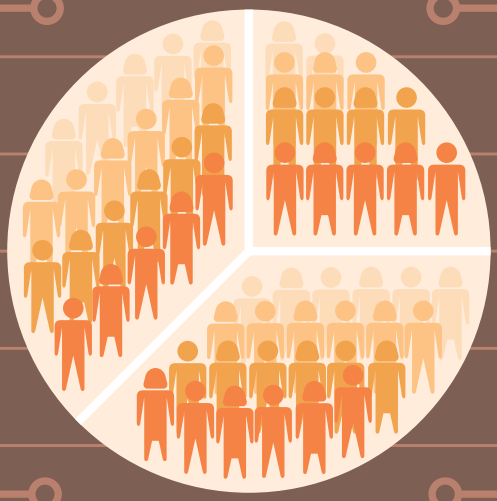


Pensamiento estadístico para docentes de bachillerato

Santiago Inzunza Cázares

COLEGIO DE BACHILLERES DEL ESTADO DE SINALOA
Organismo Público Descentralizado

Pensamiento estadístico para docentes de bachillerato

Santiago Inzunza Cázares

COLEGIO DE BACHILLERES DEL ESTADO DE SINALOA
Organismo Público Descentralizado

COLEGIO DE BACHILLERES DEL ESTADO DE SINALOA
Organismo Público Descentralizado

Pensamiento estadístico para docentes de bachillerato

Santiago Inzunza Cázares

Primera edición

ISBN: 978-607-59343-3-4

Av. Independencia No.2142 Sur. Col. Centro Sinaloa, C.P.80129,
Culiacán, Sin. Tel. 01(667)758-68-30

Versión digital en *www.cobaes.edu.mx*
Culiacán Rosales, Sinaloa, octubre de 2023
Edición con fines culturales, no lucrativos

Maquetación y diseño: **Ito Contreras**

Hecho en México / Printed in Mexico

Presentación

En las últimas tres décadas, los contenidos de estadística y probabilidad han ocupado espacios en los programas de estudio de matemáticas en todos los niveles educativos como ninguna otra temática; esto por la importancia de la estadística como herramienta metodológica y transversal a las demás ciencias, pero además por la importancia que está teniendo la alfabetización y el pensamiento estadístico en la sociedad actual, caracterizada por muchos expertos, como la sociedad de la información y del conocimiento.

Entre las causas principales que han generado la expansión de la estadística como campo de estudio y aplicación, destaca la revolución de los datos, impulsada por el creciente desarrollo de las tecnologías de la información y de las comunicaciones. Ello a su vez ha generado un fenómeno conocido como cuantificación o datificación de la sociedad, caracterizado por la necesidad de expresar y representar el comportamiento de diversos fenómenos en términos cuantitativos mediante gráficas, tablas, porcentajes, promedios, correlaciones, modelos, variabilidad, y otras medidas que definen el comportamiento de los datos provenientes de muestras, poblaciones y experimentos aleatorizados.

Por otra parte, la versión tradicional de la estadística la identifica como una rama de las matemáticas que utiliza fórmulas y procedimientos para la organización, visualización y cálculo de medidas en los datos con propósitos descriptivos, predictivos o inferenciales. Sin embargo, con el desarrollo de la tecnología y sus aplicaciones en la educación, han surgido herramientas de software para realizar gran parte del trabajo estadístico que ocupaba la mayor atención en el pasado; como consecuencia, se ha generado un cambio de enfoque hacia el desarrollo de habilidades de razonamiento, pensamiento y alfabetización estadística, considerados los nuevos aprendizajes en la educación estadística.

En este contexto, los docentes requieren cambiar esa visión tradicional por una visión más aplicada ligada a las raíces históricas de la estadística, que consiste en la solución de problemas del entorno. De tal forma, en el presente libro partimos de la perspectiva de que nuestros alumnos de bachillerato son usuarios de la estadística, por lo que nos centraremos en generar comprensión de sus conceptos y métodos, en lugar de su aplicación formal como rama de las matemáticas, que caracteriza a un curso universitario.

El primer capítulo presenta una visión general de la estadística, su campo de estudio y su metodología. Se introducen los nuevos aprendizajes

en la educación estadística, la idea de variabilidad, las ramas en que clasifica la estadística y los tipos principales de estudios de estadísticos. En el segundo capítulo se plantea el ciclo de investigación estadística que se requiere para resolver un problema, mismo que constituye la base de un modelo para el desarrollo del pensamiento estadístico propuesto Wild & Pfannkuch (1999)1. Los capítulos tres, cuatro y cinco abordan diferentes fases de este ciclo.

En el capítulo tres, se estudia la recolección y producción de los datos de manera general a través de muestreos probabilísticos y no probabilísticos, se identifican sus ventajas y desventajas respecto al censo, así como algunos errores comunes que se pueden generar con el muestreo. Cuando ya se dispone de los datos, es necesario pasar a su organización, presentación y visualización, para lo cual en el capítulo cuatro se abordan las distribuciones de frecuencia, y diferentes tipos de gráficas para datos cualitativos y cuantitativos. En el capítulo 5 se avanza en el análisis de los datos, mediante el cálculo de resúmenes numéricos, también conocidos como medidas descriptivas de los datos, para el caso de promedios y variabilidad.

Hasta antes de iniciar el capítulo seis, se han estudiado métodos estadísticos para un análisis básico de datos que provienen de una sola variable. Es necesario dar un paso más el en análisis considerando dos variables, ello da lugar al estudio de gráficas de dispersión y la correlación lineal, así como a la construcción de modelos de regresión lineal simple.

En el capítulo siete, nos introducimos al estudio de las leyes básicas del azar mediante el estudio de la probabilidad. Se abordan los diferentes enfoques sobre la probabilidad, su escala de medición y algunas propiedades, para concluir con el tema de variables aleatorias y distribuciones de probabilidad; en particular con la distribución binomial y la distribución normal.

Finalmente, en el capítulo ocho introducimos la idea de inferencia estadística, el rol del muestreo y la incertidumbre debida a la variabilidad en una inferencia, para pasar y dar una idea general de los conceptos básicos de los métodos de estimación de parámetros y pruebas de significación.

Santiago Inzunza Cázares
Colegio de Bachilleres del Estado de Sinaloa

Contenido

Capitulo 1

La estadística, su campo de estudio y metodología

La estadística requiere una forma diferente de pensar, porque los datos no son sólo números, más bien son números en contexto.

David Moore

Estadísticas son "un retrato de la realidad": INEGI

Recalcó que para la construcción de las políticas públicas se requieren datos duros

GUILLERMO CASTAÑARES

En el Día Mundial de la Estadística, la cual se celebra cada cinco años desde 2010 por la Asamblea General de las Naciones Unidas (ONU), dicha herramienta de estudio es imprescindible para la toma de mejores decisiones, de acuerdo con Julio Santaella, presidente del INEGI.

"La estadística es muy importante para los mexicanos porque a partir del retrato de la realidad es como se puede tomar mejores decisiones informadas y no con base a ocurrencias", señaló.

Todas las oficinas de estadística del mundo que están afiadas a la ONU se unirán para "conectar el mundo con datos en los que podemos confiar", que es el lema de esta celebración.

Santaella agregó que para el desarrollo nacional se necesita poder planear la política pública, y para ello se tiene que utilizar información estadística, que es la que coordina el INEGI.

El funcionario destacó que es muy importante la confiabilidad de los datos y resaltó tres puntos en los que ha trabajado el INEGI.

En primer lugar, son más de 37 años de experiencia con una trayectoria de profesionalismo, por lo que tienen un equipo muy preparado.

Dos, el entorno institucional "que contamos desde hace más 12 años, es importante, porque ya tenemos una autonomía constitucional, lo cual, nos protege de interferencias políticas puesto que la misma ley nos manda que seamos veraces, oportunos y pertinentes, brindando la información que se requiere".

Por último, destacó que los aspectos metodológicos a los que se apegan son estándares internacionales, lo que garantiza información de calidad, "estamos obligados con fechas fijas de publicación", dijo.

DÍA MUNDIAL DE LA ESTADÍSTICA. Julio Santaella, presidente del INEGI

16
ENCUESTAS
A establecimientos lleva a cabo el INEGI de manera regular, anual y especial

38
AÑOS
Cumplirá el próximo 25 de enero de 2021 el principal organismo estadístico del país.

1.1 Introducción

En los últimos años la estadística ha tenido un crecimiento notable en los programas de estudio, desde la educación básica hasta la universidad, debido a su importancia como herramienta metodológica para el análisis de diversos fenómenos que son cuantificables por medio de datos, pero además por la importancia que está teniendo el razonamiento, la alfabetización y el pensamiento estadístico en la

sociedad actual. Estas habilidades son cada vez más necesarias para dar sentido a información expresada en lenguaje estadístico sobre diversos hechos de interés para los ciudadanos, las ciencias y las profesiones.

Una de las principales causas que ha generado el crecimiento de la estadística como campo de estudio es la *revolución de los datos*, impulsada por el aumento en el desarrollo de las tecnologías de la información y de las comunicaciones, que han hecho posible la generación y almacenamiento de grandes cantidades de datos que requieren ser convertidos en información útil para la toma de decisiones. Ello a su vez ha derivado en un fenómeno conocido como *cuantificación o datificación de la sociedad*, originado por la necesidad de expresar y representar el comportamiento de diversos fenómenos en términos cuantitativos, mediante gráficas, tablas, porcentajes, promedios, correlaciones, modelos, variabilidad y otras medidas que definen el comportamiento de los datos provenientes de muestras, poblaciones y experimentos aleatorizados.

En este contexto han surgido, de manera reciente, términos como *Big Data (grandes volúmenes de datos)* y *Open Data (datos abiertos)* relacionados con la abundancia y disponibilidad de datos que se generan diariamente, como resultado de las actividades productivas y de comunicación de las personas, las empresas, las ciencias y el gobierno. Ambos términos están estrechamente ligados al estudio de la estadística.

El *Big Data* es un término que refiere a la gestión y análisis de grandes cantidades de datos; que plantea la necesidad de métodos estadísticos y herramientas tecnológicas avanzadas. Para fijar ideas sobre el potencial del *Big Data*, imaginemos la cantidad de información que los usuarios de Google generan en un solo día de acceso a su plataforma, o la información que transmite un satélite de comunicaciones desde y hacia la tierra. En la actualidad, muchas empresas y gobiernos utilizan técnicas de *Big Data* en sus procesos de recopilación, análisis y procesamiento de sus datos.

En cuanto a los *datos abiertos*, se trata de una iniciativa que ha surgido en diversos países en la última década, con el propósito de abrir el acceso a la información y los datos recolectados por el gobierno y organismos no gubernamentales a cualquier persona, sin restricciones de derechos de autor u otros mecanismos de control. En Estados Unidos se ha creado la página www.data.gov para colocar sus datos abiertos; de igual forma en México se puede acceder a datos abiertos en la página www.datos.gob.mx. En ambos casos, la idea fundamental que emerge es que los datos son fuente de información importante para conocer el comportamiento de un fenómeno.

También se ha vuelto cada vez más común la publicación de estudios de opinión y de mercado en los medios de comunicación, los cuales requieren que las personas

tengan cierto nivel de alfabetización estadística para su adecuada interpretación. En este sentido, se ha vuelto realidad lo que señaló Herbert George Wells (1866-1946) hace casi un siglo, "algún día el pensamiento estadístico será tan necesario como la habilidad para leer y escribir".

En la sociedad de la información y el conocimiento, por mucha experiencia que tenga una persona sobre un tema particular, no es posible tomar decisiones con base en anécdotas o creencias. En cambio, los datos proporcionan una base racional y un fundamento más sólido para el conocimiento de un fenómeno. La estadística, como ciencia de los datos, ha desarrollado *métodos para la recolección*, análisis e interpretación de los mismos, razón por la cual se ha convertido en una valiosa herramienta en diversas áreas de la actividad humana.

1.2 Los nuevos aprendizajes en la educación estadística

Tradicionalmente la estadística ha sido conceptualizada como una rama de las matemáticas que utiliza fórmulas y procedimientos para la organización, visualización y cálculo de medidas en los datos obtenidos (de una o varias variables) con propósitos descriptivos, predictivos o inferenciales. Con el desarrollo de la tecnología y sus aplicaciones en la educación, han surgido potentes herramientas de software para realizar gran parte del trabajo estadístico que antaño ocupaba una mayor atención, como consecuencia, se ha generado un cambio de enfoque hacia el desarrollo de habilidades de razonamiento, pensamiento y alfabetización estadística, considerados los nuevos aprendizajes en la educación estadística.

La *alfabetización estadística* es definida por Gal (2002)[1] como la habilidad para comprender, interpretar, comunicar y evaluar en forma crítica información estadística. A la alfabetización estadística también se le conoce como "estadística para la ciudadanía". Todos los ciudadanos requerimos estar alfabetizados estadísticamente para comprender y dar sentido a los cúmulos de información que diariamente se vierten en los medios de

Figura 1. Periódico Excélsior 11/03/2019

1 Adults' Statistical Literacy: Meanings, Components, Responsibilities. *International Statistical Review*, 70(1).

comunicación (ver figura 1), y de esta manera desenvolvernos adecuadamente en una sociedad permeada por los datos.

El *razonamiento estadístico* es definido por Garfield (2002)[2] como el uso de conceptos estadísticos y comprensión de información estadística; incluye interpretaciones, representaciones y resúmenes de datos. Por su parte, el *pensamiento estadístico* de acuerdo con Moore (1997)[3] involucra comprender: la necesidad de los datos, el diseño para la producción de los datos, reconocimiento de la variación, cuantificación de la variación y explicación de la variación. En este libro, utilizaremos los modelos de pedagogía estadística más conocidos para desarrollar las habilidades anteriores.

1.3 Clasificación de las áreas de estudio de la estadística

Tradicionalmente se ha dicho que la estadística se divide en dos grandes áreas: *estadística descriptiva (análisis de datos) y estadística inferencial (inferencia estadística*). Sin embargo, esta clasificación no considera una importante área de la estadística, que es aquella que proporciona los métodos para la *recolección de los datos*. A continuación, describiremos brevemente el campo de acción de cada una de estas tres áreas, y en capítulos posteriores se profundizará en sus conceptos y métodos.

• Recolección de los datos

El proceso de diseño para la obtención de los datos es una de las etapas más importantes del ciclo de investigación estadística. Existen diversos métodos basados en *teoría del muestreo* y *diseño de experimentos* que permiten la recopilación de datos, ya sean de muestras, poblaciones o experimentos. En el área de ciencias sociales, las muestras y los censos son más utilizados, mientras que, en ingeniería, ciencias físicas y naturales, los experimentos son la principal fuente de recolección de la información.

• Análisis de datos

Tradicionalmente esta área de la estadística ha sido conocida como *estadística descriptiva*, cuando el objetivo era precisamente la descripción de los datos mediante una serie de medidas numéricas, gráficas y tablas. El desarrollo de la tecnología computacional ha revolucionado este sector de la estadística, y ha hecho posible un enfoque exploratorio en el análisis que va mucho más allá de la mera descripción de los datos.

2 The Challenge of Developing Statistical Reasoning. *Journal of Statistics Education*, 10(3). jse.amstat.org/v10n3/garfield.html

3 New Pedagogy and New Content: The Case of Statistics. *International Statistical Review*, 65(2).

El análisis de datos contempla los métodos que permiten organizar, resumir, explorar y visualizar los datos para identificar patrones de comportamiento. Los recursos estadísticos comúnmente utilizados son las tablas estadísticas, gráficas, medidas descriptivas o resúmenes numéricos. Sin pérdida de generalidad, un análisis básico de datos contempla al menos tres aspectos:

1. Organizar y presentar los datos en tablas estadísticas y distribuciones de frecuencia.
2. Explorar y visualizar los datos mediante diversas gráficas (diagramas de barras, histogramas, diagramas de caja, diagramas de dispersión, entre otras)
3. Calcular medidas descriptivas de los datos (centralidad, variabilidad, correlación, entre otras).

• Inferencia estadística

La inferencia estadística es una de las áreas de mayor aplicación de la estadística, mediante la utilización de sus métodos se pueden obtener conclusiones significativas acerca de toda una población, con base en la información que proporcionan los datos de una sola muestra o un experimento; métodos que hacen uso del azar para seleccionar los elementos de la muestra o asignarlos a los diversos tratamientos de un experimento, lo que permite el uso de la teoría de la probabilidad para evaluar la confiabilidad y margen de error de los resultados obtenidos.

Se pueden identificar tres características clave que son parte de una inferencia estadística:

1. Un enunciado de *generalización* que va más allá de los datos.
2. Uso de *datos de una muestra aleatoria* como evidencia para apoyar esta generalización.
3. Un lenguaje probabilístico que expresa la *confiabilidad y precisión* de la generalización.

1.4 Variabilidad: razón de ser de la estadística

La *variabilidad* está en todos lados. No solo los individuos varían en una misma característica, es también el caso para las mediciones repetidas de un mismo individuo u objeto, para los resultados de muestras seleccionadas de una misma población o de individuos sujetos a distintos tratamientos. En suma, la variabilidad es una característica intrínseca de los datos. La estadística emplea métodos para describir la variabilidad, predecirla, controlarla y explicarla, puesto que no es posible eliminarla.

Para fijar ideas, consideremos el pronóstico del clima para la ciudad de Culiacán correspondiente a una semana en particular. Supongamos que nos piden

un reporte sobre las temperaturas que se esperan en la semana (ver figura 2), ¿qué temperaturas tomaríamos en consideración, si éstas varían de un momento a otro y día con día?, ¿la temperatura máxima de un día de la semana, o la temperatura mínima?, ¿un rango de temperaturas o un promedio?, ¿o una combinación de promedio y rango?, ¿qué otras medidas pueden ser útiles para describir el comportamiento de las temperaturas? La estadística nos proporciona métodos para hacer una descripción numérica de la variabilidad de las temperaturas, de tal manera que una persona pueda tomar una decisión relacionada con la temperatura de la ciudad.

Figura 2. Pronóstico del clima para la ciudad de Culiacán

• Escalas de medición y tipo de variables

Los elementos de una población poseen características o atributos que interesa estudiar y que varían de un elemento a otro, es decir, presentan variabilidad. Dichas características reciben el nombre de *variables estadísticas* o simplemente *variables*. Las variables se pueden medir y los resultados reciben el nombre de *datos*. Las variables se clasifican de acuerdo a la escala de medición utilizada para obtener la información, lo cual a su vez determina el tipo de análisis que se va utilizar. Por ejemplo, algunas gráficas pueden ser utilizadas sólo para ciertos tipos de variables; lo mismo ocurre con el cálculo de medidas descriptivas, por lo que es importante distinguir la escala de medición que será utilizada para medir una variable.

• Variables cualitativas o categóricas

Son aquellas variables que al ser medidas producen datos que se pueden clasificar en categorías o grupos. Hay dos tipos de variables categóricas: nominales y ordinales. Las *variables nominales* clasifican los resultados en categorías. Por ejemplo, la variable "género" puede producir dos resultados: masculino y femenino. La variable "estado civil" presenta cinco categorías: soltero, casado, viudo, divorciado, unión libre. Con frecuencia, y sobre todo cuando se utiliza un software para el análisis de los datos, es necesario codificar las categorías de las variables para facilitar el procesamiento. El código, aunque es numérico, sólo es una etiqueta para representar los datos.

Variable: estado civil

1	2	3	4	5
Soltero	Casado	Viudo	Divorciado	Unión libre

Con los datos que resultan al medir este tipo de variables, sólo se pueden realizar operaciones como conteos y sumas para calcular frecuencias y proporciones en un análisis estadístico.

Variable: género

Los datos que proceden de *variables ordinales* se pueden clasificar y ordenar en categorías. Por ejemplo, el salario de una persona puede ser clasificado en bajo, medio y alto, o bien, el nivel de escolaridad puede ser clasificado en primaria, secundaria, preparatoria, universidad; el nivel de satisfacción sobre un servicio puede medirse en una escala del 1 al 10, donde el valor de la escala aumenta con el grado de satisfacción.

Con los datos que resultan al medir este tipo de variables, se pueden realizar operaciones como conteos y sumas para calcular frecuencias y proporciones, ordenamientos para calcular un promedio.

• Variables cuantitativas o numéricas

Son aquellas variables que al ser medidas producen valores numéricos. Por ejemplo, la estatura, el peso, el ingreso salarial, la edad, temperatura, entre otras. Hay dos tipos de variables cuantitativas: discretas o continuas. Las *variables discretas* producen valores numéricos enteros y generalmente proceden de procesos de conteo. Por su parte, las *variables continuas*, pueden producir fraccionarios y por lo general proceden de procesos de medición.

Variable: edad (años)

Con los resultados de este tipo de variables se pueden realizar operaciones aritméticas como sumas, multiplicaciones y divisiones, que permiten calcular promedios, porcentajes, correlaciones entre otros descriptores estadísticos.

Actividad de aprendizaje

1. Considera una parte del cuestionario que INEGI utiliza en la Encuesta Nacional de Gastos en los Hogares (EnGasto). Identifica para cada pregunta del cuestionario el tipo de variable que se involucra.

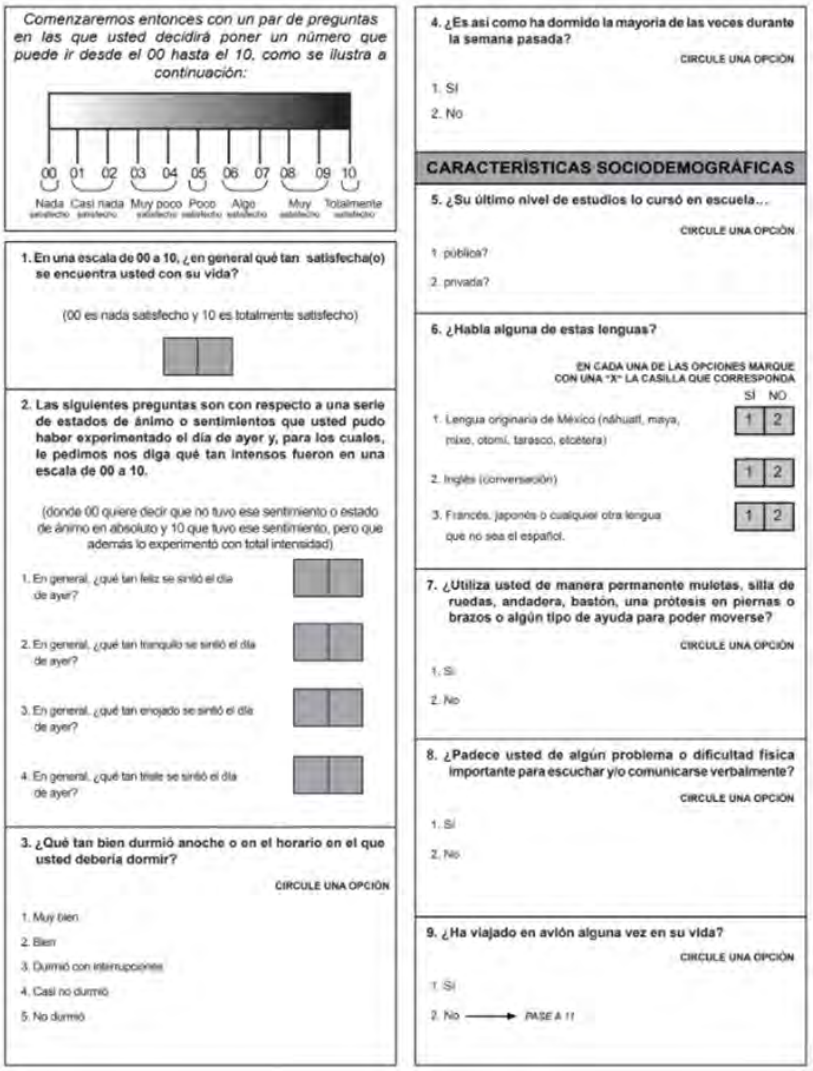
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

ENCUESTA NACIONAL DE GASTOS DE LOS HOGARES (ENGASTO)

Módulo de bienestar autorreportado

1.5 Diferentes tipos de estudios estadísticos

Existen diferentes tipos de estudios estadísticos, los cuales se pueden clasificar en estudios observacionales y experimentales.



Comenzaremos entonces con un par de preguntas en las que usted decidirá poner un número que puede ir desde el 00 hasta el 10, como se ilustra a continuación:

00 01 02 03 04 05 06 07 08 09 10

Nada satisfecho · Casi nada satisfecho · Muy poco satisfecho · Poco satisfecho · Algo satisfecho · Muy satisfecho · Totalmente satisfecho

1. En una escala de 00 a 10, ¿en general qué tan satisfecha(o) se encuentra usted con su vida?

(00 es nada satisfecho y 10 es totalmente satisfecho)

2. Las siguientes preguntas son con respecto a una serie de estados de ánimo o sentimientos que usted pudo haber experimentado el día de ayer y, para los cuales, le pedimos nos diga qué tan intensos fueron en una escala de 00 a 10.

(donde 00 quiere decir que no tuvo ese sentimiento o estado de ánimo en absoluto y 10 que tuvo ese sentimiento, pero que además lo experimentó con total intensidad)

1. En general, ¿qué tan feliz se sintió el día de ayer?
2. En general, ¿qué tan tranquilo se sintió el día de ayer?
3. En general, ¿qué tan enojado se sintió el día de ayer?
4. En general, ¿qué tan triste se sintió el día de ayer?

3. ¿Qué tan bien durmió anoche o en el horario en el que usted debería dormir?

CIRCULE UNA OPCIÓN

1. Muy bien
2. Bien
3. Durmió con interrupciones
4. Casi no durmió
5. No durmió

4. ¿Es así como ha dormido la mayoría de las veces durante la semana pasada?

CIRCULE UNA OPCIÓN

1. Sí
2. No

CARACTERÍSTICAS SOCIODEMOGRÁFICAS

5. ¿Su último nivel de estudios lo cursó en escuela...

CIRCULE UNA OPCIÓN

1. pública?
2. privada?

6. ¿Habla alguna de estas lenguas?

EN CADA UNA DE LAS OPCIONES MARQUE CON UNA "X" LA CASILLA QUE CORRESPONDA

	SÍ	NO
1. Lengua originaria de México (náhuatl, maya, mixe, otomí, tarasco, etcétera)	1	2
2. Inglés (conversación)	1	2
3. Francés, japonés o cualquier otra lengua que no sea el español.	1	2

7. ¿Utiliza usted de manera permanente muletas, silla de ruedas, andadera, bastón, una prótesis en piernas o brazos o algún tipo de ayuda para poder moverse?

CIRCULE UNA OPCIÓN

1. Sí
2. No

8. ¿Padece usted de algún problema o dificultad física importante para escuchar y/o comunicarse verbalmente?

CIRCULE UNA OPCIÓN

1. Sí
2. No

9. ¿Ha viajado en avión alguna vez en su vida?

CIRCULE UNA OPCIÓN

1. Sí
2. No → PASE A 11

• Estudios observacionales

Como su nombre lo indica, estos estudios consisten en observar a los elementos (de una muestra o población) y medir sus características de interés, buscando alterar lo menos posible sus condiciones. Es decir, no se requiere manipular las variables consideradas en un estudio, sino solo observarlas y medirlas. Las encuestas de opinión, los estudios de mercado, los censos poblacionales, son ejemplos de este tipo de estudios. De la misma manera se procede cuando el propósito es con fines comparativos; los estudios observacionales obtienen información de los grupos que se desean comparar, observando y midiendo las variables de interés, para después realizar los cálculos de los indicadores y ver si hay alguna diferencia en los grupos.

• Estudios experimentales

Estos estudios consisten en imponer deliberadamente una condición o tratamiento sobre los elementos con el propósito de observar sus respuestas. Es decir, se busca observar cómo responde una variable (*variable de respuesta*), cambiando otra o más variables de manera intencionada (*variables explicativas*). Los elementos sobre los cuales se realiza el experimento se les denomina *unidades experimentales*. Si las unidades

son personas se les denomina *sujetos* y una determinada condición experimental se denomina *tratamiento*. Para fijar ideas sobre los dos tipos de estudio consideremos lo siguiente:

Se desea probar que las dietas ricas en frutas y verduras pueden reducir el riesgo de algunos tipos de cáncer y otras enfermedades crónicas. Si se elige un estudio observacional, se podría comparar un grupo de personas que han llevado la dieta basada en frutas y verduras por algún tiempo, con un grupo que no ha llevado la dieta. Se analiza la incidencia de enfermedades en los dos grupos y se observa si existen diferencias entre ellos.

Si optamos por un estudio experimental, lo que procede es localizar individuos que deseen participar de manera voluntaria, asignarlos al azar en dos grupos, los que llevarán la dieta basada en frutas o vegetales y los que llevarán una dieta distinta. La asignación al azar garantiza que exista cierta uniformidad en ambos grupos. Los estudios experimentales nos permiten determinar si las diferencias observadas en los grupos fueron debidas a las condiciones impuestas (si tuvo efecto o no la dieta), los estudios observacionales generalmente no.

Nota histórica

La estadística surgió de recopilar información de personas, como nacimientos, defunciones, matrimonios en el siglo XVII. Uno de sus precursores fue *John Graunt*, quien en 1662 utilizó estadísticas oficiales para estimar la población de Londres. Desde muy joven, Graunt se interesó por las tablas de mortalidad que se publicaban cada semana, las cuales contenían el número de muertos que se presentaban en cada parroquia y otros recuentos, como nacimientos y bautizos.

Con dichos datos, Graunt publicó el libro titulado *Natural and Political Observations made upon the Bills of Mortality*, el cual fomentó el estudio de las estadísticas de vida en todo el Continente Europeo; por todo ello, a John Graunt le considera el fundador de la Estadística. Algunos siglos después, a la estadística se le relacionó con la probabilidad, ciencia que estudia el azar; en conjunto, probabilidad y estadística, constituyen una herramienta poderosa para estudiar el azar en los datos.

Para tu reflexión

En la década de los años 90 era muy frecuente que después de la menopausia las mujeres tomaran hormonas (estrógenos) cuando su producción natural disminuía. Las mujeres que tomaban hormonas parecía que reducían su riesgo de ataques del corazón del 35% al 50%, y los riesgos de tomar hormonas parecían pequeños comparados con los beneficios. En el año 2002, los Institutos Nacionales de Salud en los Estados Unidos declararon que éstos resultados eran erróneos, por lo que el uso de hormonas después de la menopausia de detuvo inmediatamente. Si ambas recomendaciones estaban basadas en estudios extensivos, ¿qué fue lo que pasó?

La evidencia a favor del uso de hormonas después de la menopausia se obtuvo de diversos *estudios observacionales* que comparaban la salud de mujeres que estaban tomando hormonas con la salud de otras mujeres que no las estaban tomando; pero las mujeres que estaban tomando hormonas eran muy diferentes de las mujeres que nos las estaban tomando: eran de mayor nivel económico y más educadas, por lo que veían a los doctores con más frecuencia y hacían diversas actividades para mantener su salud, así que no es sorprende que tuvieran menos ataques al corazón.

En el año 2002 se realizaron diversos *estudios experimentales* con mujeres de diferentes edades y confirmaron que tomar hormonas no reduce el riesgo de ataques al corazón. En un experimento como los que se realizaron, las mujeres son asignadas en forma aleatoria a algunos de los tratamientos (tomar pastillas con hormonas o tomar pastillas de igual apariencia y sabor, pero sin hormonas, conocidas como placebos), así que todas las mujeres tienen la misma probabilidad de ser asignadas a un tratamiento u otro, sin importar su condición, con ello se elimina el sesgo y la diferencia entre los resultados –si esta existe- solo se explica por causa de los tratamientos.

(Fuente: Moore, D., 2006). Aprendiendo de los datos. En Roxy Peck y colaboradores. La estadística: una guía de lo desconocido

Evaluación del capítulo

1. La Universidad de Oxford realizó un estudio donde se concluye que la *Dexametosona* constituye un eficiente tratamiento contra el Covid-19 en pacientes que reúnen ciertas condiciones. Un total de 2,104 infectados recibieron seis milígramos del medicamento una vez al día por vía oral o por inyección intravenosa durante 10 días, cuya evolución fue comparada con la de 4,321 que recibieron los cuidados convencionales. Al comparar los resultados tras 28 días, la mortalidad se redujo de 41 a 28 por ciento entre quienes requirieron respirador, y de 20 a 25 por ciento menor entre aquellos pacientes que solo necesitaron oxígeno. ¿De qué tipo de estudio estadístico se trata?

a) Estudio observacional
b) Estudio longitudinal
c) Estudio Experimental

2. Comprender, interpretar, comunicar y evaluar en forma crítica y reflexiva información estadística (por ejemplo, gráficas y promedios) que se presenta en los medios de comunicación y en reportes gubernamentales, es una habilidad que se denomina:

a) Alfabetización estadística
b) Razonamiento estadístico
c) Pensamiento estadístico

3. Las preguntas de una encuesta de satisfacción por la prestación de un servicio, muestran las siguientes opciones de respuesta (ver figura). Las variables que se están evaluando, son variables:

a) Nominales
b) Ordinales
c) Numéricas

4. Elige tres características clave que son parte de una inferencia estadística.

a) Un enunciado que va más allá de los datos disponibles
b) Los datos pueden ser recolectados a través de experimentos o muestras
c) Uso de datos de una muestra aleatoria como evidencia para apoyar una generalización
d) Un lenguaje probabilístico que expresa la confiabilidad de una generalización
e) Las gráficas permiten identificar patrones en los datos

5. Organizar y presentar los datos en tablas de frecuencia, b) explorar y visualizar los datos mediante diversas gráficas (diagramas de barras, histogramas, diagramas de caja, diagramas de dispersión, entre otras), c) calcular medidas descriptivas de los datos (centralidad, variabilidad, correlación, entre otras). Las tres fases anteriores caracterizan a una rama de la estadística conocida como:

a) Inferencia estadística
b) Recolección de datos
c) Análisis de datos

6. Realiza una lectura a algunos artículos que se reportan en la bibliografía sobre los términos Big Data, Revolución de los datos, Datos abiertos, y relaciona la forma como están impactando en la estadística actualmente.

