to Frequentist and Bayesian Approaches* (2nd Ed.). Routledge.
Rust, J., Kosinski, M., & Stillwell, D. (2021). *Modern Psychometricts: The Science of Psychological Assessment* (4th Ed.). Routledge.
Rutherford, A. (2011). *ANOVA and ANCOVA: A GLM Approach* (2nd Ed.). Wiley.
Salkind, N. J. (Ed.). (2007). *Encyclopedia of Measurement and Statistics* (Vols. 1-3). Sage.
----- (2017). *Statistics for People Who (Think They) Hate Statistics* (6th Ed.). Sage.
Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating Causal Effects using Experimental and Observational Designs.* American Educational Research Association.
Schunk, D. H. (2012). *Teorías del aprendizaje: Una perspectiva educativa* (6.ª ed.). Pearson.
Shultz, K. S., Whitney, D. J., & Zickar, M. J. (2020). *Measurement Theory in Action* (3rd Ed.). Routledge.
Tabachnick, B. G. & Fidell, L. S. (2019). *Using Multivariate Statistics* (7th Ed.). Pearson.
The Windows Club. (2021, June 18th). *What is the Maximum Number of Columns & Rows in Excel Worksheet*. Retrieved June 28th, 2021, from https://www.thewindowsclub.com/what-is-the-maximum-number-of-columns-rows-in-excel-worksheet
Thorndike, E. L. (1918). The Nature, Purposes and General Methods of Measurements of Educational Products. *The Seventeenth Yearbook of the National Society for the Study of Education* (Part II: The Measurement of Educational Products; pp. 16-24).
Toye, F. (2015). Not Everything that Can Be Counted Counts and Not Everything that Counts Can Be Counted (Attributed to Albert Einstein). *British Journal of Pain, 9*(1), 7.
Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company.
Universo de Fórmulas. (2021). Recuperado el 27 de abril de 2021, de https://www.universoformulas.com/estadistica/descriptiva/percentiles/
VandenBos, G. R. (Coord.). (2015). *APA Dictionary of Psychology* (2nd Ed.). American Psychological Association.

Weschler, D. (1997). *The Weschler Adult Intelligence Scale* (3rd Ed.). The Psychological Corporation.
Wild, C. J. & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review, 67*(3), 223-248.

Recursos de internet

American Statistical Association (https://www.amstat.org/#)
Excel Video Training (https://support.microsoft.com/en-us/office/excel-video-training-9bc05390-e94c-46af-a5b3-d7c22f6990bb)
JASP (https://jasp-stats.org/)
Quick Start (https://support.microsoft.com/en-us/office/create-a-workbook-in-excel-94b00f50-5896-479c-b0c5-ff74603b35a3?wt.mc_id=otc_excel)

YouTube

Chan, K. (2016). *Use Excel 2016 to Find Summary Statistics for Quantitative Data* (https://youtu.be/vWX4ZH4k8To)
Dunaetz, D. (2016). *How to Calculate Descriptive Statistics in Excel 2016 for Mac using the Data Analysis Toolpak* (https://youtu.be/25Lb8TA_Qts)
Flores, K. R. (2014). *Estadística descriptiva en Excel* (https://youtu.be/1IuCBFuNam4)
Jardín, E. (2016). *Descriptive Statistics in Excel* (https://youtu.be/4_9vGqQaCFk)
Pozo, S. (2017). *Estadística descriptiva, media, moda, mediana, de, varianza con Excel* (https://youtu.be/jIw8nOfwDh4)