Bibliografía recomendada

- Instituto Nacional de Estadística y Geografía. https://www.inegi.org.mx
- Uso de gráficas y resúmenes estadísticos en conferencia de prensa sobre Covid-19 https://www.pscp.tv/w/1rmxPAOrprmKN?t=6m53s
- Big Data. https://www.michaelpage.es/advice/empresas/desarrollo-profesional/big-data-la-revolución-de-los-datos-masivos
- El mundo de la estadística. http://www.worldofstatistics.org
- Biografía de John Graunt. https://apuntesdedemografia.com/2015/04/28/john-graunt-primera-tabla-de-mortalidad/
- Datos abiertos. https://opendatahandbook.org/guide/es/what-is-open-data/
- Revolución de los datos. https://www.csic.es/es/actualidad-del-csic/la-revolucion-de-los-datos

Capitulo 2

El ciclo de resolución de los problemas estadísticos

Algún día el pensamiento estadístico será tan necesario como la habilidad para leer y escribir.

Herbert George Wells

2.1 Introducción

Nos proponemos iniciar el estudio de la estadística como lo hace un profesional de la estadística, analizando problemas reales, planteando preguntas para responder con los datos, discutiendo las diferentes formas que existen para su recolección, seleccionando los métodos más adecuados para el análisis de los datos y su interpretación.

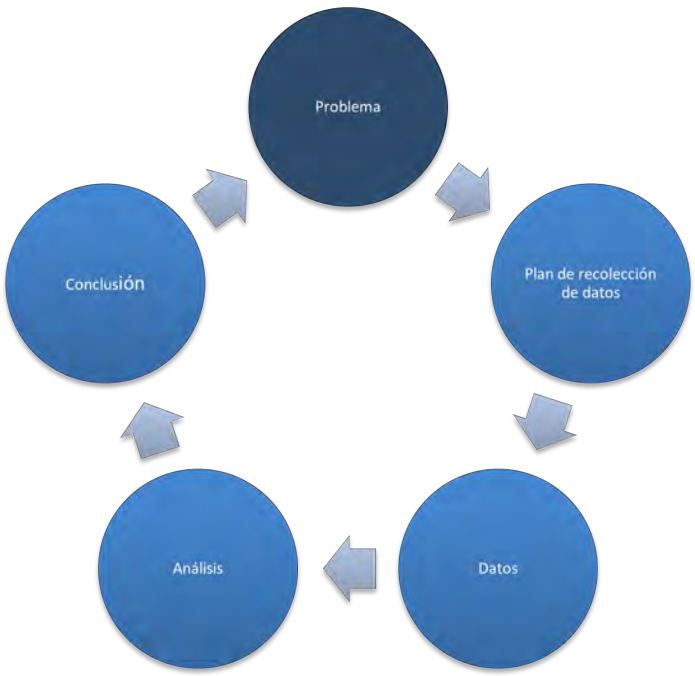


Figura 1: Ciclo de investigación estadística (Problema-Plan-Datos-Análisis-Conclusiones)

Para responder las preguntas adecuadamente se requiere seguir un proceso sistemático conocido como *ciclo de investigación estadística*, que inicia con el planteamiento del problema y termina con la presentación de resultados y conclusiones. Dicho ciclo de investigación se puede visualizar en la Figura 1.

Una mirada a la metodología que utiliza el Instituto Nacional de Estadística y Geografía (INEGI) en México, nos puede dar una idea sobre la importancia que se otorga al proceso de planeación para la recolección de los datos mediante encuestas. El INEGI ha definido un proceso estándar para

realizar encuestas por muestreo que contempla las siguientes fases:

1. Planeación
2. Diseño conceptual
3. Diseño de la muestra
4. Diseño de la captación y el procesamiento
5. Captación
6. Procesamiento
7. Presentación de resultados

Otro modelo más general para la resolución de problemas estadísticos como medio para el aprendizaje de la estadística es el que propone la Asociación Americana de Estadística (ASA por sus siglas en inglés) en el documento *Lineamientos para la Evaluación y Enseñanza en Educación Estadística* (GAISE, por sus siglas en inglés); involucra cuatro componentes:

1. **Formular preguntas sobre un problema:**
   - Tener claro el problema a resolver
   - Formular preguntas que puedan ser contestadas con los datos
2. **Recolectar los datos:**
   - Diseñar un plan para recolectar los datos
   - Emplear el plan para recolectar los datos
3. **Analizar los datos**
   - Seleccionar los métodos gráficos y numéricos apropiados
   - Usar los métodos para analizar los datos
4. **Interpretar los resultados**
   - Interpretar el análisis
   - Relacionar la interpretación con las preguntas planteadas

Como puede verse, la estadística no se limita solo al análisis de los datos, memorización de fórmulas y procedimientos, como usualmente se piensa. La estadística es la *ciencia de los datos*, y constituye una herramienta metodológica de gran utilidad parar resolver problemas de muchas ciencias y profesiones.

En los siguientes capítulos del libro abordaremos con detalle y a profundidad las diferentes etapas de un ciclo de investigación estadística y los conceptos y métodos estadísticos que intervienen en cada una de ellas, para dar forma a la solución de un problema estadístico.

2.2 Planteamiento de un problema estadístico

Es el punto de partida del ciclo de investigación estadística. En esta etapa se reflexiona sobre un problema que puede ser resuelto con herramientas y métodos de la estadística.

Antes de pasar a la siguiente etapa debes cerciorarte de haber comprendido bien el problema. En particular se requiere que identifiques y definas lo siguiente:

1. Identificar la *población objetivo*.
2. Identificar las *variables* de estudio.
3. Plantear algunas *preguntas de investigación* acerca de las variables elegidas.

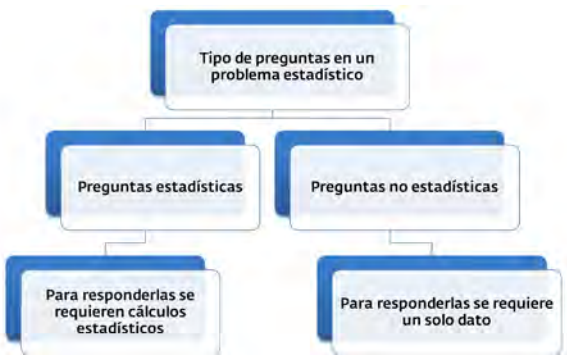
El planteamiento de un problema puede empezar con una idea general sobre algún tema, y poco a poco se puede ir acotando y traduciendo en ideas formales que generan un problema factible de investigar. Consideremos el caso de la Encuesta Nacional sobre Disponibilidad y Uso de TIC en Hogares que realiza el INEGI en México.
Población objetivo: Todas las personas de 6 años o más que permanentemente residen en los hogares ubicados en el territorio nacional
Algunas *variables* de interés:

1. Uso de computadora, laptop o tableta en los últimos tres meses dentro o fuera del hogar.
2. Tiempo que utiliza la computadora en un día.
3. Razones por las que no utiliza computadora.
4. Frecuencia con la que utiliza la computadora.
6. Género de la persona encuestada.
6. Edad de la persona encuestada.

Posibles *preguntas de investigación:*

1. ¿Qué porcentaje de personas mayores de 6 años utiliza computadora?
2. ¿Cuál es el tiempo promedio que una persona utiliza computadora?
3. **¿Con qué** frecuencia utilizan las personas la computadora?
4. ¿Existen diferencias por grupo de edad de las personas en cuanto al tiempo de uso de la computadora?
5. ¿Existe diferencia de género en cuanto a la frecuencia de uso de la computadora?

Es importante distinguir entre preguntas estadísticas y no estadísticas. Las preguntas estadísticas involucran variabilidad en los datos. Por ejemplo, se tiene información de los precios de vehículos último modelo de diferentes marcas que se venden en México, la pregunta ¿qué vehículo es más económico? es una pregunta no estadística porque puede



ser respondida con un solo dato. En cambio ¿cuál es el peso promedio de los vehículos? y ¿cuál marca de vehículos es más económica? son preguntas estadísticas, porque los precios varían y se requiere el cálculo de medidas estadísticas.

Las preguntas de investigación que hemos planteado son todas preguntas estadísticas. Por ejemplo, en la primera pregunta se involucra a la variable edad de las personas, e interesa calcular un porcentaje. La segunda involucra la variable tiempo de uso de la computadora e interesa calcular un promedio. La quinta pregunta involucra a las variables género y frecuencia de uso de la computadora, e interesa comparar las frecuencias de uso por género de la persona.

Ejemplo de preguntas no estadísticas en el contexto de este problema pueden ser: ¿cuál es el mayor número de horas que una persona utiliza una computadora al día?, ¿cuál es la edad más alta de las personas que utilizan computadora? En ambos casos las preguntas involucran variabilidad y se pueden responder con un solo dato, sin necesidad de cálculos estadísticos.

2.3 Recolección de los datos

En esta fase se recopilan los datos para responder las preguntas que se plantearon en la definición del problema. Se pueden identificar dos niveles: el plan para la recolección de los datos y proceso de recolección de los datos.

• Plan de recolección de los datos

Según sea el proyecto estadístico que se ha planteado, los datos se pueden recopilar principalmente de tres fuentes: internet, mediciones u observaciones directa y encuestas. Los datos de internet provienen de fuentes secundarias, porque fueron recopilados por otras personas; en el caso de mediciones directas y encuestas, los datos son de fuentes primarias, en tanto fueron obtenidos de manera directa por el interesado.

En internet hay datos disponibles prácticamente de cualquier tema (por ejemplo: datos gubernamentales, clima, finanzas, contaminación); algunas colecciones de datos están en bases de datos para ser analizados de manera directa a través de algún software estadístico, en otros casos se requiere recopilar los datos y prepararlos para su procesamiento.

En el caso de encuestas, el investigador diseña el cuestionario que van a responder los sujetos de estudio (cara a cara, en línea o vía telefónica). El diseño de un cuestionario para una encuesta requiere de mucho

cuidado para que los datos obtenidos sean de buena calidad. Hay toda una teoría para el diseño de encuestas que van más allá de los propósitos de este libro.

• El proceso de recolección de los datos

Una vez que se ha construido el instrumento de recolección de datos (cuestionario) se requiere definir la estrategia para la recolección de los datos. Una decisión importante en este momento consiste en determinar si se va estudiar la población completa (censo) o si se va recurrir a una parte de ella (muestreo). El uso de muestras en estudios estadísticos es bastante frecuente, dado que comúnmente las poblaciones suelen ser muy grandes y complejas. Si se decide por el muestreo, es importante tener en cuenta tres aspectos que en conjunto se denominan *diseño de la muestra*:

- La población objetivo,
- El tamaño de la muestra,
- La forma de seleccionar la muestra de la población.

En los siguientes capítulos abordaremos, en lo general, los principales métodos de muestreo (probabilísticos y no probabilísticos). Los métodos de muestreo probabilísticos son preferidos, porque hacen posible la generalización de los resultados de la muestra a la población objetivo, y además permiten el cálculo de la confiabilidad y el margen de error, dos indicadores que deben acompañar los resultados de cualquier estudio muestral.

Un elemento más a considerar en esta etapa de diseño de la muestra consiste en definir la precisión de los datos, posibles codificaciones, depuración (en caso de ser necesario) y captura en un software estadístico para su posterior análisis estadístico.

2.4 Análisis de los datos

Un análisis básico puede incluir ordenamiento de los datos, construcción de tablas estadísticas y de frecuencias, diversas gráficas, cálculo de medidas descriptivas y en algunos casos puede incluir la estimación o una prueba de significación sobre un parámetro poblacional. Las gráficas ocupan un lugar importante en el análisis de los datos, pues permiten visualizar patrones y tendencias en los datos. Las tablas complementan el análisis gráfico, mientras que el cálculo de medidas descriptivas precisa de forma contundente el comportamiento de los datos.

El análisis de los datos ocupa una parte importante de un curso de estadística; en nuestro caso, sólo es una componente más del ciclo de investigación estadística. En los próximos capítulos abordaremos los métodos para organizar los datos en distribuciones de frecuencias y tablas estadísticas, cómo construir las gráficas más apropiadas para datos cualitativos (gráficas de barras y circulares) y cuantitativos (histogramas, diagramas de caja, diagramas de puntos, diagramas de tallo y hoja,

gráficas de línea), y finalmente, los métodos estadísticos para analizar la tendencia central, variabilidad, forma y relaciones entre los datos.

La tecnología computacional es un elemento importante que ha cambiado la forma de analizar los datos en las últimas décadas. En este libro nos apoyaremos fundamentalmente en software estadístico (*Geogebra*, Excel y applets) para realizar los cálculos y construir gráficas de los datos. De tal forma, el esfuerzo de cálculo lo dejaremos a la tecnología, y nos centraremos en aspectos conceptuales de los datos y en su interpretación, con el propósito de contribuir a la alfabetización y pensamiento estadístico.

2.5 Conclusiones del problema

Las conclusiones son la última etapa del ciclo de investigación estadística. En esta parte se deben interpretar las tablas, gráficas y medidas estadísticas calculadas sobre los datos para responder las preguntas que se han planteado al inicio del ciclo.
El contexto del problema juega un papel muy importante en la interpretación, pues los datos son números rodeados de un contexto. De tal forma, en los razonamientos e interpretaciones se deben entrelazar los conceptos estadísticos necesarios con el contexto de los datos.

Actividad de aprendizaje

Analiza la encuesta realizada por el periódico El Financiero en marzo de 2020, para conocer la opinión de los habitantes de la Ciudad de México sobre las medidas de distanciamiento social tomadas por el gobierno para combatir la expansión del coronavirus.

Metodología: Encuesta en la Ciudad de México realizada por vía telefónica a 400 adultos el 19 y 20 de marzo de 2020. Se hizo un muestreo probabilístico de teléfonos residenciales y celulares en 16 Alcaldías. Con un nivel de confianza de 95%, el margen de error de las estimaciones es de +/-4.9 por ciento.

https://www.elfinanciero.com.mx/nacional/apoyan-9-de-cada-10-capitalinos-distanciarse-socialmente

Identifica las etapas del ciclo de investigación estadística en la realización de la encuesta, en particular responde lo siguiente:

- **Planteamiento del problema**
  - ¿Cuál es el problema que aborda la encuesta?

- ¿Cuál fue la población objetivo?
- ¿Qué variables estadísticas se consideraron en el estudio?

- **Recopilación de los datos**
  - ¿Cómo se recopilaron los datos?
  - ¿Cuál fue el tamaño de la muestra?
- Análisis de los datos
  - ¿Qué gráficas se utilizaron para analizar los datos?
  - ¿Qué medidas estadísticas se calcularon con los datos?
- **Conclusiones**
  - ¿Cuáles son las conclusiones de la encuesta?
  - ¿Qué tanta confiabilidad tiene los resultados?

Para tu reflexión

En el año 2000 surge en Inglaterra el proyecto *Census at School*, convertido ahora en un proyecto internacional que implementan varios países (Estados Unidos, Nueva Zelanda, Canadá, Australia, entre otros). El objetivo de *Census at School* es fomentar entre los estudiantes de educación básica y bachillerato mayor motivación por el estudio de la estadística y su utilidad para resolver problemas con datos reales, y de esta manera desarrollar su alfabetización y pensamiento estadístico utilizando el ciclo de resolución de los problemas estadísticos.

El insumo principal *Census at School* son datos de los mismos estudiantes recopilados por ellos a través de una encuesta en línea que ellos completan, generando así una gran base de datos de todos los países participantes en el proyecto. Los estudiantes y profesores pueden acceder a los datos a través de una liga de internet que realiza un muestreo aleatorio de la población. Los datos son reales y de contextos significativos para ellos. Se disponen además diversos materiales de enseñanza que están basados en las fases del ciclo de resolución de problemas estadísticos.

En el caso de Nueva Zelanda, a partir del año 2007 su currículo de estadística a nivel preuniversitario expresa de forma explícita tres líneas en las que se desarrollan los aprendizajes esperados: investigaciones estadísticas, alfabetización estadística y probabilidad, en las que el eje principal es el ciclo de investigación estadística que hemos expuesto en este capítulo. Con ello, Nueva Zelanda se ubica como un país líder en educación estadística.

Evaluación del capítulo

1. Identifica los elementos que forman parte del planteamiento de un problema estadístico.

a) Identificar la población objetivo
b) Identificar las variables de estudio
c) Plantear algunas preguntas de investigación sobre las variables del estudio
d) Construir tablas y gráficas con los datos

e) Calcular promedios y porcentajes con los datos

2. Identifica los elementos que forman parte de la recolección de los datos a través de muestras:

a) La población objetivo
b) Plantear algunas preguntas de investigación sobre las variables del estudio
c) El tamaño de la muestra
d) La forma de seleccionar la muestra de la población.
e) Calcular promedios y porcentajes con los datos

3. En el siguiente conjunto de preguntas, identifica las preguntas estadísticas

a) ¿Cuál es la estatura promedio de los mexicanos?
b) ¿Qué porcentaje de estudiantes mexicanos tiene computadora portátil?
c) ¿En qué universidad estudia el alumno con la mayor estatura en México?
d) ¿Cuál es el modelo de computadora portátil de mayor costo en México?
e) ¿Cuál es la edad mediana de las personas que han enfermado de Covid en México?

4. Ordena las cinco etapas del ciclo de investigación estadística

a) Conclusiones
b) Problema
c) Datos
d) Análisis
e) Plan de recolección de datos

Bibliografía recomendada

- Potencial de los proyectos para desarrollar motivación, competencias de razonamiento y pensamiento estadístico.
  https://revistas.ucr.ac.cr/index.php/aie/article/view/29874/29886
- Estadística con proyectos
  https://www.ugr.es/~batanero/pages/ARTICULOS/Libroproyectos.pdf
- El papel de los proyectos en la enseñanza de la estadística
  https://www.ugr.es/~batanero/pages/ARTICULOS/CEIO.pdf

Capitulo 3

La recolección y producción de los datos

Una deficiente recopilación y producción de datos puede conducir a resultados erróneos.

Michael Shaughnessy

3.1 Introducción

Los datos son la materia prima del trabajo estadístico, no son solamente números como usualmente se les considera, más bien son números con contexto y significado; en este sentido el trabajo estadístico va mucho más allá de la manipulación de números y gráficas, implica entre otras cosas plantearse preguntas en torno a un problema en particular, diseñar un plan para la recolección de los datos, analizar los datos y visualizar sus patrones de comportamiento para responder las preguntas.

Los patrones de comportamiento en los datos son sensibles a la forma como fueron recolectados, así que un diseño inadecuado para la recolección de los datos puede con facilidad conducir a resultados sesgados, por ello la etapa de recolección de los datos es fundamental en el proceso de resolución de un problema estadístico, pues de poco servirían los métodos más refinados de análisis estadístico si los datos están sesgados de origen y no son representativos del fenómeno que se está estudiando.

Los datos generalmente son recolectados de muestras, censos y experimentos. La elección de un muestreo sobre un censo depende de los propósitos del problema, ligado a ello se encuentran factores de costo y tiempo. Consideremos dos casos para ejemplificar lo anterior, los cuales han sido tomados de la *revista aregional* (www.aregional.com) especializada en temas de economía regional.

- Un estudio realizado en 2010 sobre las finanzas de los estados de la República Mexicana señala que más del 71% de las entidades federativas del país destina más de la tercera parte de sus recursos para el pago de salarios y prestaciones a su personal. Dicho estudio reporta además otros importantes indicadores del

gasto administrativo de las entidades.

- Los municipios mexicanos invierten en promedio el 62% de su presupuesto en gastos administrativos. Dicha información fue obtenida de una muestra de 70 municipios con calificación de riesgo crediticio.

En el primer caso se trata de un censo, dado que la información fue obtenida de la totalidad de las entidades federativas del país; en el segundo caso se trata de un muestreo pues se consideraron solo 70 municipios. La decisión del muestreo sobre el censo, en el segundo caso, seguramente se debió a la gran cantidad de municipios que existen en el país.

Otra fuente importante de datos en muchas áreas de estudio son los experimentos, sobre todo en las ciencias naturales, áreas de la salud e ingenierías. Un experimento consiste en imponer una o más condiciones (también llamadas tratamientos) a los elementos de estudio, con el objeto de observar sus respuestas y determinar posibles diferencias y sus causas.

3.2 Métodos de muestreo

Uno de los principios estadísticos que debe reunir una muestra para que sea apropiada para realizar generalizaciones a una población, consiste en utilizar el azar en la selección de sus elementos. Una muestra seleccionada bajo este principio se conoce como *muestra probabilística o aleatoria*. Cuando la selección no se realiza al azar, sino atendiendo otros principios como la conveniencia o la facilidad, se denomina *muestra no probabilística*.

Por lo general se pretende que la *muestra* seleccionada sea *representativa* de la población. Esto quiere decir, que la muestra refleje de la manera más precisa posible las características de la población de la que proviene. Las muestras representativas se caracterizan por aportar buenos datos y para obtenerlas es importante seguir principios estadísticos apropiados, de lo contrario, las conclusiones obtenidas no serán confiables y los resultados no tendrán validez. Cuando una muestra no es representativa de la población se dice que es una *muestra sesgada*.

Un caso famoso de una muestra sesgada

En 1936, en los Estados Unidos, la revista Literary Digest en un intento por predecir los resultados de la elección presidencial, envío diez millones de tarjetas sorteadas a gentes cuyos nombres figuraban en listas de directorios telefónicos, clubes y propietarios de coches. El pronóstico fue

que Roosevelt obtendría el 40.9% de los votos, mientras que su oponente Landon, recibiría el 57%. Sin embargo, unas semanas después, en la elección real, Roosevelt obtuvo el 60.7 % de los votos y ganó la elección presidencial. La equivocación consistió en que la revista seleccionó una muestra que no era representativa de la población de los Estados Unidos, pues se trataba de gente rica que tenía teléfono, coches y asistía a clubes sociales, y se excluyó a la mayoría de la población, que por esas épocas sufría problemas económicos derivados de la Gran Depresión.

Pasos previos a la toma de una muestra

Es muy importante que antes de proceder a seleccionar una muestra, se realice toda una etapa de planeación que involucra, entre otras cosas, la definición de variables que interesa medir y los instrumentos adecuados de medición. Por ejemplo, si estamos interesados en realizar un estudio de opinión mediante una encuesta, será necesario primeramente diseñar un cuestionario con preguntas bien definidas y las opciones de respuesta bien precisas. Si el propósito es medir otro tipo de variables, como el peso, la estatura, el instrumento de medición no sería un cuestionario, sino más bien, sería una báscula o una cinta métrica.

En casos donde no se involucra personas, sino otro tipo de elementos, como objetos o animales, de la misma manera, se requiere definir las variables a medir y contar con instrumentos adecuados. Por ejemplo, si se desea evaluar la calidad del agua en un lugar determinado, las variables de interés podrían ser la cantidad de impurezas por unidad de volumen, porcentaje de cloro u otras cosas. En resumen, aspectos cómo: ¿a quién le pregunto?, ¿cómo pregunto? y ¿a cuántos les pregunto?, deben estar perfectamente definidos antes de la selección de la muestra.

3.3 Muestreos no probabilísticos

Estos muestreos no involucran el azar en la selección de la muestra. Por lo tanto, su principal defecto es que generan muestras con sesgos y los resultados no pueden ser generalizados a toda la población; además, no permiten conocer el error

de muestreo y la confiabilidad de los resultados. Suelen utilizarse con el propósito de tener una idea sobre el comportamiento de un fenómeno, o satisfacer la curiosidad de saber qué opina la gente sobre cierto tema. Veremos a continuación algunos tipos de muestreo que corresponden a esta categoría, los cuales con frecuencia se utilizan en los noticieros televisivos, revistas de entretenimiento y encuestas por internet.

• Muestreo de voluntarios

En este tipo de muestreo, la muestra se auto-selecciona cuando un grupo de personas deciden participar en forma voluntaria ante un cierto llamado. Un ejemplo de ello es cuando una revista, un periódico o un programa de televisión, piden a sus lectores o televidentes contesten alguna pregunta. La muestra está formada sólo por aquellas personas que se toman la molestia de contestarla o porque tienen los medios para ello.

Este tipo de muestreo es evidente que produce muestras sesgadas, pues representa a personas motivadas y con opiniones muy firmes que se toman la molestia de responder a un llamado, o porque esperan recibir un premio, por lo tanto, no es representativa de toda la población.

Un claro ejemplo de ello ocurrió en Estados Unidos, cuando la periodista Ann Landers[1*], preguntó a sus lectores lo siguiente: "Si pudieras retroceder en el tiempo, ¿volverías a tener hijos? Unas semanas después fueron publicados los resultados bajo el siguiente encabezado: *El 70% de los padres dice que no vale la pena tener hijos.* Desde luego, estos resultados no indican que esa sea la opinión de todos los padres estadounidenses, dado que los padres que respondieron, muchos de ellos estaban molestos con sus hijos. Unos meses más tarde, una encuesta de opinión estadísticamente bien diseñada determinó que el 91% de los padres volvería a tener hijos.

Es importante señalar que muchos años antes que se utilizaran los métodos probabilísticos para seleccionar muestras, este tipo de muestreo se utilizaba para conocer la opinión de las personas sobre algún tema de interés. Lo que hacían era enviar cantidades enormes de cuestionarios por correo, con la esperanza de recibir de regreso un gran número de ellos que representaran la opinión de la población. Sin embargo, a pesar de recibir cantidades grandes de respuestas, seguía siendo un muestreo sesgado, pues depende de voluntad de los entrevistados.

• Muestreo por conveniencia

Consiste en seleccionar la muestra por su facilidad de acceso o economía. Por ejemplo,

1 Tomado de *Estadística Aplicada Básica* de David Moore

si se desea realizar un estudio de mercado para conocer la opinión de los consumidores sobre algún producto en particular o su opinión sobre un cierto tema, la muestra se puede tomar acudiendo a lugares donde se sabe de antemano que asisten muchas personas, como un centro comercial, una iglesia o al centro de una ciudad. Sin embargo, las personas que acuden a estos lugares pueden no ser representativas de la población.

Este tipo de muestreo, al igual que el de voluntarios, produce muestras sesgadas, pues las personas que se tomaron en cuenta por facilidad de acceso pueden tener características muy distintas respecto al resto de la población.

• Muestreo por cuotas

Este tipo de muestreo utiliza algunos elementos del muestreo de voluntarios y del muestreo por conveniencia; su diseño es más complejo y por lo que general arroja muestras más representativas, lo que lo convierte en el método más utilizado en la categoría de los muestreos no aleatorios.

Para seleccionar una muestra por cuotas primeramente se divide la población en grupos, utilizando algunas variables con influencia en los resultados (por ejemplo: género, edad, nivel socioeconómico, en el caso de encuestas), enseguida se define una cantidad o cuota de elementos por cada grupo que van a ser sujetos de estudio, cuando es posible se define esta cuota de forma proporcional al tamaño del grupo en la población.

Por ejemplo, si en una población hay 30% mujeres y 70% hombres, y se desea obtener una muestra de 1200 personas, 360 deberán ser mujeres y 840 deberán ser hombres. Por último, se seleccionan cada uno de los elementos de manera voluntaria o por conveniencia hasta completar la cuota de cada grupo. Si un estudio requiere la selección de 100 mujeres y 120 hombres para ser encuestados, se puede salir a la calle y preguntarles su edad y encuestarlas hasta completar la cuota, lo mismo podría hacerse vía telefónica o internet.

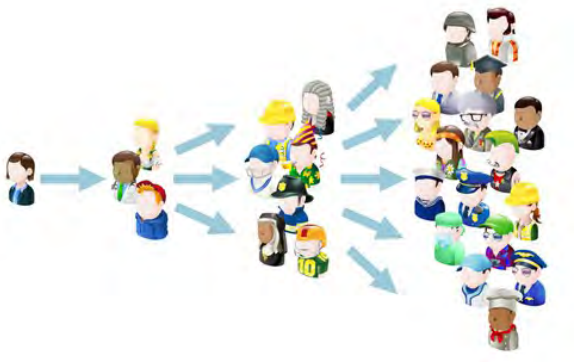
Este muestreo tiene similitud con el muestreo aleatorio estratificado, el cual analizaremos más delante, la diferencia estriba en la selección de los participantes. En el muestreo aleatorio por lo general se dispone de un listado de los participantes (marco muestral) cada uno con probabilidad conocida de ser seleccionado, mientras que en el muestreo por cuotas se van seleccionando los elementos de forma no aleatoria comprobando si cumplen cierto criterio hasta completar la cuota propuesta.

De esta manera, aunque la selección de los elementos no es aleatoria, al menos guarda la misma proporcionalidad de los grupos respecto a la población. La edad y el género son dos variables que con frecuencia se toman como referencia para formar

grupos en muchos estudios estadísticos, por tener influencia en los resultados. Por ejemplo, en un estudio sobre opinión electoral, los hombres y las mujeres pueden tener diferencias por uno u otro candidato, lo mismo entre los más jóvenes y las personas de mayor edad.

• Muestreo de bola de nieve

Este método de muestreo consiste en identificar un grupo inicial de individuos que tienen la característica de interés en la investigación, para que ellos mismos inviten a otros participantes entre sus conocidos siempre que reúnan la característica de interés. Ello permite que la muestra vaya creciendo –de ahí su nombre de bola de nieve- a medida que los participantes seleccionados invitan a sus conocidos. Este método es muy útil para acceder a poblaciones de baja incidencia y a individuos de difícil acceso, por ejemplo, pacientes con una rara condición o enfermedad, personas que coleccionan algún objeto en particular. En estos casos puede resultar más efectivo seleccionar la muestra a través de conocidos que mediante un método aleatorio.



Entre sus ventajas está acceder a poblaciones de difícil acceso, es económico y sencillo, requiere poca planificación. Cómo todos los métodos no aleatorios, no garantiza representatividad de la muestra y no es posible conocer la precisión y confiabilidad de los resultados.

Resumiendo, los muestreos de voluntarios, de conveniencia, de cuotas y bola de nieve, no permiten generalizar sus resultados más allá del ámbito de donde fueron tomados los datos. Los muestreos no probabilísticos tienen la ventaja de la rapidez y el bajo costo, sin embargo, el sesgo en que incurren supera con mucho las ventajas.

3.4 Muestreos aleatorios o probabilísticos

En una muestra compuesta por voluntarios las personas escogen responder, en una muestra por conveniencia y por cuotas es el encuestador quien elige, en una muestra por bola de nieve unos sujetos invitan a otros. Sin embargo, como señalamos anteriormente, estos tipos de muestreo no producen muestras representativas de una población ya que presentan sesgos, por lo que sus resultados son válidos sólo para las personas y el contexto en que fueron tomados. La solución estadística para evitar el sesgo, es dejar que el *azar* determine la muestra.

Una muestra elegida al azar evita que el entrevistador favorezca la elección de algún encuestado, y la auto-selección de los encuestados, como en el muestreo por voluntarios. Cuando se selecciona una muestra al azar se ataca el sesgo, ya que

todos los elementos tienen posibilidades de ser seleccionados, lo cual no sucede en los métodos no aleatorios. Los métodos aleatorios son adecuados cuando se desean realizar inferencias acerca de una población. En general presentan dos grandes ventajas que no tienen otros métodos: la disminución del sesgo y permiten determinar la confiabilidad y el error de muestreo en los resultados de la muestra.

Para seleccionar una muestra aleatoria con frecuencia es importante contar con un listado o una base de datos con los elementos de la población lo más preciso posible. Dicho listado recibe el nombre de *marco muestral*. En muchos casos se dispone de él de forma más o menos precisa. Un ejemplo de marco muestral pueden ser la lista de alumnos en una escuela o universidad, la lista de comunidades en un municipio o estado, la lista de secciones electorales que tiene el Instituto Nacional Electoral. Un ejemplo importante de la precisión que se puede obtener al seleccionar de forma eficiente y rigurosa una muestra aleatoria para predecir características de una población, se muestra en la siguiente tabla que ha sido tomada del Instituto Nacional Electoral.

El cuadro siguiente muestra un comparativo entre los porcentajes reales de participación ciudadana proporcionados por los cómputos distritales y las estimaciones mediante muestreo en las elecciones federales de 2018. Obsérvese las pequeñas diferencias en la mayor parte de los estados. A nivel nacional, la estimación señala un 62.3% de votación y el cómputo distrital reporta un 63.4%. De acuerdo con lo que hemos visto anteriormente, el 62.3% es considerado un *estadístico* porque se obtuvo de una muestra, mientras que el 63.4% se conoce como *parámetro*, ya que representa a la población de votantes.

Entidad Federativa	Cómputos distritales (A)	Estimaciones muestrales (B)	Diferencia (A - B)
Nacional	**63.4**	**62.3**	**1.12**
Aguascalientes	59.4	58.0	1.36
Baja California	52.6	51.3	1.29
Baja California Sur	58.7	56.1	2.60
Campeche	70.0	68.6	1.40
Coahuila	63.6	62.1	1.46
Colima	64.1	62.2	1.94
Chiapas	68.4	68.2	0.20
Chihuahua	54.4	53.1	1.32
Ciudad de México	70.6	70.0	0.55
Durango	57.1	55.7	1.42
Guanajuato	53.2	52.3	0.88
Guerrero	64.1	62.9	1.21
Hidalgo	65.7	65.0	0.74
Jalisco	59.2	58.0	1.21
México	67.9	67.2	0.72
Michoacán	58.5	57.2	1.32
Morelos	67.1	65.5	1.61
Nayarit	57.1	56.6	0.48
Nuevo León	55.8	54.8	0.99
Oaxaca	67.2	65.9	1.32
Puebla	68.3	67.5	0.79
Querétaro	64.7	62.2	2.45
Quintana Roo	60.2	58.3	1.89
San Luis Potosí	63.6	62.7	0.91
Sinaloa	60.5	59.4	1.14
Sonora	51.9	50.7	1.17
Tabasco	71.1	69.9	1.18
Tamaulipas	62.1	60.4	1.73
Tlaxcala	66.4	65.3	1.10
Veracruz	65.9	64.4	1.56
Yucatán	75.4	74.6	0.78
Zacatecas	65.2	64.1	1.11

Porcentajes de participación ciudadana en las elecciones federales de 2018, según fuente, por entidad federativa.

• Muestreo Aleatorio Simple

El muestreo aleatorio simple es el muestreo más común que existe, ya que además de ser un método por sí mismo, sirve de base a otros tipos de muestreo. Su característica principal es que cada elemento de la población objetivo y cada muestra de un tamaño dado, tiene la misma probabilidad de ser seleccionado.

La idea más clara de una muestra aleatoria simple, la proporciona el hecho de seleccionar al azar cierta cantidad de "papelitos idénticos" de una bolsa o canicas del mismo tamaño y textura de una urna o tómbola. A este método, aun sin conocer su nombre hemos recurrido muchas veces cuando realizamos una rifa o sorteo. Los sorteos que realiza la Lotería Nacional y los sorteos de Melate –y en general de cualquier lotería- están basados en muestreo aleatorio simple.

Tabla de dígitos aleatorios

El mecanismo de selección de los elementos o unidades de la población puede consistir en una técnica manual como el caso de la urna o tómbola, -muy usual en el caso de los sorteos, porque permiten ver físicamente los números que conforman el premio-, una tabla de dígitos aleatorios en la que están distribuidos de manera uniforme dígitos del 0 al 1, o bien un programa de computadora para generar números aleatorios con una función de probabilidad uniforme. Los dos primeros mecanismos son tediosos y requieren tiempo para su generación, lo que los hace útiles en poblaciones pequeñas. En el caso de poblaciones grandes es más apropiado generar los números aleatorios mediante un software estadístico u hoja de cálculo.

• Pasos para la selección de una muestra aleatoria simple

1. Identificar la población objetivo.
2. Se identifica o construye un marco muestral de la población.
3. Se asigna un número único a cada elemento de la población, desde 1 hasta N.
4. Se determina el tamaño de muestra que se va a seleccionar.
5. Se selecciona en forma aleatoria la cantidad de elementos señalada de la población.

Los primeros dos pasos consisten en disponer de una lista con todos los elementos de la población perfectamente identificados con un número. En muchas situaciones esto no es posible, lo que hace que el muestreo aleatorio simple no pueda ser utilizado.

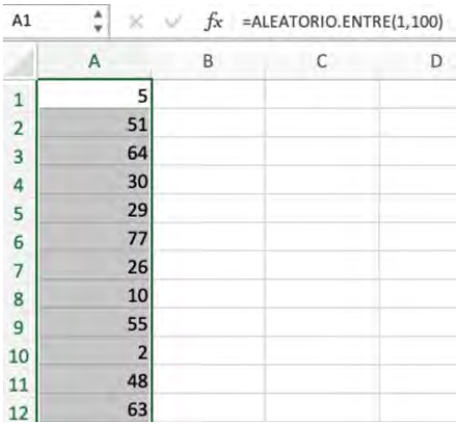
Consideremos la población de las 100 empresas más importante de México en 2017 según datos de la revista Expansión (https://expansion.mx/ranking/las-500-2017).

Tabla. Ventas de las 100 empresas más importantes de México en 2017 (millones de pesos, mdp)

No.	Empresa	Ventas (mdp)	No.	Empresa	Ventas (mdp)
1	Petróleos Mexicanos	1,079,546	51	Cuauhtémoc Moctezuma - Heineken	65,000
2	América Móvil	975,412	52	PepsiCo de México	64,160
3	Wal-Mart de México	532,384	53	Grupo Inbursa	58,591
4	Grupo FEMSA	399,507	54	MetLife México	58,060
5	Comisión Federal de Electricidad	352,106	55	Jabil Circuit de México	56,915
6	General Motors	321,905	56	GNP	54,701
7	Alfa	293,782	57	Grupo Aeroméxico	53,925
8	Fiat Chrysler	273,020	58	Grupo Lala	53,468
9	Grupo Bimbo	252,141	59	Industrias Bachoco	52,020
10	Cemex	250,909	60	Mabe	51,305
11	BBVA Bancomer	215,526	61	Lear Corporation	50,198
12	Nissan Mexicana	200,000	62	Costco de México	50,000
13	Grupo BAL	192,497	63	Grupo HSBC	49,623
14	Coca-Cola FEMSA	177,718	64	Grupo Nestlé México	49,246
15	Ford de México	177,164	65	Altos Hornos de México	48,512
16	Volkswagen	175,147	66	Grupo Sanborns	47,594
17	Grupo Banamex	154,227	67	Johnson Controls México	46,757
18	Grupo México	152,844	68	AT&T México	46,226
19	Organización Soriana	149,522	69	FEMSA División Salud	43,411
20	FEMSA(Oxxo)	137,139	70	Cultiba	43,345
21	Infonavit	128,582	71	SuKarne	42,008
22	Grupo Techint	121,471	72	Fragua Corporativo	40,572
23	Kaluz	120,851	73	P&G México	40,000
24	Grupo Banorte	120,600	74	Grupo Villacero	39,900
25	Grupo Coppel	119,944	75	Grupo Bank of America	39,367
26	Americas Mining Corporation	116,126	76	Grupo Xignux	39,251
27	Honda de México	113,000	77	Daimler México	39,191
28	Samsung Mexico	110,000	78	Alsea	37,702

29	Sam's Club	107,436	79	The Home Depot México	37,400
30	Sigma Alimentos	106,341	80	Autoliv México	37,400
31	Grupo Salinas	105,000	81	Grupo Empresarial Ángeles	37,200
32	El Puerto de Liverpool	100,442	82	Metalsa	37,000
33	Mexichem	100,041	83	Cinépolis	36,900
34	Grupo Financiero Santander México	100,040	84	Casa Ley	36,000
35	Grupo Televisa	96,287	85	Banco Azteca	35,940
36	Magna International México	95,763	86	Kimberly-Clark de México	35,660
37	Grupo Carso	95,188	87	Fresnillo PLC	35,633
38	Arca Continental	93,666	88	AXA Seguros	35,488
39	Alpek	90,192	89	Aeropuertos y Servicios Auxiliares	35,068
40	Grupo Chedraui	88,529	90	Nacional de Drogas	35,000
41	Ternium México	84,262	91	Continental Tire de México	35,000
42	Industrias Peñoles	82,142	92	Sanmina-SCI Systems de México	34,962
43	Grupo Elektra	81,242	93	DeAcero	34,900
44	Nemak	79,244	94	Banobras	33,988
45	Grupo Modelo	77,851	95	Arcelor Mittal Mexico	33,772
46	Toyota Motor Sales de México	73,600	96	Grupo Kuo	33,627
47	LG Electronics México	72,000	97	Iberdrola México	32,445
48	Gruma	68,206	98	Scotiabank Inverlat	32,250
49	Flextronics Manufacturing México	68,170	99	Televisa Telecomunicaciones	31,892
50	BMW Group México	68,000	100	Grupo Palacio de Hierro	31,160

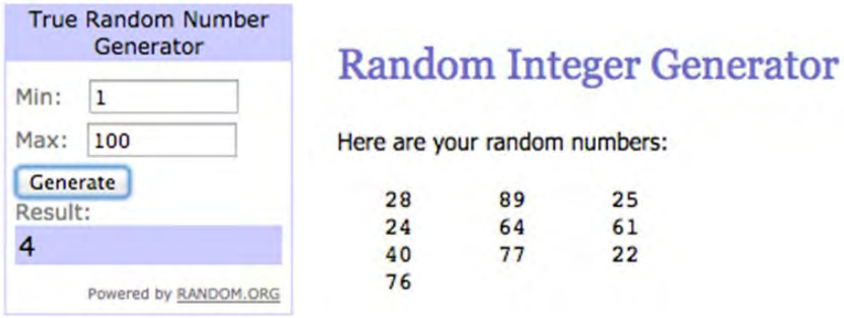
Los *elementos* o *unidades* de la población son las empresas, y el marco muestral lo constituye la lista de empresas con su número identificador (del 1 al 100). Se desea determinar una muestra aleatoria simple de 10 empresas para revisar a detalle una serie de variables que las caracteriza. Utilicemos un generador de número aleatorios para definir las empresas que serán parte de la muestra. Veamos cómo se realiza el proceso anterior en Excel.



Generador de números aleatorios con Excel

Con la función *Aleatorio.Entre (1,100)* generamos 15 números aleatorios (por si acaso algunos se repiten para considerarlos una sola vez) de los cuales seleccionamos los primeros 10, sin incluir repetidos. Las empresas seleccionadas son:

5	Comisión Federal de Electricidad
51	Cuauhtémoc Moctezuma - Heineken
64	Grupo Nestlé México
30	Sigma Alimentos
29	Sam's Club
77	Daimler México
26	Americas Mining Corporation
10	Cemex
55	Jabil Circuit de México
2	América Móvil



www.random.org

Otra opción a la que se puede recurrir para generar números aleatorios, son sitios web especializados en ello como es el caso de www.random.org. Algunos aficionados a los juegos de azar utilizan estas páginas para seleccionar los números que compran en una lotería.

• Muestreo Aleatorio Estratificado

El muestreo aleatorio estratificado consiste en dividir a la población en grupos o estratos lo más homogéneos entre sí, para posteriormente seleccionar muestras aleatorias simples de cada estrato. La combinación de las muestras simples de cada estrato forma la muestra completa.

El muestreo aleatorio estratificado es elegible cuando los elementos de la población presentan mucha variabilidad en las características que interesa medir. Por ejemplo, en algunos estudios puede existir diferencia notable sobre la opinión de un tema, según el grupo de edad, el nivel socioeconómico o el género de la persona. En estos casos es razonable dividir a la población en estratos antes de tomar la muestra. Cada grupo de edad puede ser un estrato, hombres y mujeres forman dos estratos. En el caso de las

empresas se pueden formar tres estratos: pequeñas, medianas y grandes.

Cuando es posible identificar y clasificar a los elementos de una población en estratos o categorías de las variables que más inciden en lo que se está investigando, se puede mejorar mucho la representatividad de la muestra. Por lo general el muestreo estratificado es más eficiente que el muestreo aleatorio simple, ya que garantiza que queden representados en la muestra elementos de la población con características diversas, lo cual no siempre ocurre en el muestreo aleatorio simple.

Pasos para la selección de una muestra aleatoria estratificada

1. Definir la población objetivo
2. Identificar las variables de estratificación y el número de estratos.
3. Identificar o construir un marco muestral que incluya información sobre las variables de estratificación para cada elemento en la población.
4. Dividir el marco muestral en estratos, categorías de las variables de estratificación, creando un marco muestral para cada estrato.
5. Determinar el tamaño de muestra de cada estrato.
6. Seleccionar en forma aleatoria la cantidad de elementos de cada estrato.

La cantidad de elementos seleccionados en cada estrato se puede determinar siguiendo uno de dos criterios: asignación proporcional o asignación no proporcional. En el *criterio de asignación proporcional* la cantidad de elementos seleccionados en cada estrato debe ser proporcional al tamaño relativo del estrato en la población. Es decir, si un estrato representa el 30% de la población, debe aportar el 30% de los elementos de la muestra total. La muestra total obtenida de la combinación de las muestras de cada estrato se le llama *muestra auto ponderada*. En el *criterio de asignación no proporcional*, la cantidad de elementos muestreados de cada estrato no es proporcional al tamaño del estrato respecto de la población.

Ejemplo:

En una comunidad hay 1500 personas mayores de edad, y se va a llevar a cabo un estudio para conocer su opinión sobre unas obras que se desean realizar, y que, si bien traerán algunos beneficios, también alterarán las condiciones de vida que hasta el momento han llevado sus habitantes. Se ha determinado que una muestra de 150 personas sería adecuada para el estudio. Para ello se van a formar tres estratos de acuerdo con el grupo de edad; 18-29 años, 30 a 49 años y 50 años o más. Se dispone de un listado de todos los ciudadanos (marco muestral) y además del último censo se sabe que 800 personas caen en el primer rango de edad, 300 en el segundo y 400 en el tercero. Asignar proporcionalmente al tamaño de cada estrato, los elementos de la muestra.

Grupo de edad	Población		Muestra proporcional	
	Frecuencia	Porcentaje	Frecuencia	Porcentaje
18-29 años	800	53.3%	80	53.3%
30-49 años	300	20.0%	30	20.0%
50 años o mas	400	26.7%	40	26.7%
Total	**1,500**	**100%**	**150**	**100%**

Lo que hemos hecho hasta ahora es calcular el tamaño de las muestras aleatorias simples que se deben tomar de cada estrato aplicando el criterio de asignación proporcional (pasos del 1 al 5). Lo que procede ahora es aplicar los criterios del muestreo aleatorio simple para seleccionar los elementos que serán parte del estudio, ya sea utilizando una tabla de dígitos aleatorios o un programa de computadora para generar números aleatorios.

• Muestreo Aleatorio Sistemático

En muchas situaciones es posible disponer de los elementos de una población colocados ordenadamente en una fila o anotados en una lista (marco muestral), de tal forma que se pueden formar intervalos de igual tamaño entre ellos. El muestreo sistemático consiste en seleccionar un elemento de cada intervalo para constituir la muestra. El primer elemento se selecciona en forma aleatoria simple entre los elementos del primer intervalo, y los subsecuentes elementos se seleccionan a una distancia fija o intervalo sistemático (igual a la amplitud del intervalo) del primero hasta completar el total de elementos de la muestra. Si la población tiene **N** elementos y se va a seleccionar una muestra de tamaño **n**, la amplitud de cada intervalo **k,** se determina mediante la expresión:

$$k = \frac{N}{n}$$

Pasos para la selección de una muestra aleatoria sistemática

1. Definir la población objetivo
2. Determinar el tamaño de muestra (n)
3. Identificar o construir un marco muestral para la población objetivo
4. Determinar el número de elementos en el marco muestral (N).
5. Calcular la amplitud del intervalo (k)
6. Seleccionar por muestreo aleatorio simple un elemento del primer intervalo.
7. Seleccionar los elementos subsecuentes sumando al primer elemento la amplitud del intervalo hasta completar el tamaño de muestra.

Para fijar ideas, imaginemos que tenemos una población de 60 personas colocadas en una fila y se decide seleccionar una muestra aleatoria sistemática de 12

personas. El tamaño de cada intervalo es igual a 5, dado que $k=\frac{60}{12}=5$. Un muestreo aleatorio simple entre los 5 elementos del primer intervalo arroja el número 2, con lo cual se genera el primer elemento de la muestra, los demás elementos se obtendrán de 5 en 5 a partir del elemento 2. La muestra queda integrada por las personas: 2, 7, 12, 17, 22, 27, 32, 37, 42, 47, 52 y 57 (ver figura).

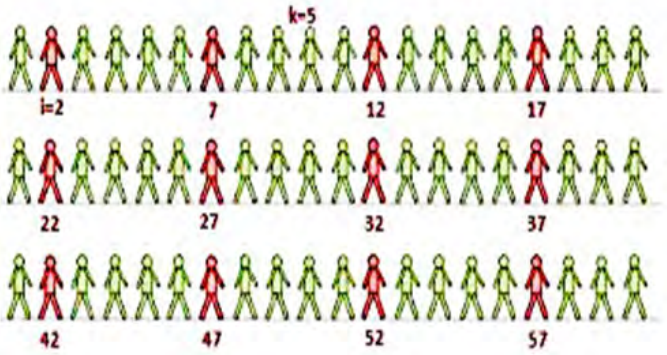


Figura. Esquema del muestreo aleatorio sistemático

Este tipo de muestreo es muy útil cuando se dispone de marcos muestrales bien definidos, como puede ser la lista de estudiantes inscritos en una escuela, el padrón de contribuyentes de un estado, la lista de colonias en una ciudad, el padrón de secciones electorales de un país, o la línea de artículos que salen en una línea de producción de una fábrica. La mayoría de las encuestas electorales realizadas en México, utilizan muestreo aleatorio sistemático para seleccionar las secciones electorales donde se aplicarán las encuestas, dado que es un marco muestral muy preciso y que cuenta con un mapa de la zona que comprende la sección electoral.

• Muestreo por conglomerados

Este tipo de muestreo consiste en identificar grupos o conglomerados que ya existen en la población, muchas veces de manera natural (por ejemplo: municipios, universidades, escuelas, empresas, edificios). Las unidades de muestreo son los conglomerados, y las unidades a estudiar son los individuos o elementos que integran dichos conglomerados. En los métodos anteriores las unidades de muestreo y las unidades a estudiar eran la misma cosa, pero no este tipo de muestreo son diferentes.

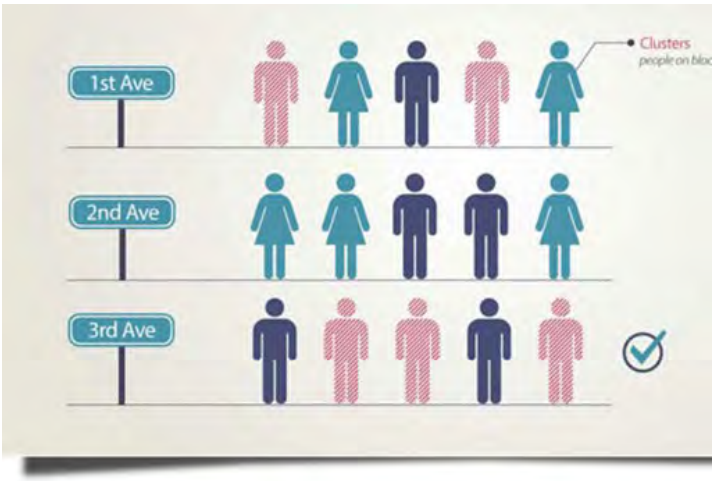


Figura: Tres conglomerados (calles)

El muestreo por conglomerados es útil cuando no se dispone o resulta muy difícil construir un marco muestral para la población objetivo, o bien la población está muy dispersa geográficamente. Este método de muestreo puede reducir los costos considerablemente sin sacrificar demasiado la precisión de los resultados, siempre que los conglomerados sean lo más homogéneos posible entre sí en la variable que se desea estudiar.

Para fijar ideas consideremos que, respecto a una universidad que tiene 50 facultades distribuidas en varias ciudades, la aplicación de un muestreo aleatorio simple, estratificado o sistemático puede resultar costoso, pues se requiere viajar

a cada una de las ciudades a estudiar a los elementos que quedaron en la muestra. Si las facultades tienen características similares, cada facultad se puede considerar un conglomerado, de tal forma que se podría hacer una selección aleatoria de unas cuantas facultades en lugar de ir a todas. En este caso las facultades serían las unidades de muestreo y los estudiantes las unidades a estudiar.

Pasos para la selección de una Muestra Aleatoria de Conglomerados

1. Definir la población objetivo
2. Determinar el tamaño de muestra.
3. Identificar o construir un marco muestral para la población objetivo
4. Determinar el número de conglomerados a ser seleccionados.
5. Seleccionar por muestreo aleatorio simple el número de conglomerados deseado.

Una vez seleccionados los conglomerados podemos estudiar a todos los sujetos que los integran, o bien, realizar un nuevo proceso de muestreo dentro del conglomerado mediante algunos de los métodos aleatorios. En este último caso estaríamos hablando de un muestreo en dos etapas o *bietápico.* La primera etapa consiste en la selección de los conglomerados, la segunda en la selección de individuos dentro del conglomerado.

3.5 Algunos errores que se pueden generar en las encuestas por muestreo

En los apartados anteriores hemos señalado que los elementos de estudio en estadística, pueden ser personas, animales, manzanas, municipios o cualquier objeto que sea susceptible de ser medido. En todos los casos se pueden cometer errores de selección o medición de los elementos. Sin embargo, cuando los elementos son seres humanos, como en el caso de las encuestas, el riesgo de cometer errores se incrementa notablemente. Es decir, no basta con que el muestreo sea aleatorio para producir buenos datos.

Se requiere, además, cuidar una serie de aspectos en la recopilación de los datos, si se quiere obtener información precisa. La encuestadora Consulta Mitofsky (http://www.consulta.mx/) lo advierte en la ficha metodológica de sus estudios: "En los estudios de opinión pública, además del error muestral, se debe considerar que pueden existir otros errores ocasionados por el fraseo de las preguntas y las incidencias en el trabajo de campo".

Veremos a continuación algunos de los principales errores y sesgos que se cometen cuando se recopila información de encuestas, dado que es una de las grandes aplicaciones que está teniendo la estadística actualmente.

• Falta de cobertura

Para realizar el muestreo es importante contar con una lista completa y precisa de

la población. Sin embargo, esto no siempre es posible, puede entonces que algunos elementos o grupos de la población no sean tomados en cuenta en el proceso de selección.

Un ejemplo sencillo de la falta de cobertura es cuando se realiza una encuesta por teléfono o internet, ya que sólo da oportunidad de participar a quienes cuentan con dichos servicios y deja fuera a quienes no cuentan con él. El muestreo de conveniencia y el muestreo de voluntarios siempre incurren en este tipo de sesgo.

• No respuesta

Esta fuente de sesgo ocurre cuando un individuo seleccionado no puede ser contactado o no quiere colaborar. Lo que usualmente se hace para reponer las no repuestas es localizar gente de la misma área para que responda. Cuando la gente contactada difiere en alguna característica de la seleccionada originalmente se puede generar un cierto sesgo.

• Sesgo de respuesta

Los encuestados pueden mentir, especialmente cuando se les pregunta sobre temas delicados o considerados tabú, como podrían ser cuestiones ilegales, pago de impuestos, cuestiones relacionadas con sexo, discriminación entre otras. El encuestador debe ser hábil para plantear las preguntas y conseguir las respuestas en forma adecuada.

• Efecto del redactado

El redactado de las preguntas puede influir enormemente en las respuestas de los encuestados. La confusión o la falta de comprensión de términos pueden introducir un sesgo muy fuerte. Lo que frecuentemente se hace para reducir o eliminar este efecto es poner a prueba el cuestionario antes de la recopilación real de los datos y ver cómo interpretan y responden las preguntas los encuestados. Otra alternativa adicional es capacitar a los encuestadores.

Sin embargo, los problemas de redacción no solamente se dan a partir de la ambigüedad de la pregunta, sino por algún comentario previo, o simplemente la redacción de esta. Por ejemplo, suponga que se desea conocer la opinión de la gente sobre el matrimonio homosexual. Analice la siguiente pregunta: ¿Está usted de acuerdo con que las parejas homosexuales, quienes se han desviado de la condición natural del humano y la doctrina de la iglesia, puedan contraer matrimonio legal ante la ley?

Si esta pregunta se la hacemos a una persona católica u homofóbica, se puede predecir la respuesta. El texto subrayado sesga la respuesta de las personas, porque intenta hacer un llamado de conciencia. La pregunta debe ser: ¿Está usted de acuerdo con la legalización del matrimonio homosexual? De esta manera se debe tener presente

que la pregunta debe ser clara, concisa, no debe tener comentarios adicionales propios del encuestador, con la finalidad de evitar el sesgo y que la información recabada sea veraz.

3.5 ¿Cuál es el tamaño adecuado de una muestra?

Una pregunta fundamental cuando se utiliza una muestra para recopilar los datos de una investigación es: ¿qué tamaño de muestra se debe seleccionar para obtener conclusiones válidas sobre una población?, tres factores influyen en la decisión: el *tamaño de la población*, el *error de muestreo* que está dispuesto a aceptar y la *confiabilidad* que se desea en el método de muestreo.

En general, si se desea mayor precisión en los resultados del muestreo se debe aumentar el tamaño de muestra, en el extremo que se desee una precisión absoluta, el tamaño de la muestra debe ser igual a la población. Sin embargo, muestras muy grandes tienen mayor costo y requieren mayor tiempo para la recopilación de los datos.

En el caso del muestreo aleatorio, la probabilidad nos ayuda a determinar el tamaño de muestra requerido para un error y una confiabilidad especificada. En los próximos capítulos abordaremos el cálculo del tamaño de muestra.

Una propiedad importante del muestreo que lo vuelve muy útil consiste en que, en poblaciones muy grandes, el tamaño de muestra que se requiere representa un pequeño porcentaje de la población, es decir, si una población consta de 1 millón de elementos, bastaría estudiar menos de 1000 elementos para obtener una buena precisión en los resultados, incluso si la población estuviera integrada por millones de elementos. A medida que aumenta el tamaño de la población la muestra va creciendo, pero no de forma proporcional y tiende a estabilizarse. Esta propiedad se puede comprender con ayuda de la teoría de la probabilidad.

3.6 Algunas creencias erróneas sobre el muestreo

1. *Si un proceso verdaderamente aleatorio es utilizado para seleccionar una muestra de la población, la muestra resultante será como la población, pero más pequeña. En otras palabras, una muestra aleatoria es como una réplica en miniatura de la población.*

Si realmente una muestra aleatoria fuera una réplica en miniatura de la población, los resultados de la muestra serían exactamente iguales a los resultados de la población, y se podrían utilizar para estimar sin error valores poblacionales. Por ejemplo, la proporción de fumadores en una muestra sería igual a la proporción de fumadores en la población. Quienes presentan esta creencia son indiferentes al error de muestreo que existe como consecuencia de la variabilidad muestral. La inferencia estadística

con ayuda de la probabilidad es lo que determina el margen de error que se puede presentar en una estimación de parámetros poblacionales.

2. *El tamaño de la muestra debe ser proporcional al tamaño de la población. Poblaciones más grandes requieren muestras más grandes.*

Muchas personas consideran incorrectamente que una población grande requiere de una muestra muy grande. Si bien es cierto que poblaciones más grandes requieren muestras grandes, esto no es proporcional y para un determinado tamaño de la población el tamaño de muestra se estabiliza, es decir, no tiene sentido incrementar el tamaño porque no se aumenta significativamente la precisión, y como una muestra tiene costo económico y en tiempo, no tiene sentido incrementarla. Un investigador que tiene esta concepción errónea tenderá a recolectar muestras más grandes de lo requerido, por otra parte, una persona podría malinterpretar un estudio que involucra una muestra pequeña respecto al tamaño de la población.

Consideremos el caso de dos encuestas de opinión realizadas en México en 2017. La empresa Parametría realizó un estudio de opinión sobre la Ley de Seguridad Interior aprobada por la Cámara de Diputados y el Senado. La población objetivo eran las personas mayores de 18 años con credencial para votar residiendo en la vivienda seleccionada, dicha población la conforman millones de mexicanos. Se utilizó un tamaño de muestra de 800 personas y se reporta una confianza estadística de 95% y un máximo margen de error de 3.5% en los resultados de la encuesta. http://www.parametria.com.mx/carta_parametrica.php?cp=5015

Consulta Mitofksy realizó una Encuesta Nacional sobre Percepción de Inseguridad Ciudadana en México y consideró como población objetivo a los mexicanos mayores de 18 años residentes en el territorio nacional. Se utilizó un tamaño de muestra de 1000 personas y se reporta una confianza estadística de 95% con un máximo error de muestreo de 3.1%. Obsérvese en ambos estudios que la muestra es muy pequeña respecto al gran tamaño de la población, sin embargo, la confianza estadística en los resultados es alta y el margen de error se considera aceptable en teoría de encuestas.

3.7 Ventajas y desventajas del muestreo respecto al censo (población)

Ventajas:

- Necesitamos estudiar menos individuos, necesitamos menos recursos (tiempo y dinero).
- La manipulación de datos es mucho más simple. Si con una muestra de 1000 personas tengo suficiente, ¿para qué quiero analizar millones de datos?

Desventajas:

- Introducimos error (controlado) en el resultado, debido a la propia naturaleza del muestreo y a la necesidad de generalizar resultados.
- Tenemos el riesgo de introducir sesgos debido a una mala selección de la muestra. Por ejemplo, si la forma en que selecciono individuos para la muestra no es aleatoria, mis resultados pueden verse seriamente afectados.

3.8 Codificación, depuración y captura de los datos en un software estadístico

La parte final de la recolección de los datos consiste en el proceso de captura de los mismos en un software estadístico o una hoja de cálculo. En muchas ocasiones es necesario establecer códigos para los datos con el fin de facilitar su captura, sobre todo si se trata de muchos datos. Por ejemplo, la variable *género* en una encuesta produce los datos *masculino* y *femenino*, que podrían ser codificados como 1 y 2 respectivamente. La depuración es necesaria cuando hay cuestionarios que no son respondidos en forma completa o que las personas eligieron dos respuestas en lugar de una, por mencionar algunos casos. Estos cuestionarios la mayoría de las veces se eliminan y no se capturan.

Actividad de aprendizaje

El periódico El Economista en su edición del 28 de julio de 2014, con base en información de América Economía (http://rankings.americaeconomia.com/2012/las-500-empresas-mas-grandes-de-america-latina/ranking-500-america-latina-1-50.php) publicó la lista de las 500 principales empresas de América Latina, con datos sobre sus ventas, sector de actividad, utilidades, activos, país de procedencia, entre otros. Por cuestiones de espacio reproducimos a continuación la lista de las primeras 20 empresas. Si se revisa la lista de las 500 empresas puede observarse que se trata de una población muy heterogénea en las variables consideradas, pues hay empresas muy grandes, grandes, medianas y pequeñas.

Se desea seleccionar una muestra aleatoria de la población de empresas. Discute con tus compañeros lo siguiente:

a) ¿Cuál sería el inconveniente de utilizar una muestra aleatoria simple?

b) ¿Si se utilizará un muestreo aleatorio estratificado ¿qué variables podrías considerar para realizar los estratos?

EMPRESA	PAÍS	SECTOR / RUBRO	VENTAS 2011 US$ Millones	VENTAS 2010 US$ Millones	VARIACIÓN VENTAS 11/10 (%)	UTILIDAD NETA 2011 US$ Millones	UTILIDAD NETA 2010 US$ Millones	ACTIVO TOTAL 2011 US$ Millones
PETROBRAS	BRA	Petróleo/Gas	130,171.7	128,000.0	1.7	17,759.4	21,119.5	319,410.4
PDVSA	VEN	Petróleo/Gas	124,754.0	94,929.0	31.4	4,496.0	3,202.0	182,154.0
PEMEX	MÉX	Petróleo/Gas	111,734.6	103,814.2	7.6	-6,559.1	-3,843.3	109,936.2
VALE	BRA	Minería	55,014.1	49,949.0	10.1	20,158.7	18,047.1	128,896.0
AMÉRICA MÓVIL	MÉX	Telecomunicaciones	47,700.1	49,220.7	-3.1	5,940.3	7,378.6	67,797.8
PETROBRAS DISTRIBUIDORA	BRA	Petróleo/Gas	39,654.0	39,655.8	-0.0	675.0	750.0	9,086.0
ODEBRECHT	BRA	Multisector	33,659.0	28,203.3	19.3	24.0	1,486.0	51,124.0
ECOPETROL	COL	Petróleo/Gas	33,194.8	21,610.7	53.6	7,801.1	4,194.3	46,585.9
JBS FRIBOI	BRA	Agroindustria	32,944.2	33,042.7	-0.3	-40.4	-181.7	25,275.0
WAL-MART DE MÉXICO Y CENTROAMÉRICA	MÉX	Comercio	27,309.8	27,195.8	0.4	1,595.5	1,583.1	16,133.7
ULTRAPAR	BRA	Petróleo/Gas	25,941.6	25,496.2	1.7	452.5	459.3	7,326.3
CBD - GRUPO PÃƒO DE AÃ‡ÚCAR	BRA	Comercio	24,839.8	19,260.4	29.0	382.9	433.6	18,002.5
TECHINT	ARG	Siderurgia/Metalurgia	24,105.0	19,092.0	26.3	0.0	0.0	31,364.0
IRANGA PRODUTOS DE PETRÓLEO	BRA	Petróleo/Gas	22,461.0	21,795.5	3.1	356.0	312.0	4,084.0
EMPRESAS COPEC	CHI	Multisector	21,132.0	12,149.8	73.9	932.7	1,013.8	20,094.9
COMISIÓN FEDERAL DE ELECTRICIDAD	MÉX	Energía Eléctrica	20,931.1	20,601.3	1.6	-1,230.9	65.5	64,987.0
GERDAU	BRA	Siderurgia/Metalurgia	18,875.6	18,841.2	0.2	1,069.3	1,285.9	26,645.6
BRASKEM	BRA	Petroquímica	17,686.4	15,301.2	15.6	-280.0	1,137.5	19,913.7
ELETROBRÁS	BRA	Energía Eléctrica	17,625.2	17,893.8	-1.5	1,989.9	1,349.1	86,972.2
CODELCO	CHI	Minería	17,515.3	16,065.9	9.0	2,055.4	1,876.3	20,834.9

Tabla: Principales empresas de América Latina

Para tu reflexión

Entre las primeras investigaciones sobre comprensión del muestreo destacan las realizadas por los psicólogos Daniel Kahneman y Amos Tversky, quienes reportan que muchas personas evalúan la probabilidad de una muestra basando sus juicios en la similitud de la muestra con la población, aunque la muestra sea pequeña. Esta idea es muy frecuente, incluso entre personas con formación estadística y se denomina *heurística de representatividad*; como consecuencia se genera un sesgo conocido como *creencia en la ley de los pequeños números*. En resumen, las personas que hacen uso de esta heurística consideran que, aunque la muestra sea pequeña, debe reflejar las características de la población.

El muestreo en apariencia es una idea sencilla, pues una vez seleccionada la muestra se procede al cálculo de resúmenes numéricos de los datos, los cuales nos informan con cierta precisión sobre características de la población de la que fue extraída la muestra. Sin embargo, una propiedad intrínseca del muestreo es la variabilidad, de tal forma que es posible que los resultados que aporta una muestra puedan estar alejados de las características de la población. Es decir, una muestra puede no ser representativa de

la población, aunque sea seleccionada al azar. En resumen, en una muestra aleatoria se conjugan las ideas de *representatividad* y *variabilidad*, dos ideas complejas que están en la base de la comprensión de los métodos de inferencia estadística, mismas que constituyen una barrera conceptual para muchos estudiantes.

El muestreo es un concepto fundamental en estadística que debe ser objeto de enseñanza desde los niveles escolares básicos hasta el nivel universitario, incrementando gradualmente su nivel de formalización.

Evaluación del capítulo

1. En una encuesta que apareció en la cuenta de Twitter del Senado de México en el marco de la entrada del T-MEC, se hizo una pregunta con tres opciones, obteniendo en un momento dado, los resultados que se muestran a continuación:

Este tipo de muestreo corresponde a un muestreo de:

a) Voluntarios
b) Conveniencia
c) Aleatorio simple

2. Es un método de muestreo que identifica a un grupo inicial de individuos que tienen una característica en común que interesa investigar, la cual por lo general es una característica rara en la población. Los individuos del grupo inicial invitan a otros participantes que tengan la misma característica. La muestra va creciendo a medida que los participantes seleccionados invitan a sus conocidos.

a) Muestreo estratificado
b) Muestreo de voluntarios
c) Muestreo "bola de nieve"

3. Explica cuál es la diferencia entre un método de muestreo probabilístico y un muestreo no probabilístico.

4. Menciona tres errores que se pueden cometer cuando se recopila información a través de encuestas por muestreo.

5. Ordena los pasos a seguir en la selección de una muestra aleatoria simple

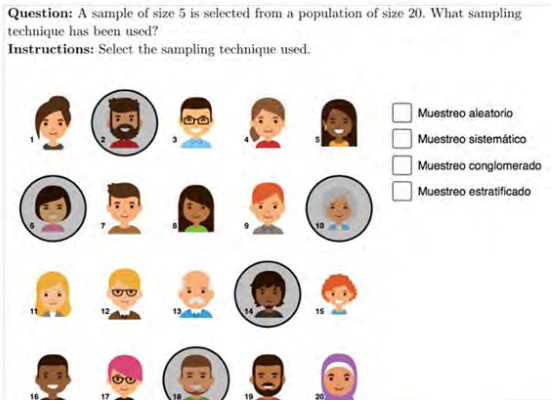
- Se determina el tamaño de muestra que se va a seleccionar.
- Se asigna un número único a cada elemento de la población, desde 1 hasta N.
- Se selecciona en forma aleatoria la cantidad de elementos señalada de la población.
- Identificar la población objetivo.
- Se identifica o construye un marco muestral de la población.

6. Cuando los datos se recopilan de una población considerando cada uno de sus elementos. El proceso se denomina:

- Muestreo
- Censo
- Experimento

7. En la siguiente figura se muestran personas (encerradas en un círculo) que fueron seleccionadas por muestreo aleatorio. Analiza la figura y selecciona el tipo de muestreo utilizado.

8. Explica por qué un muestreo no probabilístico puede producir sesgos en la selección de los datos.



Bibliografía recomendada

- ¿Qué es el muestreo y por qué funciona?
  https://www.netquest.com/blog/es/blog/es/muestreo-que-es-porque-funciona
- Tipos de muestreo (applets)
  https://www.geogebra.org/m/txbjqn47
- Sobre la población y muestra en poblaciones empíricas
  https://cuedespyd.hypotheses.org/2353
- Tipos de muestreo para investigaciones sociales
  https://www.questionpro.com/blog/es/tipos-de-muestreo-para-investigaciones-sociales/
- ¿Cómo trabaja el muestreo aleatorio?
  https://www.pewresearch.org/fact-tank/2017/05/12/methods-101-random-sampling/
- ¿Qué son las encuestas no probabilísticas?
  https://www.pewresearch.org/fact-tank/2018/08/06/what-are-nonprobability-surveys/

Capitulo 4

Organización, presentación y visualización de los datos

La excelencia en gráficas estadísticas consiste en ideas complejas comunicadas con claridad, precisión y eficiencia.

Edward Tufte

4.1 Introducción

Una vez que los datos han sido recolectados, el siguiente paso consiste en organizarlos, resumirlos y presentarlos para visualizar su comportamiento, en forma de tendencias, centralidad, variabilidad y agrupamientos. La organización, resumen y presentación de los datos, por mucho tiempo se ha conocido como *estadística descriptiva*. Con el surgimiento de la tecnología computacional se han hecho posible análisis más elaborados que van más allá de la descripción de los datos, por lo cual resulta más apropiado el nombre de *análisis de datos*, o más específicamente como *análisis exploratorio de datos*.

Sin pérdida de generalidad, un análisis básico de los datos involucra tres actividades:

1. Construcción de *representaciones tabulares (distribuciones de frecuencias)*
2. Construcción de *representaciones gráficas*
3. Cálculo de *medidas descriptivas*

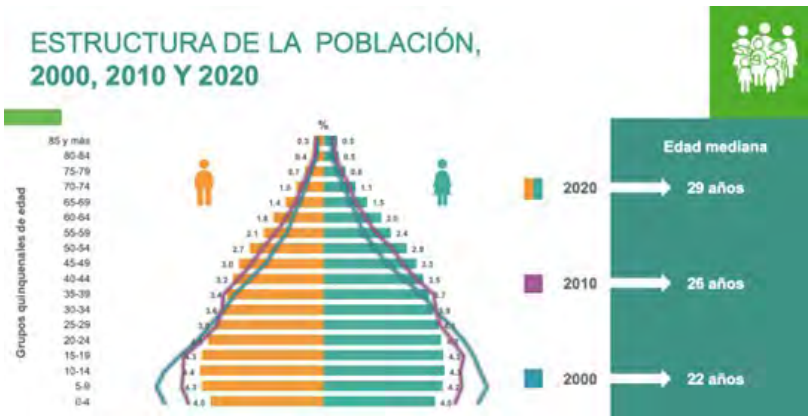
En este capítulo abordaremos los primeros dos métodos de análisis; el cálculo de medidas descriptivas será tema del próximo capítulo.

Una mirada a los diferentes bancos de datos y publicaciones del Instituto Nacional de Estadística y Geografía (INEGI), revela la importancia que tienen las representaciones tabulares y las gráficas para resumir y presentar información estadística obtenida de censos y muestreos (ver gráfica tabla 1 y gráfica 1).

Tabla 1: Población de México por edad y sexo 2020

Grupo quinquenal de edad	2020	
	Hombres	Mujeres
Total	61,473,390	64,540,634
0 a 4 años	5,077,482	4,969,883
5 a 9 años	5,453,091	5,311,288
10 a 14 años	5,554,260	5,389,280
15 a 19 años	5,462,150	5,344,540
20 a 24 años	5,165,884	5,256,211
25 a 29 años	4,861,404	5,131,597
30 a 34 años	4,527,726	4,893,101
35 a 39 años	4,331,530	4,688,746
40 a 44 años	4,062,304	4,441,282
45 a 49 años	3,812,344	4,130,069
50 a 54 años	3,332,163	3,705,369
55 a 59 años	2,692,976	3,002,982
60 a 64 años	2,257,862	2,563,200
65 a 69 años	1,706,850	1,938,227
70 a 74 años	1,233,492	1,413,848
75 a 79 años	847,898	966,684
80 a 84 años	523,812	651,552
85 a 89 años	283,351	375,894
90 a 94 años	107,358	159,448
95 a 99 años	36,615	58,590
100 años y más	6,644	11,651
No especificado	136,194	137,192

Una revisión de periódicos y otros medios de comunicación escrita, muestra el amplio uso que tienen estos métodos en el análisis y presentación de los datos. Es por ello que la habilidad para interpretar la información presentada por medio de tablas y gráficas se ha convertido en una importante competencia estadística en la sociedad moderna, denominada *alfabetización estadística*.



Gráfica 1: Población de México por edad y sexo (2015)

4.2 Un problema introductorio

El Sistema de Monitoreo Atmosférico de la Ciudad de México realiza a cada hora del día mediciones sobre diversos contaminantes en diferentes zonas de la ciudad. La unidad de medida de la contaminación se conoce y expresa como IMECAS (Índice Metropolitano de la Calidad del Aire). La calidad del aire está en función de la cantidad de IMECAS como se describe en el Cuadro 1:

Cuadro 1: Interpretación del índice de calidad del aire

Índice de calidad del aire	Condición
0-50	Buena
51-100	Regular
101-150	Mala
151-200	Muy mala
Mayor a 200	Extremadamente mala

Los datos que se muestran en el cuadro 2 representan los valores máximos de contaminación por ozono que se presentaron en cada una de las zonas de la ciudad durante el mes de diciembre de 2017.

Cuadro 2: Imecas Máximos Diarios de Ozono por Zonas
Diciembre 2017

Día	Zonas				
	Noroeste	Noreste	Centro	Suroeste	Sureste
1	46	63	67	65	73
2	63	115	101	108	94
3	106	104	111	118	112
4	106	117	115	119	120
5	80	80	103	102	107
6	63	86	80	104	113
7	100	65	49	51	45
8	33	34	35	31	37
9	36	44	46	43	44
10	42	46	55	46	47
11	65	82	48	48	59
12	44	43	46	45	45

13	102	108	109	107	104
14	117	102	114	117	110
15	76	45	49	53	41
16	27	29	29	29	33
17	41	47	51	65	94
18	113	100	106	112	100
19	104	92	104	108	86
20	86	100	80	105	104
21	92	96	86	94	107
22	98	82	80	102	67
23	73	80	96	105	103
24	78	103	110	94	117
25	71	57	92	73	88
26	50	63	82	76	104
27	84	73	104	96	102
28	76	78	69	90	88
29	63	57	59	67	63
30	80	88	82	94	86
31	103	105	108	104	103

Fuente: http://www.aire.cdmx.gob.mx/default.php?opc='aqBjnmU='

El **problema** que está detrás de los datos consiste en investigar el comportamiento de los niveles de contaminación por ozono que se presentaron en el mes de diciembre de 2017 en la Ciudad de México. Algunas **preguntas** que podrían ser planteadas en el contexto del problema son las siguientes:

1. ¿Cómo se comportan los niveles de contaminación por ozono en cada una de las zonas?
2. ¿Qué porcentaje de días del mes se tuvo una buena calidad del aire en la zona Noroeste?
3. ¿Cuál es la zona de la ciudad de México con mayor nivel de contaminación por ozono en el período de tiempo considerado?

El **plan de recolección de los datos** está definido por la norma para medir la calidad del aire. Existe un instrumento de medición que registra la cantidad de IMECAS en la zona cada hora del día, de tal forma se tienen 24 mediciones diarias por cada zona. Se selecciona el valor máximo registrado en un día por cada zona y se registra en el cuadro **1.** Daremos respuesta a las preguntas anteriores utilizando los métodos de análisis de

datos que veremos en este capítulo: distribuciones de frecuencias y representaciones gráficas.

4.3 Distribuciones de frecuencias

Una distribución de frecuencias es una representación en forma de tabla que muestra la frecuencia con la que ocurren los datos en las categorías o intervalos en que se ha dividido la variable de estudio. En su proceso de construcción se pueden identificar tres etapas:

1. Definir los intervalos o categorías de la variable.
2. Determinar las frecuencias absolutas de cada intervalo o categoría de la variable.
3. Transformar las frecuencias absolutas en frecuencias relativas o porcentajes.

Los criterios para definir los intervalos o categorías de una distribución (paso 1) están ligados al tipo de variable de la que provienen los datos.

• Variables cualitativas o categóricas

Estas variables se presentan por niveles o categorías. En nuestro problema de la contaminación, los niveles de la calidad del aire (bueno, regular, malo, muy malo y extremadamente malo) son un ejemplo de este tipo de variables. Consideremos el caso de la zona Noroeste (ver tabla 2).

Tabla 2: Calidad del aire zona Noroeste de la Ciudad de México. Diciembre 2017

Calidad del aire	Frecuencia (Días del mes)
Buena	8
Regular	15
Mala	8
Muy mala	0
Extremadamente mala	0
	31

Obsérvese que la distribución consta solo de dos columnas: la primera consiste en los niveles que definen la calidad del aire, la segunda contiene las frecuencias absolutas obtenidas del conteo directo de los datos. Una distribución más completa que contiene frecuencias en porcentajes y acumuladas se muestra en la tabla 3:

Tabla 3: Calidad del aire zona Noroeste de la Ciudad de México. Diciembre 2017

Calidad del aire	Frecuencia (Días del mes)	Frecuencia relativa (%)
Buena	8	26%
Regular	15	48%
Mala	8	26%
Muy mala	0	0%
Extremadamente mala	0	0%
	31	**100%**

El cálculo de las frecuencias relativas en porcentaje se determina dividiendo la frecuencia absoluta correspondiente entre el total de datos y se multiplica por 100. Por su parte, las frecuencias acumuladas se calculan sumando a la frecuencia anterior, la frecuencia del siguiente intervalo hasta completar el 100%.

$$\text{Frecuencia relativa } (\%) = \frac{\text{frecuencia absoluta}}{\text{total de datos}} \times 100$$

• Variables cuantitativas o numéricas

Las mediciones de una variable cuantitativa generalmente producen un amplio rango de datos. Retomemos los datos de los niveles de ozono y en particular los de la zona Centro, para mostrar cómo se procede cuando se presenta esta situación. Si colocamos los datos desde el más pequeño hasta el más grande como si fueran categorías nos quedaría la siguiente clasificación:

Valores de Imecas por Ozono	Frecuencia
29	
30	
.	
.	
114	
115	

Como vemos, sería una tabla demasiado grande y no serviría para resumir de manera adecuada los datos. Por lo general, las tablas deben tener pocas categorías o intervalos para obtener una buena visualización del comportamiento de los datos. Lo más adecuado en estos casos es construir intervalos de datos que se comportan como categorías.

Una regla práctica para determinar la amplitud de los intervalos de clase consiste en calcular el rango de los datos y dividirlo entre el número de intervalos deseados. Generalmente no se desean muchos intervalos, para que la tabla y la gráfica que se deriven de ella puedan ser más fáciles de interpretar. El rango es la diferencia entre el dato mayor y el dato menor.

$$\text{Amplitud} = \frac{\text{Rango}}{\text{Número de intervalos}}$$

Niveles de Ozono en la zona Centro

67	101	111	**115**	103	80	49	35	46	55	48	46
109	114	49	**29**	51	106	104	80	86	80	96	110
92	82	104	69	59	82	108					

- **Primer paso: Calcular el rango.**
  Rango = dato mayor – dato menor
  = 115 – 29 = 86
- **Segundo paso: Definir la cantidad de intervalos que se desean**
  Seleccionaremos 5 intervalos
- **Tercer paso: Calcular la amplitud de los intervalos**

$$\text{Amplitud} = \frac{86}{5} = 17.2$$

Como los datos son enteros, podemos redondear la amplitud a un valor entero para que los intervalos no queden fraccionarios. Si redondeamos al entero menor tendremos 5 intervalos de tamaño 17, lo que nos da un nuevo rango (rango calculado) de 85, el cual es inferior al rango de los datos que es 86, ello provocaría que algunos datos queden fuera de clasificación. Si redondeamos al entero mayor 18, el rango calculador es de 90, el cual es mayor al rango de 86. Por lo tanto, optamos por 5 intervalos de amplitud 18. La distribución que resulta se muestra en la tabla 4.

Tabla 4: Valores máximos diarios de ozono en el Centro de la Ciudad de México. Diciembre 2017

Valores de Ozono	**Frecuencia** (Días del mes)	**Frecuencia relativa**
29-46	4	13%
47-64	6	19%
65-82	7	23%
83-100	3	10%
101-118	11	35%
Total	**31**	**100%**

Condiciones que debe cumplir una distribución de frecuencias

1. Los intervalos o categorías de una distribución de frecuencias deben cumplir con dos requisitos:

a) Las categorías o intervalos deben ser *mutuamente excluyentes*.

b) Las categorías o intervalos deben ser *exhaustivas*.

El primer requisito implica que las categorías no deben traslaparse, es decir, al momento de clasificar un dato éste debe ser colocado en una y sólo una categoría. El segundo criterio implica que no debe haber huecos entre categorías subsecuentes, y que éstas deben ser lo suficientemente amplias para capturar cualquier dato de la variable en cuestión.

2. Una distribución de frecuencia debe ser consistente, esto es:

a) La suma de las frecuencias absolutas debe ser igual al total de datos.

b) La suma de las frecuencias relativas porcentuales debe ser igual al 100%

Elementos que debe contener una distribución de frecuencias

1. El título o encabezado debe describir claramente el contexto de los datos que se presentan.

2. Etiquetar dentro de las tablas todas las variables y las unidades en las cuales son medidas.

3. La fuente de donde se obtuvieron los datos debe colocarse al pie de la tabla.

4. Siempre que sea posible, utilizar frecuencias absolutas y relativas (porcentajes) para una mejor descripción de los datos.

5. La tabla debe ser consistente, esto es, la suma de renglones y columnas deben coincidir con el total datos o porcentajes**.**

Título

Calidad del aire zona Noroeste de la Ciudad de México.
Diciembre 2017

Variable

Calidad del aire	Frecuencia (Días del mes)
Buena	8
Regular	15
Mala	8
Muy mala	0
Extremadamente mala	0
	31

Categorías de la variable

Valores de las categorías

Consistencia

Fuente: Sistema de Monitoreo Atmosférico de la Ciudad de México.

Fuente

4.4 Uso de tecnología para construir distribuciones de frecuencias

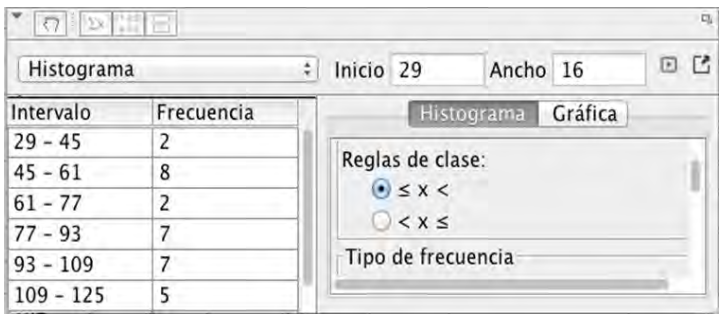
El uso de software estadístico y hojas de cálculo es de gran ayuda en la construcción de distribuciones de frecuencia, en algunos casos estas se construyen casi de manera automática con solo indicar el número de intervalos, pero en otros casos se requiere introducir varias funciones para realizar los cálculos. En el caso de *Geogebra*, la distribución de frecuencias se construye proporcionando el dato de inicio del primer intervalo y el ancho del mismo (ver tabla 5).

Tabla 5. Valores máximos diarios de ozono en el Centro de la Ciudad de México. Diciembre 2017

Intervalo	Frecuencia
29 - 47	4
47 - 65	6
65 - 83	7
83 - 101	3
101 - 119	11

Obsérvese la similitud de la distribución de frecuencias construida por *Geogebra* con la que construimos en la tabla 4. Se tiene la apariencia que los intervalos se traslapan en el límite superior e inferior, pero no es así, ya que en las reglas de clase se establece que el intervalo toma el valor izquierdo, pero no el derecho, esto es: $(a \leq x < b)$. (. Otras posibles distribuciones de frecuencia pueden ser generadas con solo cambiar el ancho del intervalo.

Tabla 6: Valores máximos diarios de ozono en el Centro de la Ciudad de México. Diciembre 2017



Intervalo	Frecuencia
29 - 45	2
45 - 61	8
61 - 77	2
77 - 93	7
93 - 109	7
109 - 125	5

4.5 Representaciones gráficas y visualización de datos

El desarrollo que ha alcanzado tecnología computacional en la actualidad, ha puesto a disposición de los profesionales de la estadística un amplio repertorio de

representaciones gráficas que facilitan la visualización de los datos. Además de la rapidez y precisión con la que las gráficas pueden ser construidas y modificadas, la mayoría de las herramientas de software permiten visualizar y explorar el comportamiento de los datos en forma interactiva y dinámica, lo cual ayuda mucho a la comprensión de los datos. El poder de cálculo y visualización de la tecnología han generado un cambio de enfoque centrado en la descripción *univariada* de los datos – conocido como estadística descriptiva - a un enfoque *multivariado* y exploratorio de datos.

Cada una de las gráficas que se pueden construir y tienen sentido sobre un conjunto de datos aporta diferentes elementos de información sobre su comportamiento. Con apoyo de la tecnología es posible maximizar la información que proporcionan los datos, explorando diversas opciones gráficas y extrayendo la información que cada una proporciona sobre el problema de interés. Las gráficas que se pueden construir para un conjunto de datos dependen del tipo de variable de la que provienen, por tanto, hay gráficas para variables cualitativas y gráficas para variables cuantitativas.

4.6 Gráficas para datos categóricos o cualitativos

- Diagrama circular o de sectores
- Diagrama de barras
  - Diagrama de barras múltiples
  - Diagramas de barras apiladas

• Diagramas circulares o de sectores

Este tipo de gráficas son apropiadas cuando la variable de interés tiene pocas categorías, lo cual permite visualizar más fácilmente el comportamiento de los datos. Se pueden utilizar frecuencias absolutas o relativas, pero es más común representar frecuencias relativas y porcentajes. El área de cada sector es proporcional a la frecuencia con la que aparecen los datos en las categorías.

Un diagrama de sectores se puede construir con un compás y un transportador. Se requiere calcular primeramente los porcentajes de datos que corresponden a cada categoría, y enseguida distribuir dichos porcentajes en los 360° de la circunferencia. Cada sector se puede iluminar con un color diferente, de manera que se distingan las diferentes categorías. Sin embargo, hoy en día lo más habitual es utilizar hojas de cálculo y software de cálculo estadístico.

Para mostrar un ejemplo de diagrama de sectores retomemos los datos de la contaminación por ozono en la zona Suroeste de la Ciudad de México que se presentaron durante el mes de diciembre de 2017.

Tabla 7. Calidad del aire en la zona Suroeste de la Ciudad de México. Diciembre de 2017

Calidad del aire	Frecuencia (Días)	Porcentaje
Buena	7	23%
Regular	11	35%
Mala	13	42%
Muy mala	0	0%
Extremadamente mala	0	0%
Total	**31**	**100%**

Fuente: Sistema de Monitoreo Atmosférico de la Ciudad de México

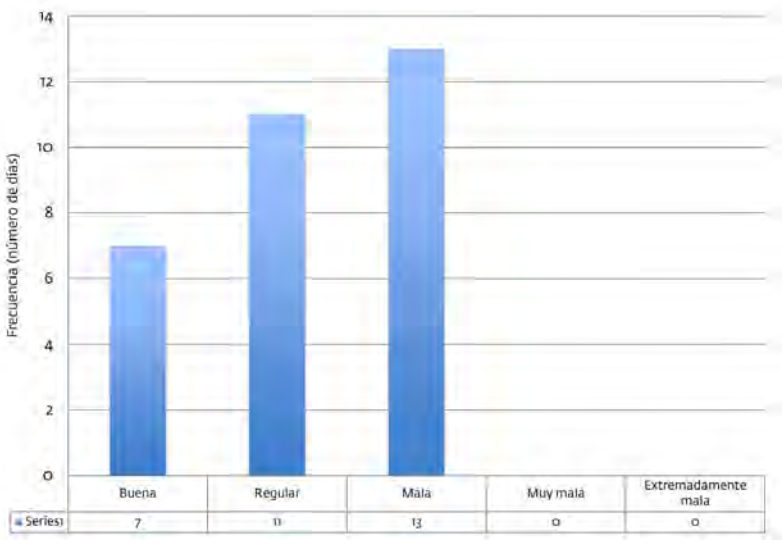
Obsérvese que el diagrama consta de tres sectores, debido a que la categoría mala y extremadamente mala de la calidad del aire, tienen frecuencia igual a cero.

• Diagramas de barras

Otra alternativa muy importante para graficar datos que provienen de variables cualitativas son las *gráficas de barras*. La gráfica consiste en un sistema de ejes, uno vertical y otro horizontal. En el eje horizontal se colocan las categorías de la variable y en el eje vertical las frecuencias (absolutas o relativas). La longitud de cada barra es proporcional a la frecuencia de cada categoría.

En algunas herramientas de software (por ejemplo, Excel) se hace la distinción entre diagrama de barras y diagramas de columnas. Se llama diagrama de columnas cuando las barras se disponen en forma vertical y diagramas de barras cuando se colocan en forma horizontal. Sin embargo, el nombre de diagrama de barras es universal independientemente de la forma como se dibujen las barras en el sistema de ejes.

Gráfica 4. Calidad del aire en la zona Suroeste de la Ciudad de México. Diciembre de 2017

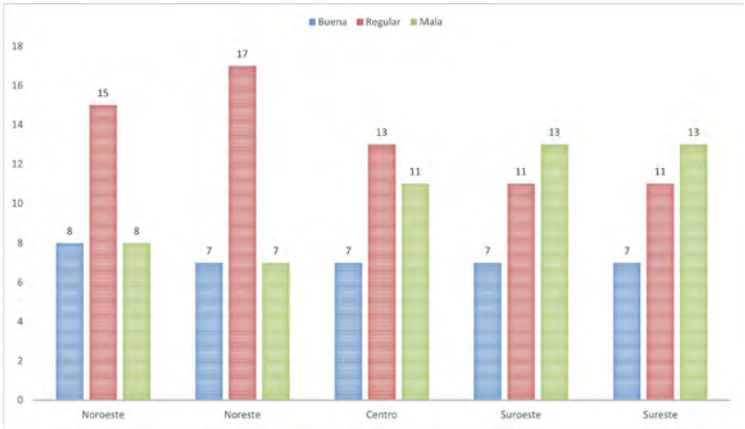


Fuente: Sistema de Monitoreo Atmosférico de la Ciudad de México.

• Diagramas de barras múltiples

En muchas ocasiones se requiere presentar información de más de una variable en una misma gráfica. Veamos de nuevo el caso de los datos sobre contaminación por ozono, en el cual se tienen dos variables: calidad del aire y zona de la ciudad. Un diagrama de barras múltiple puede ser buena opción para comparar los niveles de contaminación en cada una de las zonas (ver gráfica 5).

Gráfica 5. Calidad del aire por zona de la ciudad de México (Diciembre 2017)

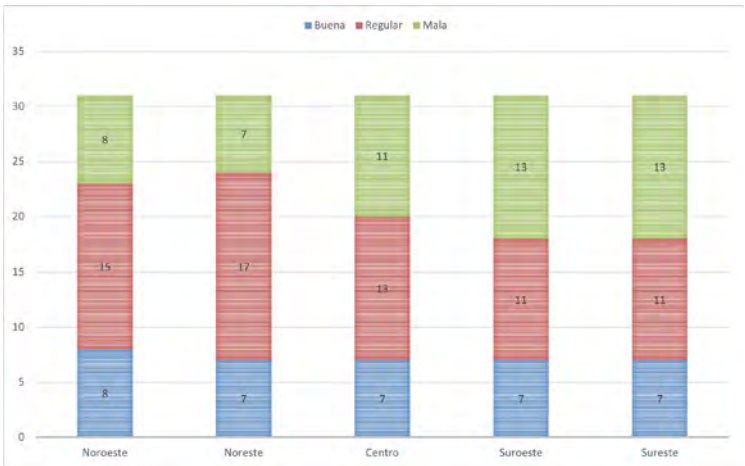


Fuente: Sistema de Monitoreo Atmosférico de la Ciudad de México.

• Diagramas de barras apiladas

Estos diagramas presentan la misma información que los diagramas de barras múltiples, pero en lugar de colocar las barras en forma contigua, se colocan una barra encima de la otra para formar una sola barra. La altura de cada sección de la barra expresa la magnitud de la frecuencia en cada categoría.

Gráfica 6. Calidad del aire por zona de la Ciudad de México (diciembre 2017)



Fuente: Sistema de Monitoreo Atmosférico de la Ciudad de México.

Elementos de un diagrama de sectores y un diagrama de barras

En una gráfica circular y en una gráfica de barras debe destacarse:

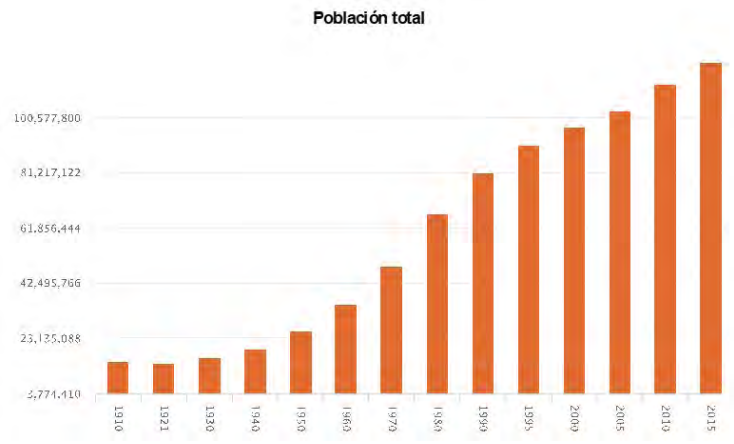
1. El título
2. La variable de interés y sus categorías
3. Las frecuencias (absolutas o relativas) de las categorías.
4. La fuente de los datos.

El título debe ser claro, explícito y resumido. Los valores de la variable deben ir perfectamente indicados en la gráfica. La fuente es una información importante que no debe faltar en cualquier tabla o gráfica, ya que permite conocer la procedencia y la forma como fueron obtenidos los datos. Finalmente, es importante revisar la consistencia de la gráfica: la suma de las frecuencias absolutas debe ser igual al total de los datos y la suma de los porcentajes debe ser igual al 100%.

Diagrama de barras vs. diagrama de sectores

Los diagramas de sectores y los diagramas de barras son esencialmente iguales, pues presentan, uno en forma de sectores circulares y otro en forma de barras, las frecuencias de los datos en cada categoría o intervalo. Un criterio de elección de un diagrama de barras sobre un diagrama circular podría ser la cantidad de categorías de la variable, cuando son muchas categorías, el diagrama circular se satura de información y dificulta la visualización de los datos; otro caso de preferencia de un diagrama de barras sobre diagrama circular puede ocurrir cuando se desea visualizar el cambio de una variable a través del tiempo (ver gráfica). En este caso un diagrama circular no es elegible como opción de representación gráfica.

Gráfica 7. Evolución de la población mexicana (1910-2015)



Fuente: INEGI

4.6 Representaciones gráficas para variables cuantitativas

1. Histogramas
2. Diagramas de caja
3. Diagramas de puntos
4. Diagramas de tallo y hoja
5. Gráficas temporales (series de tiempo)

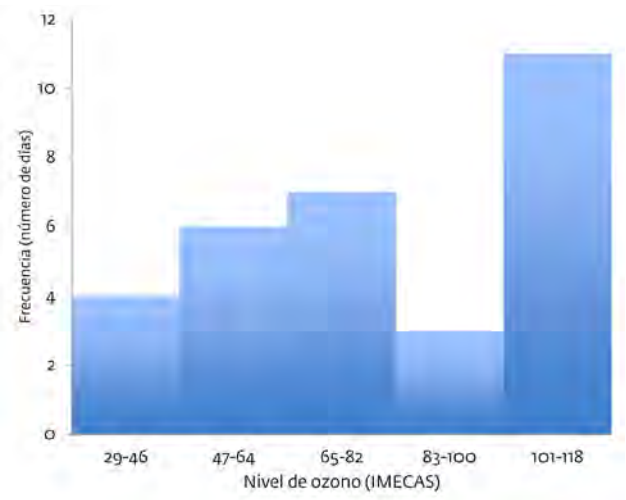
• Histograma

Esta representación gráfica es bastante usual para visualizar datos que provienen de variables cuantitativas, ya sean discretas o continuas. En el eje horizontal se colocan los valores de la variable y en el eje vertical se localizan las frecuencias de cada intervalo. El histograma se compone de una serie de rectángulos, cada uno de los cuales tiene como base un intervalo de datos y la frecuencia como altura.

Utilizaremos la hoja de cálculo Excel para construir un histograma con la distribución de los datos de contaminación por ozono de la zona Centro de la Ciudad de México (ver tabla 4).

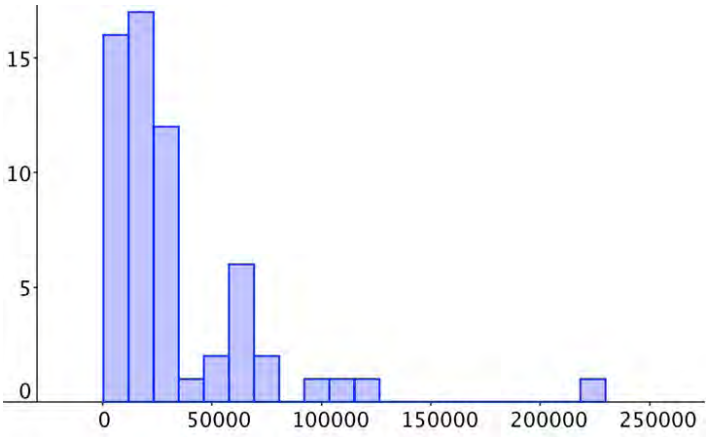
Obsérvese la similitud de un histograma con un diagrama de barras, ambos constan de un sistema de ejes y barras para representar las frecuencias. La diferencia radica que en el histograma las barras van juntas una de otra para garantizar la continuidad de la variable que está siendo representada, por su parte en el diagrama de barras, las barras se colocan separadas en tanto representan categorías de una variable no continuas. Esta diferencia no siempre es respetada en muchas gráficas que aparecen en los medios de comunicación.

Gráfica 8. Valores máximos diarios de ozono en el Centro de la Ciudad de México. Diciembre 2017



Para interpretar un histograma se debe identificar primeramente el *aspecto general de la distribución* y los datos que se desvían del conjunto, conocidos como *datos atípicos*. El aspecto general de un histograma involucra tres características de los datos: centro, dispersión y forma. En cuanto a la forma una distribución puede ser simétrica, sesgada a la derecha, sesgada a la izquierda e irregular. Consideremos la matrícula de 60 universidades mexicanas entre públicas y privadas.

Matrícula de universidades mexicanas (2016)



Fuente: http://www.execum.unam.mx

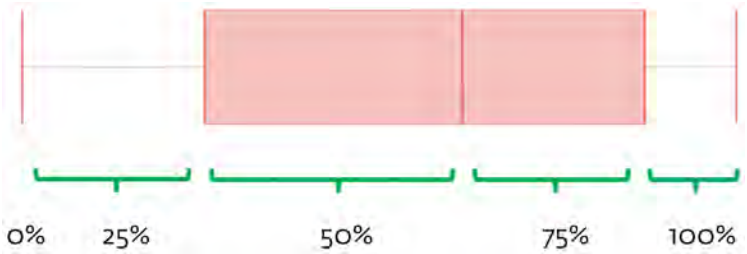
La forma del histograma es sesgada a la derecha con un dato atípico que representa la matrícula de la UNAM. El centro de los datos se ubica en 32,000 estudiantes aproximadamente. La dispersión es muy amplia ya que aparecen universidades con una matrícula muy pequeña y universidades que superan los 100,000 estudiantes, incluso una de ellas supera los 200,000 estudiantes.

• Diagrama de caja

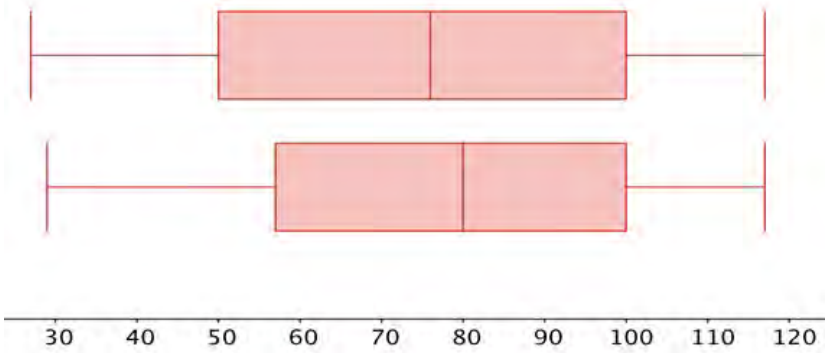
Estas gráficas son muy útiles para analizar el comportamiento de los datos por secciones; consiste en *ordenar* los datos de una variable y seccionar el rango en cuatro partes que contienen el 25% de los datos cada una, a los cuales se les denomina *cuartiles*. En un diagrama se caja se representan 5 medidas resumen de los datos:

1. Dato menor
2. Cuartil 1
3. Cuartil 2 (mediana)
4. Cuartil 3
5. Cuartil 4 (dato mayor)

Un cuartil es un valor que es mayor al 25%, 50%, 75% o 100% de los datos (según el cuartil de que se trate). Por ejemplo, el cuartil 1 es mayor al 25% de los datos (y a su vez es menor al 75% de los datos restantes), el cuartil 2 es mayor al 50% de los datos (y a su vez es menor al otro 50% de los datos). La línea que está al centro de la caja -razón por la que toma el nombre esta gráfica-, se llama *mediana*, ya que se ubica exactamente en el centro de la distribución de los datos.



Los diagramas de caja son muy útiles para establecer comparaciones de dos o más distribuciones de datos, situación que no se facilita en el caso de los histogramas. Consideremos las distribuciones de los datos de contaminación por ozono en las zonas Noroeste y Noreste de la Ciudad de México.



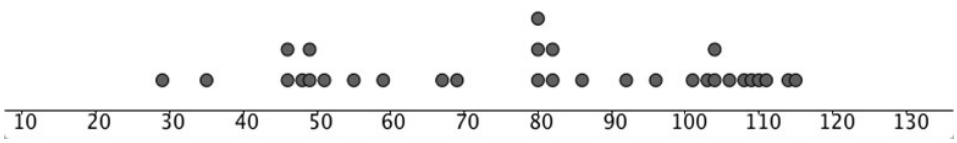
Gráfica 9. Contaminación por ozono en las zonas Noroeste y Noreste de la Ciudad de México (diciembre 2017)

La extensión de los diagramas muestra que ambas zonas tuvieron casi el mismo rango de contaminación de 29 y 30 a 118 aproximadamente. La mediana de la zona Noroeste es ligeramente menor a la de la zona Noreste, por lo que en promedio hubo menos contaminación.

• Diagrama de puntos

En una gráfica de puntos cada dato se representa mediante un punto. Para construir un diagrama de puntos se dibuja una línea horizontal y se etiqueta con el nombre de la variable, se define una escala y posteriormente cada dato se representa mediante un punto sobre la línea; cuando aparecen datos repetidos se coloca un punto encima de otro. Una gráfica de puntos para los datos de contaminación de la zona Centro se muestra en la gráfica 10.

En la gráfica de puntos se puede identificar el valor de cada dato, lo que la hace apropiada para pequeños conjuntos de datos; además de ello, se puede ver el comportamiento global de los datos en una sola mirada.



Gráfica 10. Niveles de contaminación por ozono (IMECAS)

• Diagrama de tallo y hoja

Otro importante tipo de gráficas para representar pequeños conjuntos de datos son las gráficas de tallo y hoja. Cada dato es presentado por un tallo y una hoja, usualmente el tallo consiste en todos los dígitos excepto el último, el cual es la hoja. Los datos

que representan el tallo se colocan en la parte izquierda de una línea vertical y los datos que representan la hoja se colocan del lado derecho. A continuación, mostramos un diagrama de tallo y hoja con los datos de contaminación en la zona Centro.

Gráfica 11. Valores máximos diarios de ozono en el Centro de la Ciudad de México. Diciembre 2017

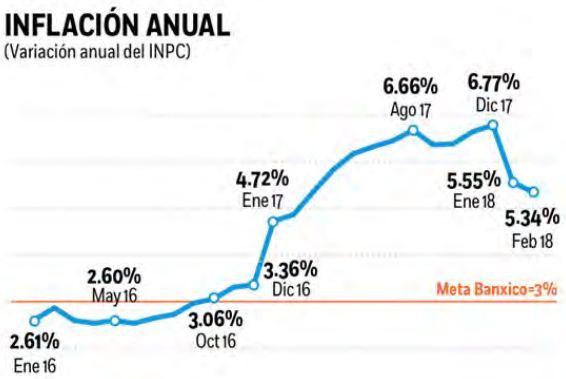
2	9						
3	5						
4	6	6	8	9	9		
5	1	5	9				
6	7	9					
7							
8	0	0	0	2	2	6	
9	2	6					
10	1	3	4	4	6	8	9
11	0	1	4	5			

El primer valor representa el dato 29, el segundo representa el dato 35, luego aparecen 46, 46, 48 y así sucesivamente. En los casos donde no aparece una hoja significa que no existe un dato con el respectivo tallo. Un diagrama de tallo y hoja tiene un aspecto similar al de un histograma colocado en posición vertical. Los datos no pierden identidad y se puede ver la variabilidad, los datos más frecuentes, los menos frecuentes, los agrupamientos y los vacíos.

• Graficas temporales (series de tiempo)

Estas gráficas permiten visualizar el comportamiento de variables respecto al tiempo. Los datos se grafican en un sistema de ejes: en el eje horizontal se colocan los valores del tiempo (horas, días, semanas, meses, años) y en el eje vertical el valor de los datos que corresponden a cada unidad de tiempo, entonces los puntos resultantes se unen mediante una línea. El patrón de comportamiento de los datos en este tipo de gráficas se denomina *tendencia*.

Gráfica 12. Comportamiento de la inflación 2016-2018



Fuente: Periódico Excélsior 9/03/2018

Al interpretar una gráfica temporal es necesario prestar atención a los siguientes aspectos:

1. *Comportamiento general o tendencia*. La tendencia puede ser creciente o decreciente en algún tramo de la gráfica o en toda su extensión.
2. *Comportamiento local (desviaciones bruscas de la tendencia).* Se refiere a cambios bruscos en la tendencia de los datos en algún punto de la gráfica, ya sea en forma de incremento o decremento.
3. *Variaciones estacionales.* Se refieren a cambios que se presentan en la tendencia en forma de ciclos que se repiten a lo largo del tiempo.

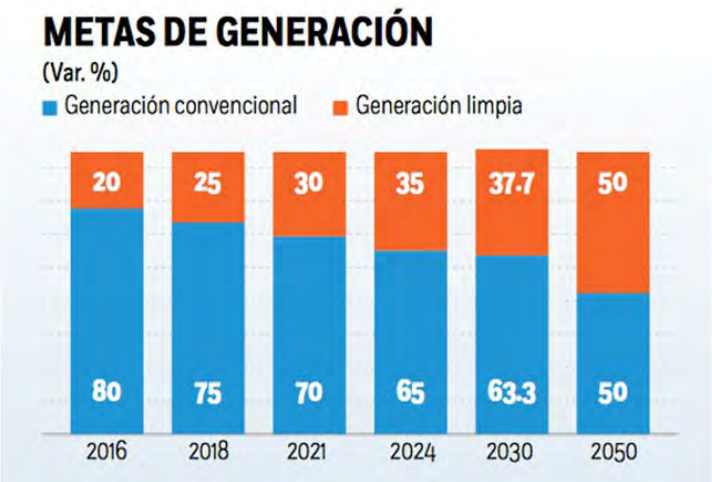
Por ejemplo, la gráfica de los usuarios de internet en Italia, Brasil y México muestra una tendencia creciente desde 1995 a 2010, no cambios bruscos en el comportamiento de la

tendencia ni variaciones estacionales. Mientras que la gráfica de la inflación muestra una tendencia creciente de enero de 2016 a diciembre de 2017, con un cambio brusco al alza de diciembre de 2016 a enero de 2017, y en enero y febrero de 2018 disminuye de manera importante.

• Gráficas combinadas

Hemos mostrado algunos casos en los que es posible colocar más de una gráfica en un mismo sistema de ejes para visualizar comportamientos de los datos y realizar comparaciones entre ellos. En este contexto son muy frecuentes los diagramas de barras múltiples y los diagramas de barra apiladas para establecer relaciones entre dos o más variables. Hay algunos casos donde además a estas gráficas se le agrega una variable temporal, lo que permite hacer comparaciones sobre el comportamiento de las variables respecto del tiempo (ver gráfica 13).

Gráfica 13. Evolución y comparación de energías convencionales y energías limpias en México



Periódico Excélsior: 16/03/2018

Obsérvese que en la gráfica 15 se presenta información de tres variables: energía convencional, energía limpia y el tiempo. El diseño de la gráfica permite contrastar el uso la energía convencional y energía limpia en un año en particular, digamos en 2016, donde la energía limpia solo representa el 20% del total, pero además permite visualizar el cambio que las variables tienen a través del tiempo, digamos que de 2016 a 2050 se espera que la energía limpia pase del 20% al 50%.

4.7 Elementos para la interpretación de tablas y gráficas

A continuación, describiremos un pequeño marco interpretativo que consta de cinco pasos para interpretar representaciones gráficas y tabulares. El marco toma en cuenta desde la observación de los elementos que integran una representación gráfica hasta el establecimiento de relaciones de comparación de categorías, así como la explicación contextual de las posibles razones para las relaciones identificadas en el comportamiento de los datos.

1. *Visualizar los elementos de la tabla o la gráfica.*

Visualiza el título, ejes, encabezados, etiquetas y fuentes de los datos para tener una idea del contexto y la calidad de los datos. Toma en cuenta si los datos fueron tomados de encuestas, experimentos o censos.

2. *Buscar el significado de los números*

Intenta comprender qué representan los números (frecuencias, porcentajes, datos simples o procesados) que aparecen en la tabla o la gráfica. Identifica el valor más grande y el más pequeño en una o más categorías de la variable o el tiempo, para obtener una impresión de los datos.

3. *Buscar diferencias en los valores de los datos*

Observa las diferencias en los valores de los datos, en su conjunto, en un renglón, en una columna, o en una parte de la gráfica. Esto puede requerir identificar cambios de la variable a través del tiempo o realizar comparaciones de una categoría con otra. Por ejemplo, el cambio en la matrícula de una escuela a lo largo del tiempo tanto en hombres como mujeres.

4. *Identifica en dónde están las diferencias en los valores de los datos*

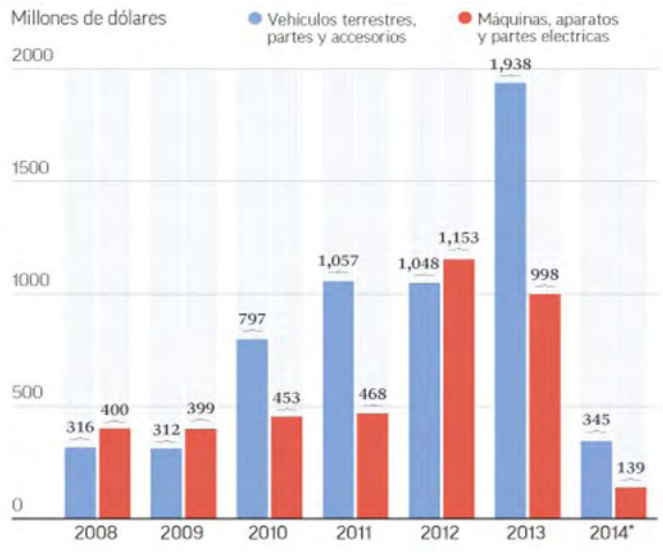
Una vez que observas alguna diferencia en los valores de los datos, identifica las relaciones que existen entre las variables. Por ejemplo, a lo largo de los años la matrícula de una escuela ha ido en aumento, pero la matrícula de mujeres ha crecido el doble que la matrícula de hombres.

5. *Identifica las razones por las cuales hay diferencias en los valores de los datos.*

Intenta buscar razones para explicar las relaciones que identificaste en los datos, considerando el contexto del cual provienen los datos y tomando en cuenta factores sociales, económicos, ambientales, entre otros.

Actividades de aprendizaje

1. Analiza en interpreta la siguiente gráfica de barras dobles y responde las siguientes preguntas:



Exportaciones de sectores clave de México a China, Japón, Corea e India (millones de dólares)

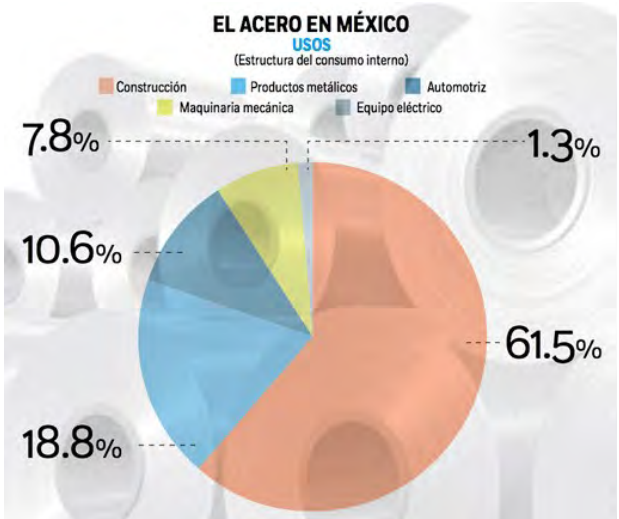
Fuente: Periódico El Financiero con datos de Banco de México.

a) ¿Qué representan los números de los ejes de la gráfica?
b) ¿En qué año hubo más exportaciones?
c) ¿En qué año hubo menos exportaciones?
d) ¿Cuál ha sido la tendencia de las exportaciones de 2008 a 2013?
e) Compara las exportaciones de vehículos terrestres, partes y accesorios con las de máquinas, aparatos y partes eléctricas en 2013. Expresa en términos porcentuales la relación.
f) Realiza una interpretación de la gráfica como si se la explicaras a alguien que sabe poco o nada de estadística.

2. Al correo de este autor llegó la siguiente gráfica anunciado el programa de apoyo del banco BBVA a todos clientes en la crisis del Covid-19 y la frecuencia de llamadas según la hora del día.

a) ¿Es esta gráfica un histograma o un diagrama de barras?
b) ¿A qué hora harías una llamada al banco, si se quiere tener mayor probabilidad que contesten rápido?
c) ¿Cuál es el intervalo de mayor número de llamadas?
d) ¿Cuántas llamadas aproximadamente recibe el banco en el intervalo de 10:00 a 14:00 horas?

3. Interpreta la siguiente gráfica de sectores sobre los usos del acero en México. Toma en consideración los elementos para la interpretación de gráficas que se describieron en el texto. Identifica la variable de interés y verifica si cumple los criterios de consistencia.



Excélsior 7/03/2018

En muchos currículos se propone que el análisis de datos haga más énfasis en la exploración que en la descripción de los datos; en dicha exploración, las representaciones gráficas ocupan un lugar fundamental. Este cambio de enfoque ha sido derivado del trabajo de John Tukey (1977) y se conoce como *análisis exploratorio de datos*. Una idea fundamental en la que se apoya este enfoque es que al utilizar diferentes representaciones de un conjunto de datos se puede facilitar la comprensión.

En un ambiente computacional dinámico e interactivo como el que proporcionan muchas tecnologías digitales actualmente, es posible cambiar de una representación a otra (por ejemplo, de una tabla a una gráfica o de una gráfica a otra gráfica) de una forma sencilla, con lo que se facilita la visualización de estructuras en los datos; también es posible identificar patrones en los datos mediante el cambio de un dato, un parámetro o la escala de una gráfica.

De acuerdo con lo anterior, el análisis exploratorio de datos puede ser una herramienta de utilidad en la generación de hipótesis, conjeturas y preguntas de investigación acerca de los fenómenos de donde los datos fueron obtenidos, y por ello se propone como medio para desarrollar una comprensión global de los datos y promover el desarrollo del razonamiento estadístico en los estudiantes.

Nota histórica: El origen de las gráficas

En el siglo XVII, en su obra La Geometría, René Descartes introduce el método de coordenadas cartesianas, el cual constituye la base para la representación gráfica de información cuantitativa. Sin embargo, es hasta finales del siglo XVIII cuando el ingeniero y economista escocés William Playfair (1759-1823) desarrolla y promueve el uso de gráficas en los negocios y la economía. Playfair fue pionero en el uso de gráficas para desplegar información cuantitativa y para comunicar relaciones entre números en una forma que textualmente no era posible, lo cual lo convierte en figura clave en la historia de los métodos gráficos.

De 1786 a 1801 escribe su libro más conocido el cual lleva el nombre *The Commercial and Political Atlas*. De 1801 a 1805 desarrolla nuevas gráficas, entre las cuales se encuentran algunas muy utilizadas aún en la actualidad, como la gráfica de barras, la gráfica circular o de sectores y la gráfica de línea para representar series temporales. La idea de representar visualmente el espacio se venía usando desde hacía siglos, pero hasta entonces a nadie se le había ocurrido representar series numéricas visualmente. Otra de sus grandes obras fue *The Statistical Breviary*, con datos económicos y demográficos europeos.

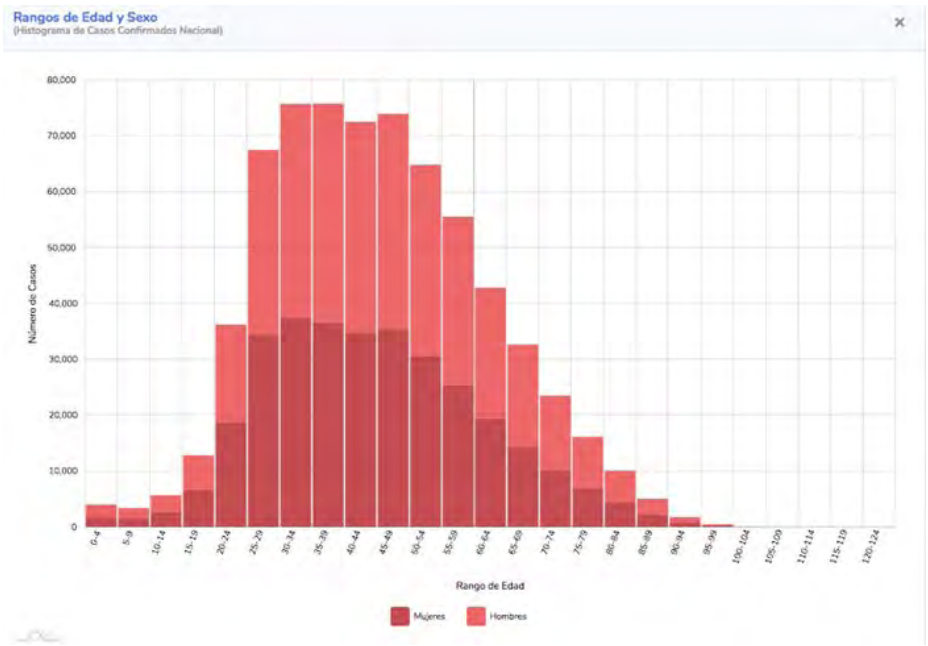
Según Playfair, una buena gráfica proporciona una explicación más adecuada de los hechos que una mera lista de datos o tablas. Sirve para simplificar lo complejo, permite al cerebro una mayor retención y es un instrumento visual de información cuantitativa. Por último, las gráficas permiten ver relaciones aparentemente inexistentes entre variables, que suelen quedar ocultas entre la multitud de datos y cifras.

Su filosofía sobre las gráficas se puede describir mediante los siguientes aspectos:

1. El método gráfico es una forma de simplificar lo tedioso y complejo.
2. Las personas ocupadas necesitan ordenamiento de ayuda visual.
3. El método gráfico es más accesible que las tablas.
4. El método gráfico apela a la vista.
5. El método gráfico apela a la mente.

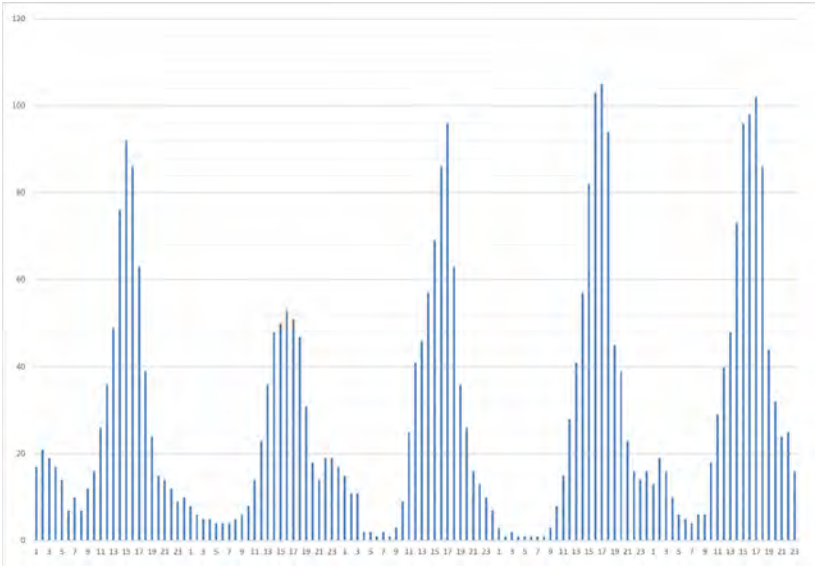
Evaluación del capítulo

1. La siguiente gráfica muestra el número de casos confirmados de Covid-19 en México registrados hasta el 16 de septiembre de 2020. Interpreta la gráfica y selecciona los argumentos que sean correctos.



a) La mayoría de los enfermos de Covid-19 tienen una edad entre 25 y 69 años.
b) No hay gran diferencia entre la cantidad de hombres y mujeres que contraen el Covid-19.
c) Aproximadamente 10,000 casos confirmados corresponden a la edad de 80 a 84 años.
d) La distribución de los casos confirmados de Covid-19 por rango de edad, es aproximadamente simétrica con un promedio de edad cercano a los 50 años.
e) Existen personas con más de 100 años de edad que han sido confirmadas con Covid-19.

2. La siguiente gráfica muestra el comportamiento de los niveles horarios de ozono en la zona Centro de la Ciudad de México del 1 al 5 de enero de 2018.



Fuente: http://www.aire.cdmx.gob.mx/default.php?opc='aqBjnmU='

Escribe brevemente el patrón de comportamiento de los niveles de ozono que identificas en la gráfica.

3. Considera que tienes menos de 100 datos numéricos de una investigación y quieres graficarlos, pero sin perder la identidad de los datos originales, ¿cuáles de las siguientes gráficas son la mejor opción?

a) Diagrama de puntos
b) Diagrama de cajas
c) Histograma
d) Diagrama de tallo y hoja

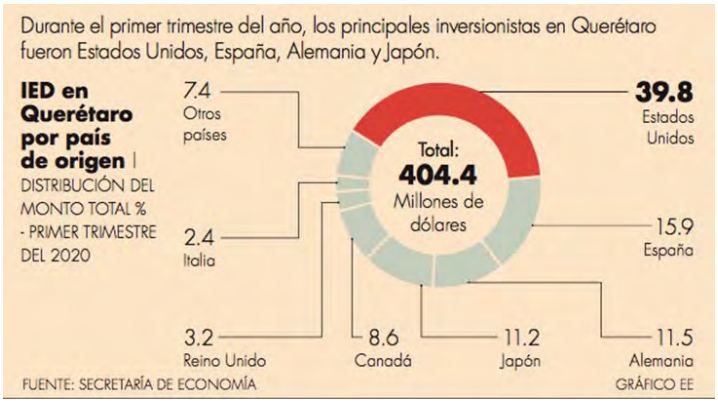
4. Selecciona los elementos que debe contener una tabla o distribución de frecuencias

a) Título o encabezado.
b) Etiquetas o nombres de las variables y las unidades en las cuales son medidas.
c) La fuente de donde se obtuvieron los datos.
d) Frecuencias absolutas y/o relativas.
e) La suma de renglones y columnas deben coincidir con el total datos o porcentajes.

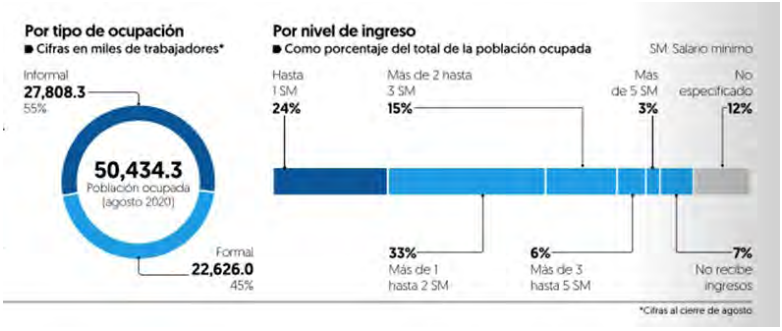
5. Interpreta la siguiente gráfica que fue publicada en el periódico El Economista y selecciona las afirmaciones que son ciertas.

a) Estados Unidos fue el país que más invirtió en Querétaro en el período considerado
c) La inversión de España en Querétaro fue de 160.95 millones de dólares
d) Los tres principales inversores fueron Estados Unidos, España y Alemania, con un 67.2% de la inversión total.

e) La inversión de España fue de 15.9 millones de dólares.



6. En el periódico El Financiero se publicaron las siguientes gráficas sobre la población ocupada en agosto de 2020 en México (las cifras están dadas en miles de trabajadores). Convierte los porcentajes por nivel de ingreso a cifras absolutas, y selecciona las respuestas que son correctas.



a) 12,104.2 trabajadores ganan menos de 1 salario mínimo.
b) 16,643.3 trabajadores ganan entre 1 y 2 salarios mínimos.
c) 7,500.1 trabajadores ganan entre 2 y 3.5 salarios mínimos
d) 3,026.05 trabajadores ganan entre 3 y 5 salarios mínimos

Bibliografía recomendada

- Uso de gráficas en conferencia de prensa sobre Covid-19 https://www.pscp.tv/w/1rmxPAOrprmKN?t=6m53s
- Día internacional de la mujer: 8 gráficas sobre la desigualdad de género. https://www.eleconomista.com.mx/economia/Dia-Internacional-de-la-Mujer-8-graficos-sobre-la-desigualdad-de-genero-20200308-0002.html
- Sitio especializado en visualización de datos. https://flowingdata.com
- Las tablas y gráficos estadísticos como objetos culturales http://www.sinewton.org/numeros/numeros/76/Articulos_02.pdf
- Estadísticas del mundo https://www.worldometers.info

Capitulo 5

Resúmenes numéricos de una distribución de datos: centro y variabilidad

El análisis exploratorio de datos busca revelar la estructura y descripciones en los datos. Observamos los números o las gráficas e intentamos encontrar los patrones.

Persi Diaconis

5.1 Introducción

Además de construir distribuciones de frecuencias y representaciones gráficas de los datos, es importante calcular resúmenes numéricos para complementar y precisar el análisis. En este apartado abordaremos dos tipos de resúmenes numéricos que describen características importantes de una distribución de datos: *medidas de centro* (*promedios*) y *medidas de variabilidad (dispersión)*.

Las medidas de centro resumen los datos en un solo valor, el cual representa el promedio de toda la distribución. Su uso es bastante frecuente en el medio escolar, medios de comunicación, información gubernamental y en investigación científica. A manera de ejemplo se presenta la siguiente información recopilada por el INEGI:

Las mujeres viven en promedio más años que los hombres; en 1930, la esperanza de vida para las personas de sexo femenino era de 35 años y para el masculino de 33. Al 2010 este indicador fue de 77 años para mujeres y 71 para los hombres, en 2016, se ubicó en casi 78 años para las mujeres y en casi 73 años para los hombres.

Con datos de 2011, la edad promedio al matrimonio de los hombres era de 29.2 años y en las mujeres es de 26.3 años.

Con datos de 2010, las mujeres mexicanas en edad reproductiva tenían un promedio de 2.3 hijos nacidos vivos.

Al año 2017, el grado promedio de escolaridad de la población mexicana de 15 años y más, fue de 9.2 años.

Las medidas de variabilidad resumen los datos en un solo valor, el cual indica el grado de variabilidad o dispersión que tienen los datos. Su uso es menos frecuente – pero no menos importante - en información gubernamental y medios de comunicación, que por lo general se limitan a informar sobre el promedio de una distribución de datos. A manera de ejemplo, obsérvese que la gráfica 1 muestra el ingreso promedio anual de los mexicanos, pero no reporta la variabilidad, lo cual sería muy importante, pues sabemos que en cuestión de salarios hay mucha variabilidad.

Gráfica 1.

Fuente: Periódico Excélsior 24/02/2018

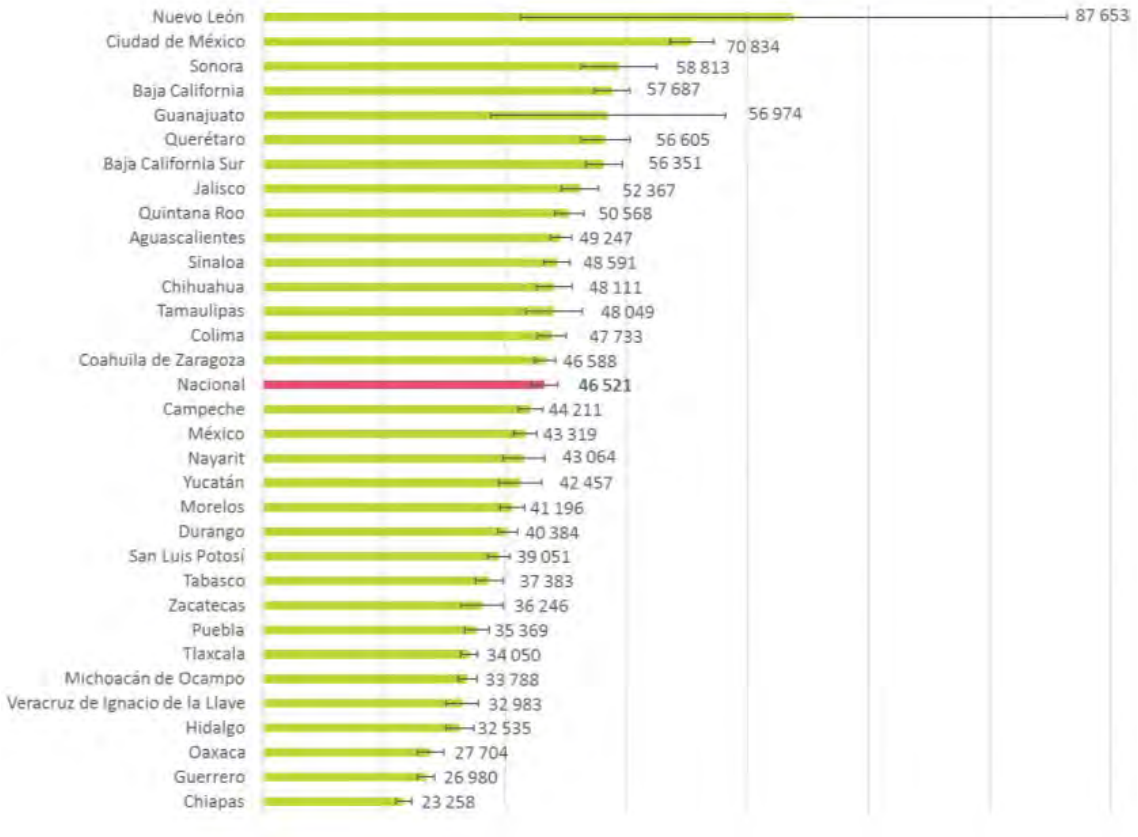
Un caso concreto de uso de medidas de variabilidad ocurre cuando se reporta el clima de una determina ciudad; se establece un *rango* de temperaturas delimitado por la temperatura máxima y la temperatura mínima, lo que permite tener una idea de la variabilidad de la temperatura a lo largo del día. Una descripción más precisa de la temperatura debe contemplar, además de la variabilidad establecida por el rango, la temperatura promedio.

Entonces, si consideramos únicamente el valor central o promedio de una distribución de datos, o si comparamos varios conjuntos de datos usando valores centrales solamente, podemos llegar a conclusiones erróneas, pues debemos considerar también la variabilidad de los datos para una descripción más completa. Existen varias maneras de calcular el promedio y la variabilidad de una distribución de datos, que analizaremos en los siguientes párrafos.

5.2 Medidas de centro de una distribución de datos

Las medidas de centro de una distribución de datos también son conocidas como promedios. El promedio más conocido y más utilizado cuando se analizan variables cuantitativas es la *media aritmética*. Sin embargo, existen otras importantes medidas de centro para analizar datos cuantitativos, como el caso de la *mediana*. Para mostrar la importancia de las medidas de centro de una distribución de datos, presentamos la gráfica 2, con los ingresos promedios trimestrales por hogar en cada entidad federativa de México en el año 2016.

Gráfica 2. Ingreso Promedio Trimestral por Hogar (En pesos, 2016)



Fuente: Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2016. INEGI

Cada dato de la gráfica representa el ingreso promedio trimestral para cada entidad en el año 2016. Es decir, se han resumido a un solo valor los ingresos de todos los trabajadores registrados en la entidad respectiva; incluso, se resumen los ingresos de todos los trabajadores del país en el ingreso promedio trimestral nacional que corresponde a $46,521. Como puede verse, el cálculo de promedios tiene la gran ventaja de resumir y simplificar un conjunto de datos en un solo número, pero la simplificación también representa pérdida de información de los datos originales.

Aquí resulta pertinente una cita del libro de Stephen Stigler *Los siete pilares de la sabiduría estadística*, "dada una cantidad de observaciones en verdad se puede obtener información ¡si se desecha información! Al calcular una simple media aritmética, descartamos la individualidad de las medidas subsumiéndolas en otra que es un resumen" (p. 13)

• Media aritmética

La media aritmética se define como la suma de todos los datos dividida entre el total de ellos. Cuando un conjunto de datos ha sido recopilado de una población, la media aritmética se representa de la siguiente manera:

$$\mu = \frac{x_1 + x_2 + x_3, \ldots, + x_N}{N}$$

Cuando los datos son recopilados de una muestra, la media aritmética tiene la siguiente expresión:

$$\bar{x} = \frac{x_1 + x_2 + x_3, \ldots, + x_n}{n}$$

Para ejemplificar el cálculo de la media, consideremos las temperaturas máximas que se pronostican para las ciudades de Culiacán y Mazatlán, durante la semana comprendida del 26 de marzo al 2 de abril de 2018.

Pronóstico de temperaturas para las ciudades de Culiacán y Mazatlán

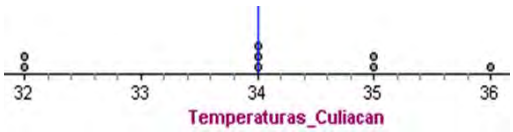
La temperatura media para Culiacán es:

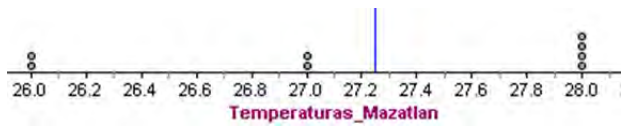
$$\bar{x} = \frac{35 + 32 + 32 + 34 + 36 + 35 + 34 + 34}{8} = 34.0 \text{ grados}$$

La temperatura media para Mazatlán es:

$$\bar{x} = \frac{27 + 26 + 26 + 27 + 28 + 28 + 28 + 28}{8} = 27.25 \text{ grados}$$

Podemos decir, entonces, que en promedio la temperatura de Culiacán en la semana fue mayor que la de Mazatlán aproximadamente en 7 grados. Un diagrama de puntos nos ayuda a visualizar los datos y la posición que ocupa la media aritmética (línea vertical). Obsérvese que dos datos quedaron por debajo de la media, tres quedaron por encima y tres datos coinciden con la media.





Observaciones importantes sobre la media aritmética

- La media aritmética no siempre coincide con alguno de los datos que se utilizaron en su cálculo.
- Para calcular la media aritmética no se requiere ordenar los datos.
- La cantidad de datos que se ubican por debajo o por encima de la media, no siempre es la misma, depende de la forma de la distribución de los datos.
- La media aritmética se puede considerar como el punto de equilibrio de los datos.
- La media es muy sensible a datos extremos o atípicos. Esto significa que datos muy alejados del conjunto, mueven a la media hacia su lado.
- Si todos los datos son iguales, la media es igual a cualquiera de ellos.

• Mediana

La mediana se define como el valor que está justamente en el centro de una *distribución ordenada* de datos. Es decir, divide a la distribución en dos partes iguales, razón por la que la mitad de los datos es menor a la mediana y la otra mitad es mayor a la mediana. Consideremos de nuevo las temperaturas de Culiacán y Mazatlán, a las que ya calculamos su media aritmética. Ordenando los datos tenemos lo siguiente:

Culiacán	32	32	34	34	34	35	35	36
Mazatlán	26	26	27	27	28	28	28	28

Mediana

La mediana se encuentra en el medio de la distribución, para el caso de Culiacán se encuentra entre los datos 34 y 34, que sumados y divididos por 2 resulta 34. Para el caso de Mazatlán la mediana se encuentra entre 27 y 28, que sumados y divididos por 2 resulta 27.5

Posición de la mediana

Con pocos datos resulta fácil identificar la posición en la que se encuentra la mediana, pero para distribuciones con muchos datos la *posición de la mediana* - no se confunda con el valor de la mediana -, se determina con la siguiente fórmula: $P = \frac{n+1}{2}$

Cuando el total de datos es par, la mediana siempre resulta en el valor único que se ubica en el centro de la distribución, pero cuando es impar, la mediana se obtiene de sumar y dividir entre 2, los dos valores centrales de la distribución.

La mediana es un resumen que se calcula en datos cuantitativos, pero también se puede calcular con datos cualitativos de tipo ordinal. Veamos un ejemplo: En la clase de estadística, el autor de este libro envió un cuestionario en línea a sus estudiantes, una de las preguntas del cuestionario era: *en general, ¿qué tan fácil te resulta trabajar en equipo?* La variable es *facilidad para trabajar en equipo* y las opciones de respuesta son categorías de una variable ordinal.

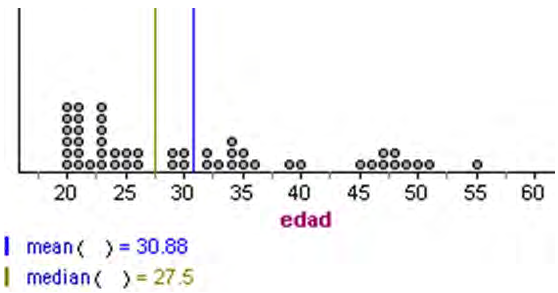
Facilidad para trabajar en equipo	Frecuencia	Orden de los datos
Extremadamente fácil	7	7
Muy fácil	15	22
Moderadamente fácil	12	34
Poco fácil	7	41
Nada fácil	4	45
	45	

Si ordenamos los datos y calculamos la posición de la mediana, $P = \frac{n+1}{2} = \frac{45+1}{2}$, esta se encuentra en la posición 23. La posición 22 corresponde a la categoría *muy fácil*, entonces la posición 23 corresponde a *moderadamente fácil*.

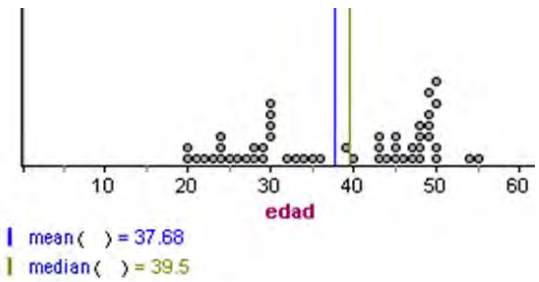
5.3 Relación entre media aritmética y mediana

La media y la mediana son medidas de centro para variables cuantitativas, y en el caso de la mediana también para datos cualitativos ordinales. La media aritmética toma en cuenta a todos los datos, un cambio en un dato o agregar datos nuevos puede cambiar significativamente el valor de la media aritmética. Por su parte, la mediana requiere el ordenamiento de los datos como primer paso para localizar el dato que está más al centro; un cambio en uno de los datos puede modificar en forma poco significativa el valor de la mediana, incluso en algunos casos puede no alterarla, por lo que es más robusta a datos extremos.

La media y la mediana pueden ser iguales si la distribución de los datos es simétrica, o muy cercanos una de otra si la distribución es aproximadamente simétrica. Si la distribución es sesgada hacia la derecha, la media se mueve lejos de la mediana en la dirección de la cola más larga y viceversa (ver gráficas de edades de dos grupos de personas).



Distribución sesgada a la derecha



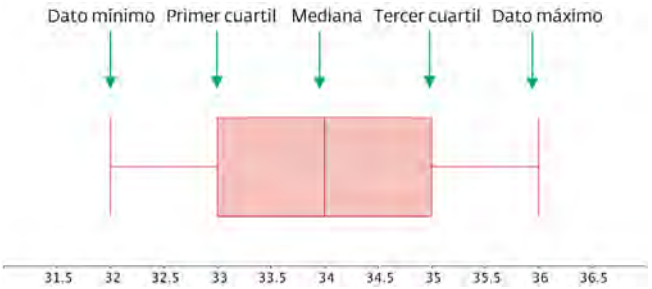
Distribución sesgada a la izquierda

La media aritmética es la medida de centro más utilizada, particularmente cuando se hacen inferencias estadísticas, sin embargo, la mediana puede ser mejor en distribuciones de datos que son asimétricas, ya que es más robusta a datos extremos.

Un ejemplo de ello ocurre cuando se analizan ingresos salariales, en estos casos la mediana es de uso común. El INEGI utiliza la mediana para reportar la edad promedio de los mexicanos, que en 2015 fue de 27 años.

5.4 La mediana, el diagrama de caja y los cuartiles de una distribución

En el capítulo anterior estudiamos los diagramas de caja, señalamos que en un diagrama de caja aparecen *cinco medidas resumen* muy importantes para analizar una distribución de datos: dato mínimo, dato máximo, primer cuartil, segundo cuartil (mediana) y tercer cuartil. Veamos el diagrama de cajas para las temperaturas de Culiacán que ya hemos discutido previamente.



Pronóstico de temperaturas para Culiacán (26 de marzo a 2 de abril 2018)

Cada cuartil representa el 25% de los datos de la distribución, así la mediana que se ubica exactamente en el centro de la distribución equivale al segundo cuartil, ya que el 50% de los datos son menores o mayores a ella. Los cuartiles permiten hacer un análisis más completo de una distribución de datos, pues además del centro, es posible analizar en cuatro segmentos el comportamiento de la distribución.

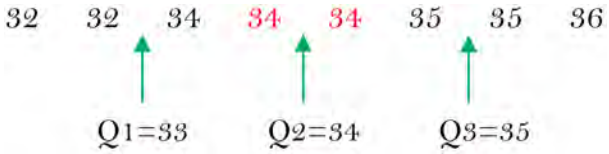
Es común que los programas de cómputo estadístico reporten las cinco medidas de un diagrama de caja como parte de un análisis básico de datos. Veamos el análisis del software *Geogebra* para las temperaturas de Culiacán.

Q1 = 33 representa el primer cuartil.
Q2 (mediana) =34 representa al segundo cuartil.
Q3 = 35 es el tercer cuartil.
Mínimo = 32
Máximo = 36

El cuartil Q1 se ubica en medio de la primera mitad de los datos, lo que resulta en 33. El cuartil Q3 está en medio de la segunda mitad, lo que resulta en 35. Los resultados concuerdan por completo con los que proporciona *Geogebra*.

n	8
Media	34
σ	1.3229
s	1.4142
Σx	272
Σx^2	9262
Mín	32
Q1	33
Mediana	34
Q3	35
Máx	36

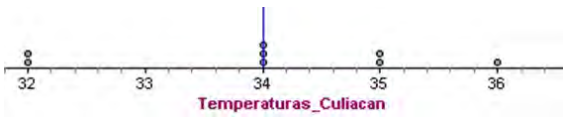
Medidas resumen sobre las temperaturas de Culiacán.

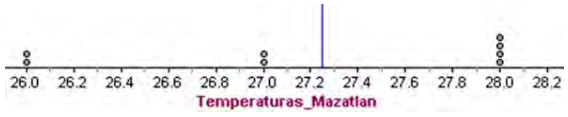


5.5 Medidas de variabilidad de una distribución de datos

Las medidas de centro no son suficientes para una adecuada descripción de los datos, ya que solo indican el valor central alrededor del cual se encuentran los datos. Es necesario calcular otras medidas que indiquen el grado de dispersión o variabilidad, para visualizar de forma más completa el comportamiento de los datos. Consideremos los datos de las temperaturas presentadas en Culiacán y Mazatlán en la semana del 23 al 30 de marzo de 2018.

Culiacán	35	32	32	34	36	35	34	34
Mazatlán	27	26	26	27	28	28	28	28





• Rango:

La medida de variabilidad más sencilla que existe es el rango. Representa la distancia entre el dato más pequeño y el dato más grande de la distribución.

Las temperaturas de Culiacán en el período considerado tienen un rango de 32-36=4 grados, las temperaturas de Mazatlán tuvieron un rango de 26-28=2 grados. Es decir, las temperaturas de Culiacán tuvieron mayor variabilidad. Obsérvese que el rango depende solo de los datos extremos, por lo que resulta muy sensible a datos atípicos; además no considera al centro de los datos, que debe ser la referencia respecto a la cual se mide la variabilidad.

Ejemplos de sus aplicaciones los encontramos cuando se reportan temperaturas máximas y mínimas, índices máximos y mínimos de contaminación, máximos y mínimos de la bolsa de valores, o en cartas de control de calidad de un producto, donde se reportan los extremos de la variable que se está midiendo en un período de tiempo. En todos estos casos el rango proporciona una visión global de la variabilidad.

PRINCIPALES ÍNDICES DEL BOLSA MEXICANA DE VALORES

Nombre	Puntos		Var. (%)	Var. (puntos)	Hora
IPC MEXICO	47.191,87	▲	+1,09%	+507,81	19:04:01
IMC30	829,74	▲	+1,13%	+9,26	19:04:01
INMEX	2.803,13	▲	+0,97%	+26,93	19:04:01
IND HABITA	50,92	▲	+0,93%	+0,47	19:04:01
IRT COMP MX	489,37	▲	+1,12%	+5,43	19:03:47

http://www.eleconomista.es/mercados/bolsa-mexicana-valores

• Rango intercuartílico

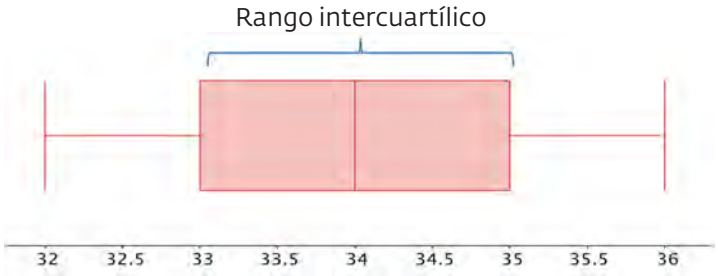
El rango intercuartílico se define como la diferencia entre el tercer cuartil (Q_3) y el primer cuartil (Q_1). Resulta menos sensible a datos extremos que el rango, porque no considera los extremos de la distribución. Se simboliza mediante IQR, entonces:

$$IQR=Q_3-Q_1$$

El rango intercuartílico de las temperaturas de Culiacán sería el siguiente:

$$IQR = 35 - 33 = 2$$

El rango intercuartílico representa la longitud de la caja (parte central) de un diagrama de caja (ver gráfica).



• Desviación estándar

La desviación estándar es la medida de variabilidad más utilizada en el análisis de datos y cumple un rol muy importante en los métodos de inferencia estadística.

Desviación

Iniciemos el análisis partiendo del concepto de *desviación*, al que definiremos como la distancia que existe entre un dato y la media aritmética de la distribución. La desviación puede ser negativa si el dato es menor que la media, positiva si es mayor que la media, e igual a cero si el dato coincide con la media (ver tabla).

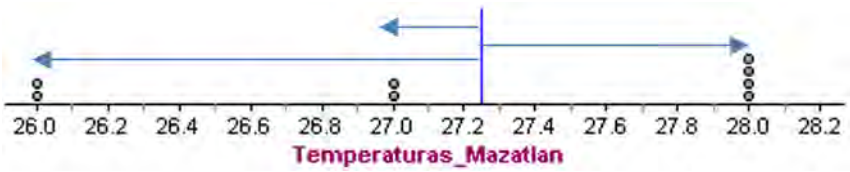


Figura. Diagrama de puntos con la media de los datos y sus respectivas desviaciones

Cálculo de desviaciones para las temperaturas de Mazatlán

Temperaturas	Media	Desviaciones
26	27.25	26-27.25= -1.25
26	27.25	26-27.25= -1.25
27	27.25	27-27.25=-0.25
27	27.25	27-27.25=-0.25
28	27.25	28-27.25=0.75
28	27.25	28-27.25=0.75

28	27.25	28-27.25=0.75
28	27.25	28-27.25=0.75
		$\sum(x-\bar{x})=0$

Obsérvese que *la suma de las desviaciones de los datos respecto a la media es igual a cero*. Esto es cierto para cualquier conjunto de datos, lo que representa una importante propiedad de la media aritmética.

Desviación cuadrática

En la búsqueda de una expresión para calcular la desviación estándar de una distribución de datos, introduciremos el concepto de *desviación cuadrática*, que consiste en elevar al cuadrado cada desviación para finalmente realizar la suma de ellas. Esto es:

$$\sum(x-\bar{x})^2=(-1.25)^2+(-1.25)^2+(-0.25)^2+(-0.25)^2+(0.75)^2+(0.75)^2+(0.75)^2+(0.75)^2=$$

$$\sum(x-\bar{x})^2=5.5$$

La suma de las desviaciones cuadráticas siempre resulta positiva, pues se obtiene de elevar al cuadrado cada una de las desviaciones y luego sumarlas.

Promedio de las desviaciones cuadráticas

Si la suma de las desviaciones cuadráticas se divide entre el total de datos menos 1, se obtiene la media de las desviaciones cuadráticas, a lo cual se le denomina **varianza** y constituye otra medida de variabilidad de los datos. $\frac{\sum(x-\bar{x})^2}{n-1}=\frac{5.5}{7}=0.79$

La varianza muestral se simboliza con la letra y la expresión para calcularla es la siguiente: $S^2=\frac{\sum(x-\bar{x})^2}{n-1}$

La raíz cuadrada a la varianza

En pasos anteriores se elevaron al cuadrado las desviaciones de los datos, se sumaron y se dividieron entre el total de datos menos 1, lo que dio lugar a una medida de variabilidad conocida como varianza; si ahora se extrae la raíz cuadrada a la varianza, se obtiene un resultado conocido como **desviación estándar**. $\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=\sqrt{0.79}=0.88$

La desviación estándar muestral se simboliza con la letra y la expresión para calcularla es la siguiente: $S=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

Para el cálculo de la varianza y la desviación estándar hay una pequeña diferencia, que se debe tener en cuenta cuando los datos provienen de muestras y cuando provienen de poblaciones.

Cuando los datos provienen de una muestra:

La desviación estándar se representa mediante la letra y la varianza mediante respectivamente.

$$S^2 = \frac{\sum(x-\bar{x})^2}{n-1} \qquad S = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Cuando los datos provienen de una población:

La desviación estándar se representa mediante la letra σ y se expresa de la siguiente manera:

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} \qquad \sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

μ es la media de la población
$\bar{x}$ es la media de la muestra
n es el total de datos en la muestra
N es el total de datos en la población
En el caso de la muestra el denominador es ***n-1*** en lugar de ***n***.

La razón de dividir entre y no entre ***n-1***, en el caso de la muestra, es porque se ha determinado que de esta manera se puede estimar, con mayor precisión, la media de una población en situaciones de inferencia estadística. Sin embargo, cuando la distribución contiene muchos datos, la diferencia entre ambas es casi imperceptible.

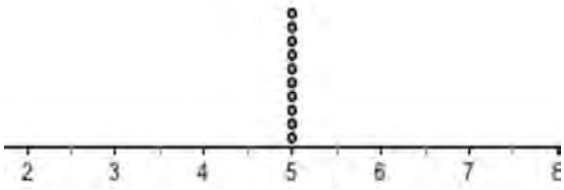
Interpretación de la desviación estándar

La interpretación de la desviación estándar no resulta tan evidente como en el caso de la media o la mediana; consideremos simplemente que es una medida de la dispersión de los datos, a mayor desviación estándar los datos están más alejados del centro de la distribución representado por la media aritmética.

Una regla empírica de mucha utilidad, que es válida para distribuciones de datos simétricas o aproximadamente simétricas, utiliza la desviación estándar y la media para caracterizar una distribución de datos.

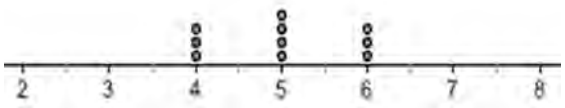
- 68% de los datos a encuentran a 1 desviación estándar de la media aritmética, esto es entre $\bar{x} - S$ y $\bar{x} + S$.
- 95% de los datos se encuentran a 2 desviaciones estándar de la media aritmética, esto es entre $\bar{x} - 2S$ y $\bar{x} + 2S$.
- 99% de los datos se encuentran a 3 desviaciones estándar de la media aritmética, esto es entre $\bar{x} - 3S$ y $\bar{x} + 3S$.

Para fijar ideas sobre la regla empírica considérese cuatro distintas distribuciones de datos, las cuales tienen la misma media aritmética y distinta desviación estándar.



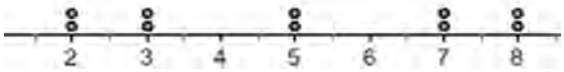
$\bar{x} = 5$

$s = 0$



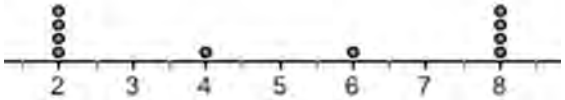
$\bar{x} = 5$

$s = 0.81$



$\bar{x} = 5$

$s = 2.40$



$\bar{x} = 5$

$s = 2.86$

En el primer caso la desviación estándar es igual a cero, pues todos los datos están concentrados en la media. En el segundo caso, la desviación estándar aumenta a 0.81. En el tercer caso los datos están aún más dispersos y la desviación aumenta a 2.40. En el último caso hay más datos alejados de la media y la desviación aumenta a 2.86.

Ideas importantes sobre la varianza y la desviación estándar

- La desviación estándar se mide en las mismas unidades de los datos, lo cual representa una ventaja de la desviación estándar respecto a la varianza que se mide en unidades al cuadrado.
- Si todos los datos son iguales, la varianza y la desviación estándar son iguales a cero, ya que no habría desviaciones de los datos respecto a la media.
- La desviación estándar es sensible a datos atípicos, unos cuantos datos atípicos pueden hacer que aumente mucho de valor.
- Una distribución muy asimétrica con una cola larga de datos tiene una desviación estándar muy grande. En estos casos puede ser más útil describir los datos por las 5 medidas resumen de los datos.
- La variabilidad no tiene que ver con la forma irregular de una distribución (en el sentido vertical) sino con la distancia de los datos al centro de la distribución. Por ejemplo, a muchas personas la primera gráfica les parece que tiene más variabilidad porque es más irregular que la segunda, lo cual no es necesariamente cierto.

Actividad de aprendizaje

En el capítulo 4 se expusieron datos sobre los índices máximos diarios de contaminación por ozono que se presentaron en las diferentes zonas de la ciudad de México en el mes

de diciembre de 2017. Los datos fueron analizados en una primera etapa por medio de distribuciones de frecuencias y algunas gráficas que mostraron el comportamiento de la contaminación. Ahora continuaremos el análisis de los datos mediante el cálculo de medidas descriptivas de centro y variabilidad.

Imecas Máximos Diarios de Ozono por Zonas

Diciembre 2017

Día	Zonas				
	Noroeste	Noreste	Centro	Suroeste	Sureste
1	46	63	67	65	73
2	63	115	101	108	94
3	106	104	111	118	112
4	106	117	115	119	120
5	80	80	103	102	107
6	63	86	80	104	113
7	100	65	49	51	45
8	33	34	35	31	37
9	36	44	46	43	44
10	42	46	55	46	47
11	65	82	48	48	59
12	44	43	46	45	45
13	102	108	109	107	104
14	117	102	114	117	110
15	76	45	49	53	41
16	27	29	29	29	33
17	41	47	51	65	94
18	113	100	106	112	100
19	104	92	104	108	86
20	86	100	80	105	104
21	92	96	86	94	107
22	98	82	80	102	67
23	73	80	96	105	103
24	78	103	110	94	117
25	71	57	92	73	88
26	50	63	82	76	104
27	84	73	104	96	102

28	76	78	69	90	88
29	63	57	59	67	63
30	80	88	82	94	86
31	103	105	108	104	103

Fuente: http://www.aire.cdmx.gob.mx/default.php?opc='aqBjnmU='

Los cálculos fueron realizados con el software *Geogebra* y los resultados se muestran a continuación:

Cuadro. Medidas estadísticas básicas proporcionadas por *Geogebra* sobre los datos de contaminación de la ciudad de México

	n	Media	σ	s	Mín	Q1	Mediana	Q3	Máx
Noroeste	31	74.7742	25.4009	25.8208	27	50	76	100	117
Noreste	31	76.9032	24.9959	25.4091	29	57	80	100	117
Centro	31	79.5484	26.1409	26.573	29	51	82	104	115
Suroeste	31	82.9355	27.6685	28.1258	29	53	94	105	119
Sureste	31	83.7419	27.0185	27.4651	33	59	94	104	120

La primera columna (n) indica el total de reglones (días del mes), la columna (Media) proporciona la media aritmética de los índices de ozono en cada zona; se observa que la zona Noroeste tuvo el menor índice con una media de 74.77 Imecas, mientras que la zona Sureste tuvo el mayor valor con una media de 83.74 Imecas. La mediana indica también que la zona menos contaminada por ozono fue la Noroeste, y las más contaminadas fueron las zona Suroeste y Sureste.

La columna (s) representa la desviación estándar de los datos y se observa que la zona Noreste tuvo menor variabilidad en los índices de ozono con un valor de 25.82 Imecas, mientras que la zona Suroeste presentó la mayor variabilidad en sus índices. El análisis de los datos muestra además los valores mínimos y máximos de Imecas en cada zona y el cuartil 1 y cuartil 2.

Para tu reflexión

¿Cuál es mejor, la media o la mediana?, suele ser una pregunta que muchos estudiantes se hacen una vez que han estudiado los dos promedios. La respuesta es: depende de la distribución de los datos. La media puede ser un buen promedio cuando la distribución de los datos no es muy sesgada o asimétrica, en caso contrario la mediana puede ser un mejor resumen del centro de la distribución.

Hay muchos casos donde se ha generado disputa sobre la conveniencia de la media o la mediana. Por ejemplo, en Inglaterra en 2005 se generó un debate importante sobre el tema, cuando se analizó si el ingreso de los hogares había disminuido o no. El Instituto de Estudios Fiscales elaboró un reporte en el cual se establecía que la "media

de ingreso real por hogar" había disminuido un 0.2% respecto del año anterior.

Cuando esto fue publicado en los medios de comunicación, algunos comentaristas criticaron al gobierno por los malos resultados. Gordon Brown, que era el canciller en ese momento, intentaba explicar que la mediana es la medida para calcular el ingreso promedio de los habitantes porque la distribución es sesgada, ya que son muchas más personas las que tienen un bajo salario que las que tienen un salario alto.

El informe del Instituto señalaba "esta es la primera vez que los ingresos han caído desde la recesión a principios de la década de 1990", cuando lo correcto hubiera sido "esta es la primera vez que los ingresos medios han disminuido". El cálculo con la mediana demuestra que el ingreso promedio en realidad aumentó ligeramente. Ante ello, o los comentaristas de los medios no sabían que era incorrecto usar la media para calcular el ingreso promedio, o deseaban criticar al gobierno aun sabiendo que el cálculo era incorrecto.

La mayoría de las personas está muy familiarizada con la media, pues desde la escuela primaria aprenden a calcularla para determinar el promedio de las calificaciones, pero debemos estar conscientes que la mediana es mucho mejor en distribuciones de datos asimétricas, como es el caso de los salarios. Para muchas personas cuando se habla de promedio viene a su mente la media, pero no hay que olvidar que la mediana también es un promedio; incluso en la hoja de cálculo *Excel*, a la media se le denomina promedio indebidamente.

Nota histórica

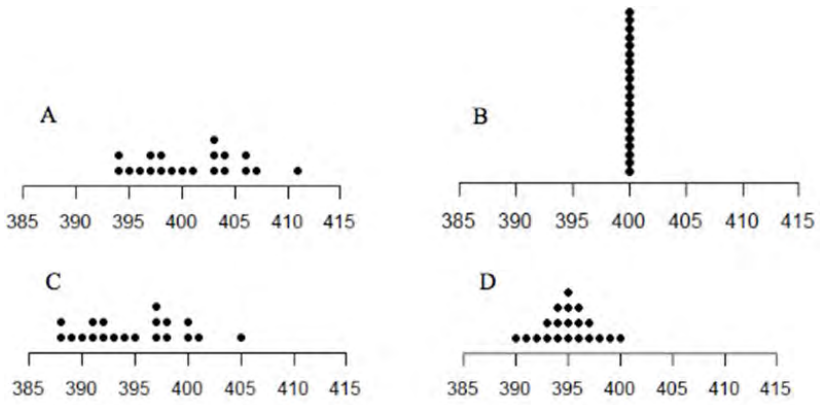
Un término menos usado para el cálculo de medidas descriptivas de un conjunto de datos es el de *agregación de datos*. En el siglo XIX se le conocía como "combinación de observaciones". Este enunciado expresa la idea que se lograba una ganancia de información más allá de lo que los valores individuales en un conjunto de datos dijeran al convertirlos en un resumen estadístico.

Calcular la media de un conjunto de datos era un paso bastante radical en el análisis, pues al sumar los datos y dividir la suma entre el total de ellos se pierde la individualidad de cada dato y el orden en que fue hecha la medición. Incluso después de que el cálculo de la media se volvió una práctica común, la idea de que descartar información puede aumentar la información no siempre fue convincente para todo el mundo.

En 1860 William Stanley Devons propuso medir los cambios en los precios por medio de un índice que se obtenía del promedio de los cambios porcentuales en distintas mercancías (similar a como se hace hoy en día para determinar el índice de precios al consumidor), sus críticos consideraron absurdo promediar los datos del hierro con los de la pimienta, señala Stigler en su libro *Los Siete Pilares de la Sabiduría Estadística*. El uso de la media aritmética como medida de agregación se empezó a generalizar a partir del siglo XVII; hoy día es la medida de centro de los datos más utilizada.

Evaluación del capítulo

1. Una fábrica de café envasa el producto en bolsas, proponiéndose un peso promedio de 400 gramos. Para verificar la calidad del proceso de envasado se seleccionan muestras aleatorias de 25 bolsas diariamente, y se registra el peso promedio en cada muestra. ¿Asumiendo que el proceso se realiza correctamente, cuál de las siguientes gráficas es más recomendable para el peso promedio en cada una de las 20 muestras seleccionadas?

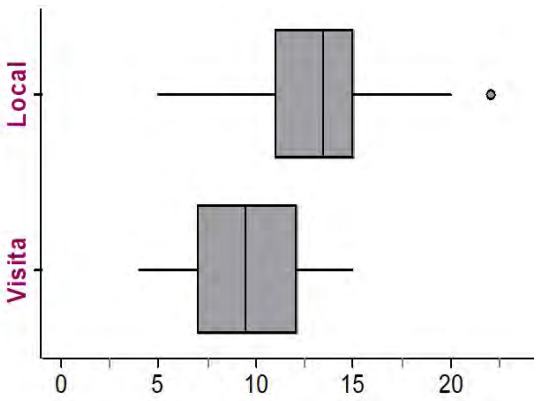


a) Gráfica A
b) Gráfica B
c) Gráfica C
d) Gráfica D

2. Un profesor aplica una prueba de ciencias de 15 ítems a un grupo de estudiantes. Para cada ítem un estudiante recibe 1 punto por respuesta correcta, 0 puntos por no respuesta y pierde 1 punto por cada incorrecta. El score total de la prueba varía de -15 puntos a 15 puntos. El profesor calcula la desviación estándar de la prueba y resulta igual a -2.30. ¿Qué conocemos a partir de dicho resultado?

a) La desviación estándar fue calculada incorrectamente.
b) La mayoría de los estudiantes recibieron puntajes negativos.
c) La mayoría de los estudiantes presentó puntajes por debajo de la media.
d) Ninguna de las respuestas de arriba.

3. Observa los siguientes diagramas de caja. Ellos muestran los puntos ganados, tanto de local como de visita, por todos los equipos de fútbol profesional en el torneo de apertura 2018. ¿Existe diferencia en los puntos ganados para las dos situaciones descritas? Explica tu respuesta.



4. Un pequeño objeto fue pesado en la misma báscula en forma separada por nueve estudiantes en una clase de ciencia. Los pesos (en gramos) registrados por cada estudiante se muestran a continuación:

6.2 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.2

Los estudiantes desean determinar de manera más exacta el peso real de este objeto ¿cuál de los siguientes procedimientos les recomendarías que utilizaran?

a) Usar la medida más común que se ha encontrado, que es 6.2

b) Usar el 6.15, dado que es el peso más exacto.

c) Sumar las 9 medidas y dividir el resultado entre 9.

d) Eliminar el 15.3, para sumar las otras 8 medidas y dividir el resultado entre 8.

5. Selecciona las afirmaciones que sean correctas.

- Para calcular la media aritmética se requiere ordenar los datos.
- La cantidad de datos que se ubican por debajo o por encima de la media no siempre es la misma, depende de la forma de la distribución de los datos.
- La media es muy sensible a datos extremos o atípicos. Esto significa que datos muy alejados del conjunto mueven a la media hacia su lado.
- Si todos los datos son iguales, la media es igual a cualquiera de ellos.
- La media y la mediana pueden ser iguales si la distribución de los datos es simétrica, o muy cercanas una de otra si la distribución es aproximadamente simétrica.
- La desviación estándar se mide en las mismas unidades de los datos. La varianza se mide en unidades al cuadrado.
- Si los datos son todos iguales, la varianza y la desviación estándar son iguales a cero, ya que no habría desviaciones de los datos respecto a la media.
- La desviación estándar es sensible a datos atípicos, unos cuantos datos atípicos la pueden hacer que aumente mucho de valor.
- Una distribución muy asimétrica con una cola larga de datos tiene una desviación estándar muy grande.

Bibliografía recomendada

- Comprensión y razonamiento de profesores de matemáticas de bachillerato sobre conceptos estadísticos básicos
  http://www.iisue.unam.mx/perfiles/articulo/2014-146-comprension-y-razonamiento-de-profesores-de-matematicas-de-bachillerato-sobre-conceptos-estadisticos-basicos.pdf
- Razonamiento estadístico de estudiantes universitarios sobre el análisis de datos en un ambiente computacional
  http://www.scielo.br/pdf/bolema/v28n50/1980-4415-bolema-28-50-1262.pdf
- Taller sobre análisis exploratorio de datos
  https://www.ugr.es/~batanero/pages/ARTICULOS/TallerAnadadi.pdf

Capitulo 6

Análisis de datos bivariados: correlación y regresión lineal

Comprender el rol de los modelos es una habilidad crítica para investigar la distribución de los datos y las relaciones entre variables.

Proyecto GAISE

6.1 Introducción

En los capítulos anteriores nos hemos enfocado en el análisis de datos que provienen de una sola variable, también conocido como *análisis univariado*. Sin embargo, es bastante frecuente que en un estudio estadístico se involucre más de una variable, con el propósito de determinar la existencia de alguna relación o asociación entre ellas. A este tipo de análisis se le conoce como *análisis multivariado*, y en particular, se le denomina *análisis bivariado* de datos al análisis de dos variables, del cual nos ocuparemos en este capítulo, También es común el título de *correlación y regresión*, e incluso *covariación* para denominar estos temas.

Al razonamiento que emerge del análisis bivariado de datos se conoce como *razonamiento covariacional*, el cual es de suma importancia en el razonamiento científico. En la vida cotidiana, la covariación también aparece con frecuencia en los medios de comunicación, por lo general a través de descripciones verbales y datos representados en forma tabular con dos o más variables en forma simultánea, por lo cual, el tema no es solo de interés de los científicos y profesionistas, sino que forma parte del bagaje de cultura estadística que deben tener todos los ciudadanos.

Un ejemplo de una nota periodística que requiere razonamiento covariacional fue publicada en el periódico Excélsior. En ella se señala que conforme aumenta la posibilidad de que Donald Trump gane la presidencia de los Estados Unidos en 2016, el peso mexicano disminuye su valor. Se proporcionan para ello datos en forma gráfica (ver figura 1).

Las variables estadísticas pueden ser cuantitativas o cualitativas, por lo que existen tres posibles combinaciones de pares de variables: *cualitativa versus cualitativa*,

cuantitativa versus cuantitativa y *cualitativa versus cuantitativa*. En este capítulo abordaremos el caso de relación entre dos variables cuantitativas.

Cuando analizamos la relación entre dos variables examinamos si ciertos valores de una variable corresponden a ciertos valores en otra variable. Por ejemplo, ¿existe relación entre el nivel de estudios de las personas y sus ingresos salariales?, ¿existe relación entre el promedio de las calificaciones de preparatoria con el resultado del examen de ingreso a licenciatura?, ¿cómo se puede describir dichas relaciones?, ¿se puede hacer una predicción sobre el nivel de ingresos conociendo el nivel de estudios de una persona? Este tipo de preguntas son naturales cuando se analizan datos bivariados. Cuando se determina que realmente existen patrones en los datos de las variables se dice que hay *relación estadística* entre las variables.

Figura 1. Periódico Excélsior 20/09/2016

Un análisis básico de relación entre dos variables cuantitativas involucra tres etapas:

1. Construir una gráfica para visualizar la relación entre las variables.
2. Calcular medidas numéricas que resumen a la relación entre las variables.
3. Construir un modelo de ajuste a los datos que permitan predecir una variable a partir de los datos de la otra variable.

La relación entre dos variables puede ser lineal, curvilínea o irregular. En este capítulo abordaremos la *relación lineal* entre dos variables, por lo que las medidas numéricas y el modelo de ajuste de los datos harán referencia a este caso.

6.2 Visualizando la relación entre dos variables cuantitativas: diagramas de dispersión

Un diagrama de dispersión es una gráfica para representar la relación entre dos variables cuantitativas. Los valores de una variable aparecen en el eje horizontal y los valores de la otra variable aparecen en el eje vertical. Para muchos propósitos – en particular cuando se desea hacer predicciones -, es importante distinguir cual variable es explicativa y cual es variable de respuesta. La *variable explicativa* se coloca en el eje horizontal y la *variable de respuesta* en el eje vertical.

Para dar contexto a lo anterior consideremos un estudio realizado por el INEGI en 2015, donde por primera vez realiza un estudio sobre *bienestar subjetivo* de los mexicanos vinculado con el *bienestar objetivo*. En este estudio se toma en cuenta no solo bienes y servicios para

medir el nivel de satisfacción con la vida, sino que considera la medición de bienes intangibles como los afectos, la familia, los amigos, sentimientos de logro y propósito en la vida entre otros.

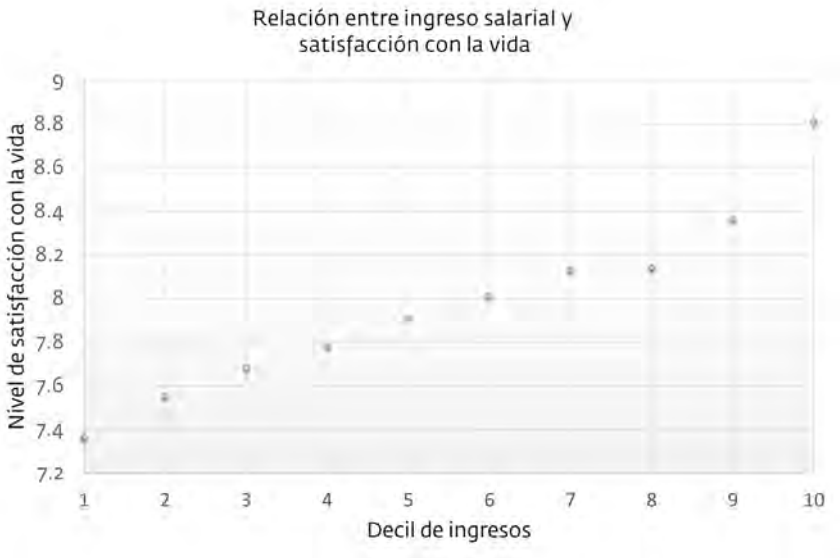
Uno de los resultados del estudio destaca la relación entre el nivel de ingresos de las personas y la satisfacción con la vida. Los ingresos se definen mediante deciles; el decil 1 representa el 10% de la población con menos ingresos, el decil 2 representa al 20% con menos ingresos, el decil 10 representa al 10% de la población con mayores ingresos (ver tabla 1).

Tabla 1. Relación entre ingreso salarial y promedio de satisfacción con la vida (2015)

Decil	Promedio de Satisfacción con la vida
I	7.36
II	7.55
III	7.68
IV	7.78
V	7.91
VI	8.01
VII	8.13
VIII	8.14
IX	8.36
X	8.81

Fuente: INEGI

Obsérvese la existencia de dos variables: ingreso salarial (expresado por medio deciles) y nivel de satisfacción con la vida (expresado como un número en una escala de 1 a 10). La *variable explicativa* es el ingreso salarial y la *variable de respuesta* es el nivel de satisfacción, porque el nivel de satisfacción depende del nivel de ingreso salarial. El diagrama de dispersión correspondiente a ambas variables se muestra en la gráfica 1 construida con *Excel*.



Gráfica 1: Relación entre ingreso salarial y promedio de satisfacción (2015)

Interpretación de un diagrama de dispersión

Para interpretar un diagrama de dispersión se deben tener en cuenta las siguientes características:

a) Dirección (positiva, negativa)
b) Forma (lineal, curvilínea, irregular)
c) Intensidad (débil, moderada, fuerte)
d) Agrupamientos
e) Puntos extremos

La dirección de la relación se refiere a cómo cambian los valores de una variable en relación con los cambios en la otra variable. Si los valores de una variable tienden a aumentar, conforme los valores de la otra variable también aumentan, se dice que entre ambas variables hay *dirección positiva*; por el contrario, si los valores de una variable tienden a aumentar, conforme los valores de la otra variable tienden a disminuir o viceversa, se dice que entre ambas variables hay *dirección negativa*.

Si la nube de puntos de un diagrama de dispersión tiende a parecerse en forma global a una línea recta pero no en forma exacta, se dice que la relación entre las variables tiene *relación lineal*. En caso contrario, podría haber una *relación curvilínea*, o bien, podría no haber una relación con forma identificable. Cuando en un diagrama de dispersión existe poca dispersión en la nube de puntos respecto a una línea recta imaginaria que pasa por el medio de la nube, se dice que hay una *intensidad fuerte* en la relación entre las variables. En caso contrario, se dice que entre las variables existe una *relación débil*. Si en el diagrama de dispersión se observan uno o más puntos alejados del resto de la nube de puntos se dice que existen *puntos extremos*. Si en vez de pocos datos se observan grupos que se separan entre sí en la nube de puntos, entonces existen *agrupamientos de datos*.

Para el caso del diagrama de dispersión sobre el nivel de satisfacción con la vida y el ingreso salarial de las personas (gráfica 1), la dirección es positiva, lo que sugiere que a mayor nivel de ingreso tiende a haber mayor nivel de satisfacción en la vida. La forma de la relación es lineal, no exactamente, pero con poca variación, mientras que la intensidad de la relación es moderada, dado que los puntos se aproximan bastante a una línea recta con pendiente intermedia. No se observan agrupamientos de datos ni puntos extremos.

6.3 Una medida numérica de la relación lineal entre dos variables: coeficiente de correlación lineal

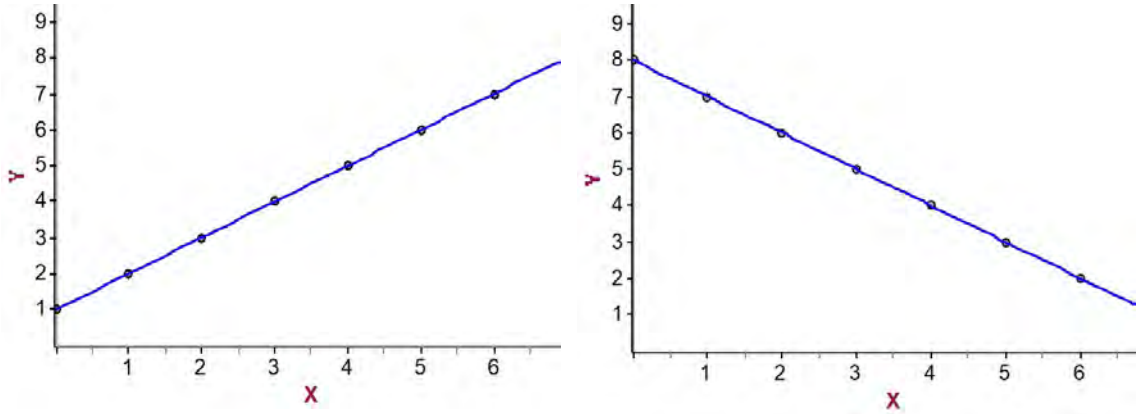
Construir un diagrama de dispersión es un buen punto de partida en el análisis bivariado de datos, pero no es suficiente. La siguiente etapa consiste en calcular medidas numéricas que describan con mayor precisión la relación entre las variables. La medida que describe en forma precisa la relación lineal entre dos variables cuantitativas

es el *coeficiente de correlación lineal*, también llamado *coeficiente de correlación de Pearson*. El coeficiente de correlación es una medida de la relación lineal entre dos variables cuantitativas. El coeficiente de correlación proporciona dos propiedades relevantes de la relación entre dos variables cuantitativas:

1. Dirección
2. Intensidad

El signo del coeficiente hace referencia a la dirección y el valor del coeficiente expresa la intensidad de la relación. Usualmente se simboliza con la letra r.

Una relación perfecta entre dos variables cuantitativas se observa cuando todos los puntos en un diagrama de dispersión caen sobre una línea recta. En este caso la desviación de cada punto respecto a la línea recta es igual a cero. Las siguientes figuras muestran los dos casos de relación perfecta que se pueden presentar:



r=1 r=-1

En el primer caso el coeficiente de correlación es igual a 1 y en el segundo diagrama es igual a -1. En la práctica es difícil encontrar una relación perfecta dada la variabilidad que caracteriza a los datos estadísticos, así que lo más común es encontrar coeficientes de correlación entre -1 y 1 pero no exactamente iguales a 1. Un coeficiente cercano a 0 indica que existe poca relación entre las variables.

Propiedades del coeficiente de correlación

- El coeficiente de correlación toma valores entre -1 y 1.
- El coeficiente de correlación no cambia cuando se cambian las unidades de medida de una o ambas variables. Por ejemplo, si dos variables se miden en centímetros y luego se convierten a pulgadas, el coeficiente de correlación es el mismo.
- El coeficiente de correlación no tiene en cuenta cuál es la variable explicativa y cuál es la variable de respuesta, si las invertimos el coeficiente no cambia.
- La correlación mide la fuerza de la relación entre dos variables respecto a una línea recta.
- La correlación es muy sensible a datos extremos.

Cálculo del coeficiente de correlación

El coeficiente de correlación involucra a su vez a diversos conceptos estadísticos, por lo que su cálculo resulta laborioso, particularmente cuando se tienen muchos datos. Mencionamos anteriormente que, en una correlación perfecta entre dos variables, los puntos quedan perfectamente colocados en una línea recta, lo que implica que la desviación de ellos a la recta de ajuste es igual a cero. Es decir, entre más cercanos estén los puntos a la línea recta habrá mayor correlación entre las variables.

En este sentido el concepto de desviación es parte importante en el cálculo del coeficiente de correlación, y como la desviación se mide respecto a un promedio, la media aritmética también es otro concepto importante que forma parte de la fórmula, así como la desviación estándar que mide la variabilidad, como se muestra a continuación:

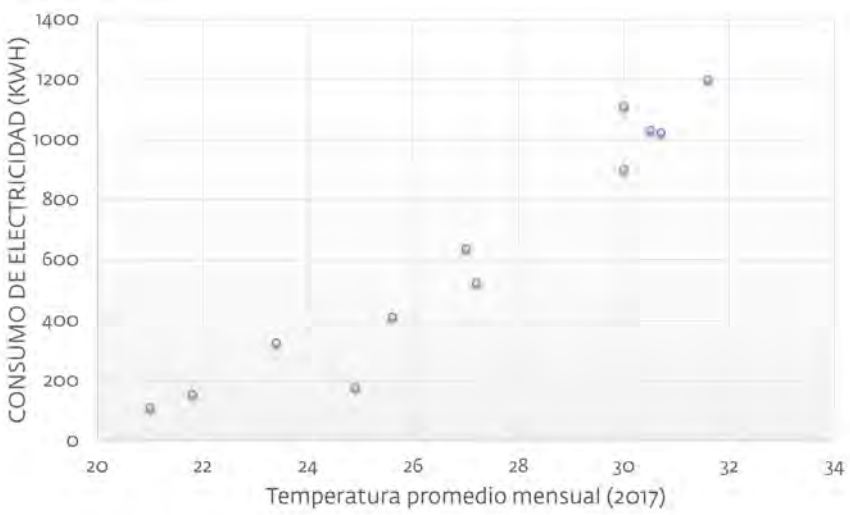
$$r = \frac{\sum \left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right)}{n - 1}$$

x: valores de la variable 1
y: valores de la variable 2
$\bar{x}$: media aritmética de la variable 1
$\bar{y}$: media aritmética de la variable 2
S_x: desviación estándar de la variable 1
S_y: desviación estándar de la variable 2
n: tamaño de la muestra

En cualquier herramienta de software, hoja de cálculo y en algunas calculadoras se encuentra disponible una función que calcula el coeficiente de correlación. Para fijar ideas consideremos los siguientes datos, que representan las temperaturas promedio mensuales que se presentaron en la ciudad de Culiacán en el año 2017 y el consumo de electricidad en el hogar del autor de este libro que vive en la ciudad.

Tabla. Temperatura promedio mensual y consumo de energía

Mes	Temperatura promedio	Consumo energía (Kwh)
Enero	21.0	111
Febrero	21.8	156
Marzo	24.9	178
Abril	25.6	412
Mayo	27.2	525
Junio	30.0	901
Julio	31.6	1,199
Agosto	30.7	1,023
Septiembre	30.0	1,111
Octubre	30.5	1,032
Noviembre	27.0	637
Diciembre	23.4	326



Los cálculos para determinar el coeficiente de correlación son los siguientes:

Variable	**Media aritmética**	**Desviación estándar**
Temperatura (x)	$\bar{x} = 26.97$	$S_x = 3.66$
Consumo electricidad (y)	$\bar{y} = 634.25$	$S_y = 403.81$

Temperatura promedio (x)	Consumo energía (Kwh) (y)	$(x - \bar{x})$	$(y - \bar{y})$	$\frac{(x - \bar{x})}{S_x}$	$\frac{(y - \bar{y})}{S_y}$	$\frac{(x - \bar{x})}{S_x}\frac{(y - \bar{y})}{S_y}$
21.0	111	-5.97	-523.25	-1.63	-1.30	2.11
21.8	156	-5.17	-478.25	-1.41	-1.18	1.66
24.9	178	-2.07	-456.25	-0.56	-1.13	0.63
25.6	412	-1.37	-222.25	-0.37	-0.55	0.20
27.2	525	0.23	-109.25	0.06	-0.27	-0.01
30.0	901	3.03	266.75	0.83	0.66	0.54
31.6	1,199	4.63	564.75	1.26	1.40	1.76
30.7	1,023	3.73	388.75	1.01	0.96	0.96
30.0	1,111	3.03	476.75	0.83	1.18	0.98
30.5	1,032	3.53	397.75	0.96	0.98	0.94
27.0	637	0.03	2.75	0.008	0.00	0.00
23.4	326	-3.57	-308.25	-0.97	-0.76	0.73
Suma						**10.5**

$$r = \frac{\Sigma\left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right)}{n - 1} = \frac{10.5}{11}$$

$$r = 0.96$$

El reporte del análisis de los datos del software *Geogebra* coincide con los cálculos que proporciona la fórmula.

MediaX	26.975
MediaY	634.25
Sx	3.6619
Sy	403.8175
r	0.9617
ρ	0.9492
nVarX	147.5025
nVarY	1793754.25
nCov	15642.375

Reporte de resultados del software *Geogebra*

Como puede observarse, los cálculos para determinar el coeficiente de correlación son laboriosos, aun para un pequeño conjunto de datos. La tecnología permite realizar dichos cálculos de manera rápida y precisa con solo introducir los datos e indicar la función correspondiente. En la hoja de cálculo Excel se dispone de una función para calcular el coeficiente de correlación, al igual que en el software *Geogebra*. El coeficiente de correlación resulta ser igual a 0.96, lo cual coincide con el cálculo que se ha realizado siguiendo la fórmula.

Actividad de aprendizaje

En la siguiente tabla se han recopilados datos de tres variables para cada una de las entidades federativas: participación porcentual al PIB, grado de escolaridad y porcentaje de la población con pobreza.

Entidad federativa	Participación al PIB 2016 (%)	Grado de escolaridad 2015	% de la población con pobreza 2016
Aguascalientes	1.4	9.7	28.2
Baja California	3.3	9.8	22.2
Baja California Sur	0.8	9.9	22.1
Campeche	2.2	9.1	43.8
Coahuila	3.7	9.9	24.8
Colima	0.6	9.5	33.6
Chiapas	1.7	7.3	77.1
Chihuahua	3.4	9.5	30.6
Ciudad de México	17	11.1	27.6
Durango	1.2	9.1	36.0
Guanajuato	4.2	8.4	42.4
Guerrero	1.4	7.8	64.4
Hidalgo	1.6	8.7	50.6
Jalisco	7.1	9.2	31.8

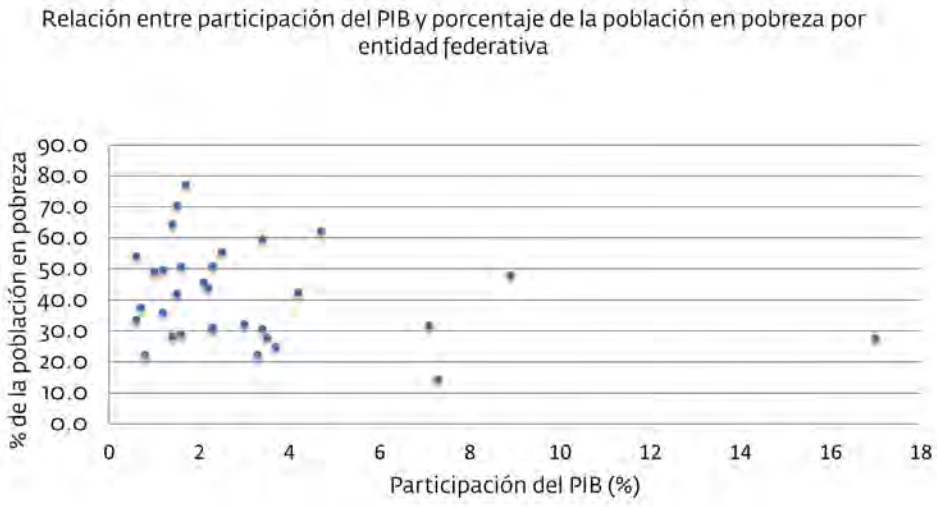
México	8.9	9.5	47.9
Michoacán	2.5	7.9	55.3
Morelos	1.2	9.3	49.5
Nayarit	0.7	9.2	37.5
Nuevo León	7.3	10.3	14.2
Oaxaca	1.5	7.5	70.4
Puebla	3.4	8.5	59.4
Querétaro	2.3	9.6	31.1
Quintana Roo	1.6	9.6	28.8
San Luis Potosí	2.1	8.8	45.5
Sinaloa	2.3	9.6	30.8
Sonora	3.5	10	27.9
Tabasco	2.3	9.3	50.9
Tamaulipas	3	9.5	32.2
Tlaxcala	0.6	9.3	53.9
Veracruz	4.7	8.2	62.2
Yucatán	1.5	8.8	41.9
Zacatecas	1	8.6	49.0

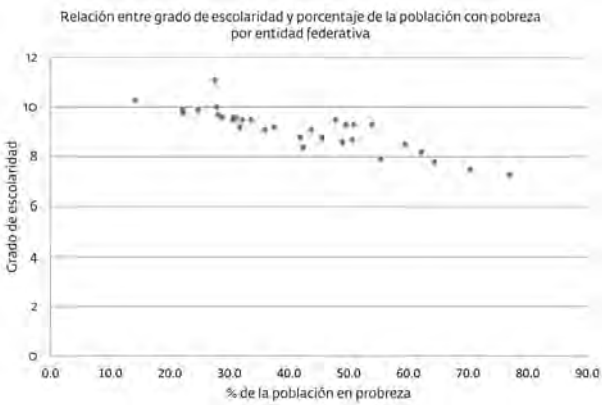
Fuente: INEGI y CONEVAL

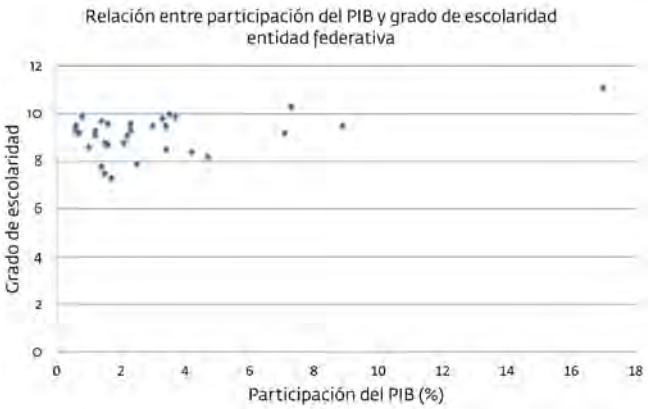
a) Interpreta los tres diagramas de dispersión tomando en cuenta la dirección de la relación, forma e intensidad, agrupamientos y datos extremos.

b) Realiza una estimación del coeficiente de correlación con solo visualizar los diagramas.

c) Calcula el coeficiente de correlación utilizando el software *Geogebra* y compara con la estimación que realizaste.



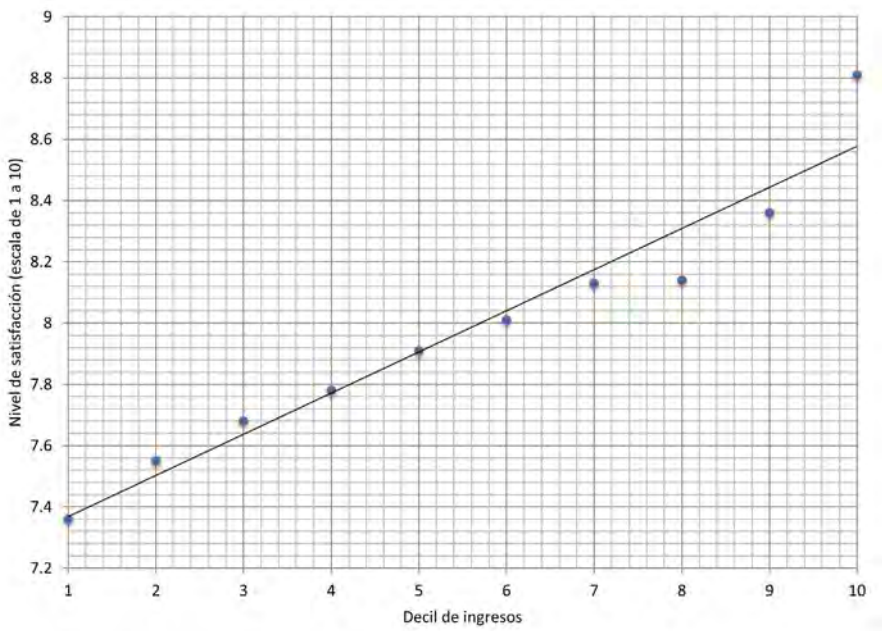




6.4 Modelación de la relación entre dos variables: regresión y predicción

Los diagramas de dispersión permiten visualizar la relación entre dos variables cuantitativas, el coeficiente de correlación lineal describe numéricamente dicha relación. Un paso más en el análisis consiste en construir un modelo matemático de la relación para *predecir* el valor de una variable a partir de los valores de la otra. A este proceso se le conoce como *regresión*, y en nuestro caso, que trata de la relación lineal entre dos variables, le denominaremos *regresión lineal simple*.

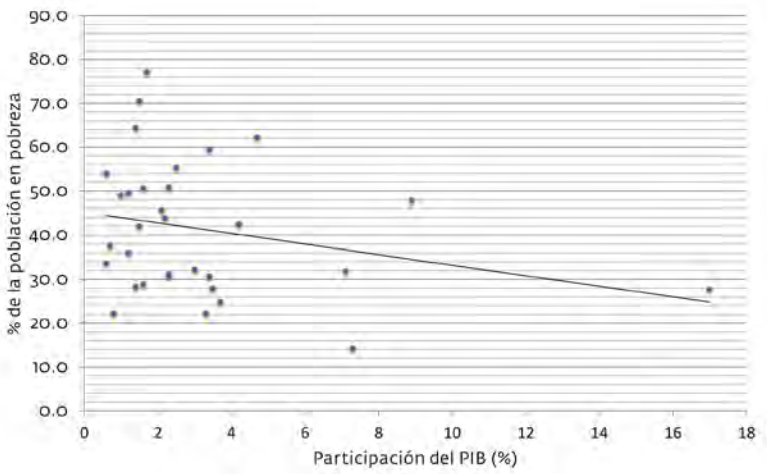
Las variables que deseamos predecir se llama *variable de respuesta*, usualmente denotada por la letra Y; la variable que proporciona los valores para la predicción se denomina *variable explicativa*, usualmente es denotada por la letra X. En el caso de la regresión lineal simple hay una sola variable explicativa y una variable de respuesta. En casos de regresión lineal múltiple puede haber diversas variables explicativas, pero también una sola variable de respuesta. Convencionalmente la variable explicativa se coloca en el eje de las X y la variable de respuesta en el eje Y del diagrama de dispersión.



Para fijar ideas, consideremos la relación entre el nivel de ingreso y satisfacción con la vida, cuyos datos

fueron obtenidos por el INEGI, los cuales hemos discutido anteriormente. El nivel de satisfacción es la variable de respuesta (Y) y el ingreso salarial es la variable explicativa (X).

El diagrama de dispersión muestra que los datos tienen un comportamiento aproximadamente lineal, que pueden resumirse mediante un ajuste a una línea recta la cual, hay que dejar claro, nunca resume la relación en forma perfecta. Sin embargo, existen casos donde la relación puede ser menos precisa que en el caso anterior. Véase por ejemplo la relación entre la participación en el PIB de los estados del país y el porcentaje de la población en pobreza (ver gráfica).



Un examen visual de la gráfica nos muestra que hay puntos que están muy alejados de la recta de regresión y la tendencia está lejos de ser lineal.

Matemáticamente es posible establecer una recta de regresión para cualquier conjunto dado de puntos en un diagrama de dispersión, sin embargo, muchas veces la recta puede no ser un buen modelo de predicción. Entonces, nos interesa establecer medidas numéricas que permitan evaluar la calidad de la predicción que se puede hacer con la recta de regresión. Una de estas medidas numéricas es el coeficiente de correlación.

MediaX	5.5
MediaY	7.973
Sx	3.0277
Sy	0.4198
r	0.9691
ρ	1
Sxx	82.5
Syy	1.586
Sxy	11.085
R^2	0.9391
SSE - Suma errores cuadrados	0.0966

Modelo de Regresión

Lineal $y = 0.13\,x + 7.23$

MediaX	3.125
MediaY	41.3625
Sx	3.221
Sy	15.2339
r	-0.251
ρ	-0.2389
Sxx	321.62
Syy	7194.195
Sxy	-381.78
R^2	0.063
SSE - Suma errores cuadrados	6741.0019

Modelo de Regresión

Lineal $y = -1.19\,x + 45.07$

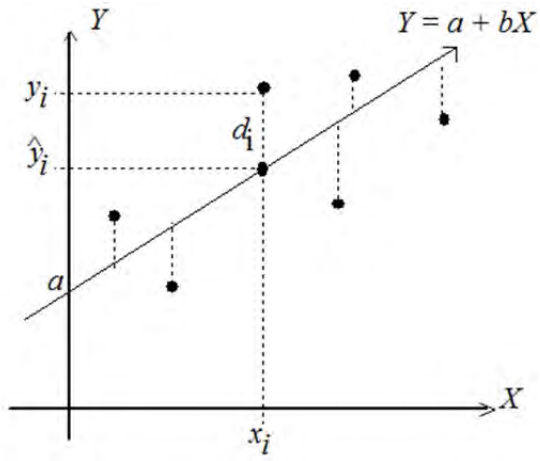
Análisis de correlación con *Geogebra*

Un análisis de regresión hecho en *Geogebra* muestra, entre otros resultados, un coeficiente de correlación de las variables ingreso y nivel de satisfacción igual a 0.96, y para el caso de las variables porcentaje de participación en el PIB y porcentaje de población en pobreza en las entidades federativas de México es de igual a -0.25. En el

primer caso se puede observar una alta correlación positiva entre las variables (muy cercana a 1), en el segundo caso la correlación negativa débil cercana a 0. El análisis anterior incorpora además la ecuación de regresión para ambos diagramas dispersión.

• Ecuación de la recta de regresión y su interpretación

Se pueden establecer muchas líneas rectas sobre los puntos de un diagrama de dispersión, pero nos interesa la línea recta que pase lo más cerca posible de todos los puntos, a esta recta se le conoce como *recta de mínimos cuadrados*. La recta de mínimos cuadrados es la línea que hace que la suma de los cuadrados de las distancias verticales () de los puntos a la línea recta sea la más pequeña de todas las posibles rectas que se pueden trazar sobre la nube de puntos.



La ecuación de la línea recta está determinada por y=a+bx**,** donde *x* es la variable explicativa mientras que *y* es la variable de respuesta. El valor de *a* representa el intercepto de la línea recta con el eje de la Y, el valor de *b* representa la pendiente de la recta y significa el cambio de la variable de respuesta (Y) por cada unidad de la variable explicativa (X).

• Interpretación de los coeficientes de la ecuación de regresión

De la figura 1 se observa que, para el caso de la relación entre ingreso salarial y satisfacción con la vida, la ecuación de regresión es y=0.13x+7.23. El valor de 0.13 significa que por cada unidad (decil de ingresos) de la variable explicativa, la variable de respuesta aumenta 0.13 unidades (porcentaje de satisfacción). En el caso de la relación entre porcentaje de participación del PIB y porcentaje de población con pobreza en cada entidad, la ecuación de regresión es y=-0.19+45.07, donde -0.19 significa que, por cada 1% de aumento de participación del PIB, la población en pobreza disminuye 0.19%. Hay fórmulas matemáticas para determinar la ecuación de regresión por mínimos cuadrados, sin embargo, en este capítulo nos apoyaremos en el uso de *Excel* y *Geogebra* para realizar los cálculos de manera automática como se muestra en los cálculos de la figura 1.

• Bondad de ajuste de la recta de regresión: el coeficiente de determinación

El coeficiente de correlación mide la intensidad de la relación entre dos variables cuantitativas, pero si el coeficiente se eleva al cuadrado, se obtiene el *coeficiente de determinación*que evalúa una propiedad importante de la relación entre dos variables, se representa por r^2.

El coeficiente de determinación r^2 se interpreta como la cantidad de variabilidad de Y que es explicada por la variable X. Dado que se obtiene de elevar al cuadrado el coeficiente de correlación y habiendo aclarado anteriormente que el máximo valor de es 1, el valor máximo de r^2 es 1, por lo tanto, valores cercanos a 1 indican que la variable X explica una buena parte de la variabilidad en la variable Y.

En la figura 1 se observa que, para la relación entre ingreso salarial y satisfacción con la vida, reporta un valor de r^2 =0.93, lo cual significa que la variabilidad en la satisfacción con la vida es explicada en un 93% por el nivel de ingresos. Ello significa que la ecuación y=0.13x+7.23 es un bueno modelo de predicción.

6.5 Resolución de un problema

La Dirección General de Evaluación Institucional de la UNAM recopila información de las universidades mexicanas y realiza comparativos de indicadores. La siguiente tabla muestra la cantidad de profesores y estudiantes que tenían 60 universidades mexicanas en 2016. Nos planteamos las siguientes preguntas para responder con los datos:

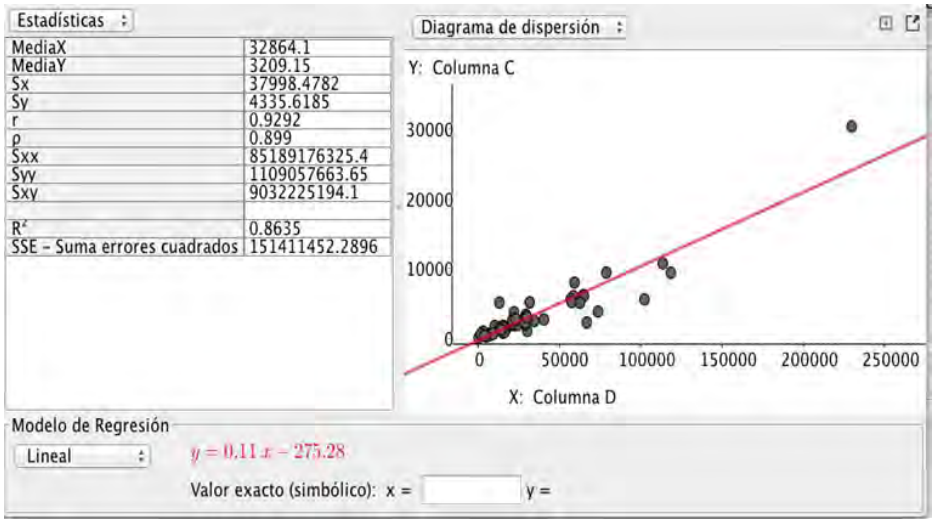
- ¿Existe alguna relación entre personal docente de las universidades y la matrícula?
- ¿Cómo es dicha relación y cómo se puede describir?
- ¿Es posible hacer una predicción de sobre el personal docente que requiere una universidad si conocemos su matrícula?

No.	Instituciones	Personal docente	Matrícula
1	Benemérita Universidad Autónoma de Puebla	2308	66992
2	Centro de Investigación y de Estudios Avanzados del IPN	610	2698
3	Colegio de Posgraduados	335	1170
4	Colegio de México	156	431
5	Instituto Politécnico Nacional	10757	113708
6	Instituto Tecnológico Autónomo de México	648	6278
7	Instituto Tecnológico de Sonora	1640	16101
8	Instituto Tecnológico de Estudios Superiores de Occidente	1829	10398
9	Instituto Tecnológico de Estudios Superiores de Monterrey	9451	79036
10	Sistema Universidad Anáhuac	3792	22245

11	Sistema Universidad del Valle de México	8017	59482
12	Sistema Universidad Iberoamericana	1921	22073
13	Sistema Universidad Lasalle	5189	31928
14	Universidad Autónoma Agraria Antonio Narro	542	5371
15	Universidad Autónoma Benito Juárez de Oaxaca	1436	17455
16	Universidad Autónoma de Chapingo	639	5814
17	Universidad Autónoma de Aguascalientes	1675	14984
18	Universidad Autónoma de Baja California	6244	64861
19	Universidad Autónoma de Baja California Sur	573	6095
20	Universidad Autónoma de Campeche	474	6820
21	Universidad Autónoma de Chiapas	2108	22075
22	Universidad Autónoma de Chihuahua	3409	29566
23	Universidad Autónoma de Ciudad Juárez	2365	28544
24	Universidad Autónoma de Coahuila	2316	24833
25	Universidad Autónoma de Guadalajara	1248	13296
26	Universidad Autónoma de Guerrero	1108	30384
27	Universidad Autónoma de la ciudad de México	830	15571
28	Universidad Autónoma de Nayarit	1366	16263
29	Universidad Autónoma de Nuevo León	5618	102407
30	Universidad Autónoma de Querétaro	2601	21176
31	Universidad Autónoma de San Luis Potosí	2082	28953
32	Universidad Autónoma de Sinaloa	3895	73877
33	Universidad Autónoma de Tamaulipas	2516	34716
34	Universidad Autónoma de Tlaxcala	1850	15473
35	Universidad Autónoma de Yucatán	988	16529
36	Universidad Autónoma de Zacatecas	1995	24380
37	Universidad Autónoma del Carmen	434	5758
38	Universidad Autónoma del Estado de Hidalgo	2911	28638
39	Universidad Autónoma del Estado de México	6085	58840
40	Universidad Autónoma del Estado de Morelos	2686	23811
41	Universidad Autónoma Metropolitana	5641	57092

42	Universidad de Colima	1426	13191
43	Universidad de Guadalajara	9429	118665
44	Universidad de Guanajuato	2824	23155
45	Universidad de las Américas Puebla	569	8093
46	Universidad de Monterrey	707	9480
47	Universidad de Quintana Roo	287	5207
48	Universidad de Sonora	2264	29375
49	Universidad del Ejército y Fuerza Aérea	1026	3220
50	Universidad Intercontinental	629	1836
51	Universidad Juárez Autónoma de Tabasco	3074	30186
52	Universidad Juárez del Estado de Durango	1637	15521
53	Universidad de Michoacana de San Nicolás de Hidalgo	2726	40641
54	Universidad Nacional Autónoma de México	30334	229831
55	Universidad Panamericana	5158	13203
56	Universidad Pedagógica Nacional	5228	57706
57	Universidad Popular Autónoma del Estado de Puebla	1291	14351
58	Universidad Regiomontana	402	4012
59	Universidad Tecnológica de México	6121	65281
60	Universidad Veracruzana	5129	62770

Las respuestas a las preguntas anteriores se pueden obtener haciendo un análisis de datos bivariados en *Geogebra*. La siguiente figura muestra los resultados.



Reporte de análisis bivariados con el software *Geogebra*

El diagrama de dispersión fue construido considerando como *variable explicativa* la cantidad de estudiantes (X) y como *variable de respuesta* el personal docente (Y), es decir, consideramos que el personal docente requerido en una universidad depende de la matrícula de estudiantes. Es importante tener claro el rol de cada de una las variables cuando se realiza análisis de regresión, pues si invertimos el rol de cada variable se da una ecuación de regresión distinta.

Interpretación de los resultados

El diagrama de dispersión muestra una relación positiva, lo cual quiere decir que a medida que se tienen más estudiantes las universidades tienen más personal docente. La intensidad de la relación medida por el coeficiente de correlación es r=0.92 , lo que indica que hay mucha relación entre ambas variables. Se observan dos grupos de universidades y una universidad que es punto extremo (UNAM). El primer grupo tiene menos de 50,000 estudiantes, el segundo grupo tiene entre 50,000 y menos de 120,000 estudiantes aproximadamente.

La forma de la relación puede ser ajustada mediante una línea recta por el método de mínimos cuadrados que reporta *Geogebra*. La ecuación es y = 0.11x - 275.28, traducida la ecuación a las variables del problema resulta:

Personal docente = 0.11 (matrícula) - 275.28

El valor de $r^2 = 0.86$, que significa la proporción de la variación de los valores del personal docente, es explicado por la matrícula, el otro 0.14 es explicado por distintos factores. El valor de es muy cercano a 1, que es el valor máximo, por lo que se considera que la ecuación de regresión que se ha construido es un buen modelo de predicción.

Si una universidad tiene 80,000 estudiantes, ¿qué personal docente le corresponde?

Personal docente = 0.11 (matrícula) - 275.28

Personal docente = 0.11 (80,000) - 275.28=8,524.72 profesores

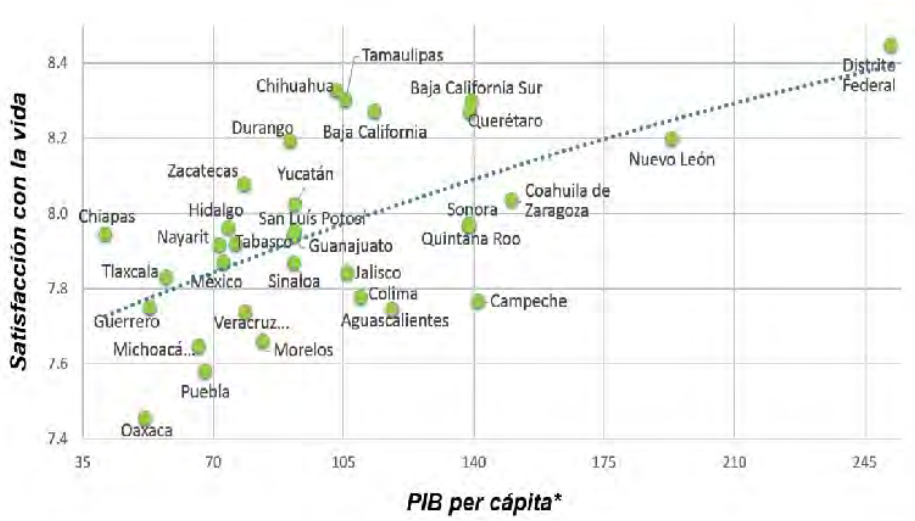
Resumiendo

La relación entre dos variables puede ser visualizada a través de un diagrama de dispersión, el coeficiente de correlación mide la intensidad o fuerza de la relación. Si la intensidad es alta (cercana a -1 o 1) el siguiente paso consiste en construir un modelo de predicción, que en nuestro caso es lineal. El valor del coeficiente de determinación nos proporciona una medida de la calidad del modelo construido. Por último, se procede a realizar predicciones para la variable de respuesta introduciendo valores de la variable explicativa en el modelo.

Actividad de aprendizaje

El siguiente diagrama de dispersión fue tomado de una publicación del INEGI. En

dicho diagrama se relaciona el PIB per cápita de cada estado del país con el nivel de satisfacción con la vida.



PIB per cápita y satisfacción con la vida por entidad federativa referente a 2013 (base 2008)

a) ¿Existe relación entre PIB per cápita y nivel de satisfacción con la vida? Interpreta el diagrama de dispersión considerando la dirección, forma, intensidad, agrupamientos y datos extremos.

b) ¿Cómo explicas el caso de Chiapas, que tiene el menor PIB per cápita, pero tiene un nivel de satisfacción mayor que varios estados con mayor PIB?

c) Realiza una estimación del coeficiente de correlación que existe entre ambas variables.

Para tu reflexión

Correlación y causalidad

Es bastante frecuente una concepción errónea que consiste en creer que, si dos variables están correlacionadas, una debe ser causa de la otra. Sin embargo, la correlación entre dos variables no significa que hay relaciones causales entre una y otra, aunque la correlación sea muy alta. Para determinar relaciones causales entre dos variables se debe recurrir a otro tipo de análisis como sería el caso de un estudio experimental.

Por ejemplo, fumar y tener cáncer de pulmón son variables que según estudios médicos están altamente correlacionadas. A través de diversos experimentos se ha mostrado que fumar es una causa posible de padecer cáncer de pulmón. Un ejemplo contrario podría ser el análisis de la contaminación en la Ciudad de México, en un período de tiempo, y el comportamiento de la Bolsa Mexicana de Valores (BMV) en ese mismo período.

Supongamos que se ha encontrado una tendencia al alza de la BMV a medida que disminuyen los índices de contaminación en la ciudad, ello quiere decir que ambas

variables están correlacionadas de manera negativa. El sentido común nos dice que no hay relación causal entre ambas variables, pues nada tiene que ver una variable con la otra.

Nota histórica

Francis Galton (1822-1911) científico inglés interesado en la teoría de la evolución de Charles Darwin (1809-1882) y destacado impulsor de la Biometría en el siglo XIX, es considerado el fundador de la teoría de la correlación y la regresión. De acuerdo con Stephen Stigler, historiador de la estadística, la teoría de la evolución estaba incompleta cuando se publicó en 1859 y seguía incompleta hasta 1882, año en que murió Darwin.

Galton fue el primero en observar la regularidad estadística, conocida como *regresión a la media*, al comparar características físicas de hijos y padres para determinar en qué medida variables como la estatura eran heredables. La regresión a la media, en este contexto, consiste en que padres, que en promedio son más altos, tienden a tener hijos en promedio más bajos y viceversa. Galton demostró que la regresión hacia la media era consecuencia de la relación imperfecta en la correlación entre las estaturas de padres e hijos. Esta misma tendencia se observa en otros contextos, como la inteligencia, el gusto por el arte, el deporte.

Sin pérdida de generalidad, la regresión a la media consiste en que cuando dos variables no están perfectamente correlacionadas los valores extremos de una de ellas se asocian con valores menos extremos de la otra.

Los métodos de regresión y correlación tuvieron un profundo impacto en el desarrollo de investigación en la biología, psicología, economía y otras ciencias durante el siglo XX, y son la base para la construcción de modelos de predicción multivariados en una diversidad de áreas del quehacer humano.

Evaluación del capítulo

1. El siguiente diagrama de dispersión muestra la relación entre la tasa de mortalidad infantil y la esperanza de vida en un grupo de países. Observa la ecuación de la recta de regresión y responde los siguientes incisos.

- El valor -0.36 de la ecuación de la recta de regresión significa:
  - **a)** El coeficiente de correlación entre las dos variables.
  - **b)** Por cada unidad (año) que aumenta la esperanza de vida, la tasa de mortalidad infantil disminuye 0.36
  - **c)** Por cada unidad que aumenta la tasa de mortalidad infantil, disminuye 0.36 años la esperanza de vida.
- El valor 79.3 de la ecuación de la recta de regresión significa:

a) La pendiente de la recta

b) El valor de la esperanza de vida cuando la tasa de mortalidad sea igual a cero

c) El coeficiente variabilidad entre las dos variables

El valor de r^2 =0.86 significa:

d) El coeficiente de correlación entre las dos variables

e) La variación de la esperanza de vida explicada por la tasa de mortalidad infantil.

f) La variación de la tasa de mortalidad infantil explicada por la esperanza de vida

- La recta de regresión trazada sobre la nube de puntos cumple con el siguiente criterio:

a) La mitad de los puntos siempre están por encima de ella y la otra mitad por debajo

b) La suma de los cuadrados de las distancias verticales de cada punto a la recta es mínima

c) La mayoría de los puntos debe estar sobre la recta de regresión

2. De las siguientes afirmaciones, selecciona las que sean correctas.

- Un coeficiente de correlación igual a 1 entre la estatura y el peso en un grupo de niños significa que la estatura está relacionada con el peso, pero un coeficiente de correlación de -1 significa que la estatura no está relacionada con el peso.
- Si un coeficiente de correlación igual a cero entre dos variables significa que no hay relación lineal entre ellas.
- Si existe una correlación alta entre dos variables cuantitativas significa que una es causa de la otra.
- La recta de regresión es la que mejor ajusta a la nube de puntos, en el sentido de que es la que proporciona una suma de mínimos cuadrados de error.
- Cuando existe una correlación alta entre dos variables generalmente los errores de estimación son pequeños.
- Un coeficiente de correlación alto entre dos variables no significa necesariamente que una variable dependa de la otra, solo indica que están relacionadas.

Bibliografía recomendada

- Análisis de datos bivariados en un ambiente basado en applets y software dinámico. http://www.revista-educacion-matematica.org.mx/descargas/Vol28/3/3.pdf
- Investigación didáctica en regresión y correlación https://www.ugr.es/~batanero/documentos/Investigacion.pdf
- Análisis de datos y su didáctica https://www.ugr.es/~batanero/pages/ARTICULOS/Apuntes.pdf
- La educación estadística en América Latina: Tendencias y perspectivas http://saber.ucv.ve/jspui/handle/123456789/4666
- Algunas notas históricas sobre la correlación y regresión y su uso en el aula https://core.ac.uk/download/pdf/20343739.p

Capitulo 7

Introducción a la probabilidad

Es extraordinario que una ciencia que empezó con la importancia de un juego se haya elevado a los más importantes objetos del conocimiento humano.

Pierre Simón de Laplace

7.1 Introducción

El azar y la incertidumbre son parte de nuestra vida diaria, desde situaciones sencillas como los juegos de azar, lanzar una moneda para elegir el lado de la cancha en un partido de futbol, realizar una rifa seleccionando una muestra aleatoria de números de una pequeña bolsa, comprar boletos de alguna lotería o sorteo, hasta situaciones más complejas, como decidir sobre la compra de algún tipo de seguro (por ejemplo: gastos médicos, vida, casa, automóvil), interpretar información meteorológica, interpretar los resultados de una encuesta basada en muestreo y análisis de riesgos de una inversión económica. En concordancia con ello, muchas notas periodísticas hacen referencia a eventos inciertos y utilizan la probabilidad, veamos algunas:

La probabilidad de que se produzca un terremoto en la misma fecha y con 32 años de diferencia es del 5% en un país como México, donde se registra una media de dos potentes sismos al año con una intensidad superior a 7 en la escala de Richter, explicó Efe Vala Hjorleifsdottir, investigadora del Instituto de Geofísica de la UNAM.

Periódico Excelsior 23/09/2017

En 2016 murieron 685,766 mexicanos, de los cuales 36,726 tenían un rango de edad de 50 a 54 años de edad; basados en estos datos, la probabilidad de muerte de un mexicano en este rango de edad es de 5.35%.

Instituto Nacional de Estadística y Geografía (INEGI)

Una muestra aleatoria de 800 mexicanos con alcance nacional tomada en febrero de 2018 reporta con una confianza de 95% y un margen de error de 3.5% que el 35% de los mexicanos considera que el tema más importante en la elección presidencial de 2018 es la inseguridad, el 32% considera a la economía y el 26% la corrupción.

Encuestadora Parametría.

Como puede verse en los enunciados anteriores, el razonamiento probabilístico es importante para comprender y dar sentido a información que se vierte a diario en los medios de comunicación. Sin embargo, como veremos más adelante, la probabilidad no es intuitiva, y más aún, se han documentado una diversidad de concepciones erróneas en torno a diversos conceptos probabilísticos.

La teoría de la probabilidad tiene un extenso campo de estudio, sin embargo, en este capítulo abordaremos sólo aspectos básicos de las leyes de la probabilidad que se requieren para estudiar la inferencia estadística, particularmente, los aspectos necesarios para comprender la aleatoriedad de los métodos de muestreo y los experimentos aleatorizados. Es decir, nuestro estudio de la probabilidad será como elemento de apoyo para la inferencia estadística.

7.2 ¿Cómo cuantificar la incertidumbre?: la idea intuitiva de probabilidad

Introduciremos la idea de probabilidad analizando un sencillo fenómeno aleatorio: el lanzamiento de una moneda. Partimos del hecho que no se puede predecir con certeza el resultado que ocurrirá en el siguiente lanzamiento, ya que variará conforme la moneda sea lanzada, bajo la posibilidad de ser "águila" o "sol". Sin embargo, los resultados de los fenómenos aleatorios se caracterizan por presentar un patrón regular (regularidad estadística) conforme estos se repiten muchas veces en las mismas condiciones.

Gráfica 7.1. Frecuencias relativas del resultado "águila" en el lanzamiento de una moneda obtenida mediante simulación.

Podemos lanzar una moneda 1,000 veces, pero en lugar de ello recurriremos a un método conocido como simulación, aprovechando el potencial de la tecnología. La gráfica 7.1 muestra el comportamiento de la frecuencia con la que aparece el resultado "águila" después de simular el lanzamiento de una moneda equilibrada 1,000 veces. En los primeros 100 lanzamientos se observa mucha variabilidad en la frecuencia relativa, pero esta tiende a estabilizarse conforme se incrementan las repeticiones; para 1,000 lanzamientos se tiene una frecuencia de águilas aproximada a 0.5. Decimos entonces que la probabilidad es muy cercana o quizá igual a 0.5. La gráfica muestra la regularidad estadística que mencionamos anteriormente, y que caracteriza a los fenómenos aleatorios.

En esta perspectiva la *probabilidad se puede considerar como la proporción o frecuencia relativa con la que ocurre un resultado después de una larga serie de repeticiones.* Desde este enfoque podemos obtener una buena estimación conforme se incrementa el número de observaciones de un fenómeno. Un hecho importante que se deduce de lo anterior es que el azar no es sinónimo de caos; el azar tiene un orden que se puede apreciar conforme un fenómeno aleatorio se repite u observa una gran cantidad de veces. Este principio se puede aplicar en el análisis de diversos fenómenos del mundo real para realizar estimaciones de la probabilidad vía frecuencias relativas.

Por ejemplo, se miden variables climatológicas relacionadas con la lluvia y se toma nota si llovió o no llovió, después de muchas observaciones es posible estimar la probabilidad de lluvia con base en dichos registros. Otro ejemplo: se registra la edad de defunción de las personas de un país por un largo tiempo, y con base en ello se puede estimar la probabilidad de fallecimiento de una persona de cierta edad. El primer ejemplo es muy importante para la predicción del clima, el segundo, para las compañías de seguros.

Abordemos ahora un enfoque distinto sobre la moneda en particular y sobre los eventos aleatorios en general. Cualquier persona a quien se le pregunte por la probabilidad de "águila" o "sol" dirá que es 0.5, pues solo hay dos resultados posibles; esto es cierto siempre que ambos resultados sean igualmente probables, lo que deriva del hecho que la moneda esté equilibrada o simétrica. No es difícil que una moneda sea exactamente simétrica, y podemos asumir de forma razonable que lo sea.

En el caso de un dado se puede hacer la misma suposición, entonces la probabilidad de una cara en particular es 1/6. En ambos casos estamos haciendo una idealización de la realidad considerando el principio de equiprobabilidad o simetría en los resultados de la moneda y el dado. Desde esta perspectiva *la probabilidad se puede considerar como el valor que se obtiene de dividir la cantidad de resultados*

favorables a un evento particular entre el total de resultados posibles del evento.

El modelo de simetría nos proporciona un modelo de probabilidad que puede ser muy útil para muchos casos, pero no para todos. Un modelo de probabilidad como el descrito tiene dos componentes esenciales: el espacio muestral, que consiste en la lista de todos los resultados posibles que el fenómeno o experimento puede tener y las probabilidades de cada uno de los resultados. Ejemplo: Al lanzar un dado no se sabe qué número caerá, pero a pesar de lo anterior, está determinado el conjunto de los posibles resultados {1, 2, 3, 4, 5, 6} y, además, la experiencia se puede repetir cuantas veces se desee en condiciones similares.

En muchas situaciones se puede aplicar el enfoque de frecuencias relativas para estimar la probabilidad de un evento, en otras ocasiones se puede utilizar el modelo de simetría o equiprobabilidad; incluso hay situaciones donde se pueden utilizar ambos; pero hay situaciones más complejas donde quizá ninguno de los dos sea posible de utilizar, ya sea porque no se puede asumir equiprobabilidad en los resultados o porque el fenómeno no puede ser observable o repetible muchas veces en condiciones idénticas.

Además, las personas con frecuencia estimamos probabilidades de manera subjetiva, con base en nuestra experiencia con un determinado fenómeno. Por ejemplo, voy de mi casa hacia mi oficina y tengo tres opciones para realizar el recorrido, de manera subjetiva evaluamos probabilidades de tráfico, para cada una de ellas, en función del conocimiento o experiencia que tenemos sobre el fenómeno. Personas con mucho conocimiento y experiencia en un campo puede hacer estimaciones subjetivas de probabilidad muy cercanas a la probabilidad verdadera de un evento. Este enfoque se conoce como *probabilidad subjetiva*.

En resumen, hemos mostrado ejemplos de tres formas distintas de asignar probabilidades a un evento aleatorio:

a) Desde las frecuencias relativas de los datos

b) Desde un modelo que asume equiprobabilidad en los eventos.

c) De acuerdo con el grado de creencia o conocimiento de una persona.

7.3 La escala de la probabilidad

El resultado de un evento aleatorio puede ocurrir pocas o muchas veces en una serie de observaciones. Los casos extremos son que el resultado no sea posible, decimos entonces que su probabilidad es cero (evento imposible); el otro extremo es que el resultado siempre ocurra, decimos entonces que su probabilidad es 1 (evento seguro).

Escala de la probabilidad y posibilidades de ocurrencia de un evento

Entre ambos valores se encuentra la probabilidad de un evento aleatorio; entre más próxima a 1 es la probabilidad se espera que ocurra con más frecuencia, por otro lado, entre más próxima a 0 sea la probabilidad se espera que ocurra con menor frecuencia. En suma, la probabilidad de un evento es un valor entre 0 y 1.

7.4 Enfoques de la probabilidad

La probabilidad puede verse a través de diferentes enfoques, lo cual también se conoce como diferentes definiciones de la probabilidad:

• Enfoque de modelos o enfoque clásico

Este enfoque está basado en la definición que estableció Laplace en su libro *Théorie analytique des probabilités en 1812: "La probabilidad de un evento es la proporción de casos favorables al número de casos posibles, siempre que los resultados sean equiprobables".*

A este enfoque también se le conoce como enfoque clásico, y está basado en el principio de simetría o equiprobabilidad en los resultados. Las probabilidades que el modelo proporciona serán aproximadas a las del experimento real en la medida que el supuesto de simetría se cumpla, si los supuestos de simetría son erróneos los resultados del modelo serán también erróneos. Consideremos el caso de una moneda deformada que tiene dos resultados que evidentemente no son equiprobables, si asumimos simetría en los resultados el modelo no proporciona resultados correctos. Otro ejemplo es el juego de cartas, que requieren no estar marcadas y bien barajadas para garantizar el principio de equiprobabilidad, de lo contrario el modelo no proporcionará resultados correctos.

En resumen, el enfoque de modelos proporciona una aproximación a la probabilidad real de un evento aleatorio en la medida que se cumpla el supuesto de equiprobabilidad en los resultados. Este enfoque tiene la característica que no requiere la realización u observación de los experimentos o fenómenos, por ello es un enfoque a priori. Es decir, no necesitamos lanzar un dado muchas veces para saber que la probabilidad de la cara con el número 5 es 1/6.

• Enfoque frecuencial o empírico

Este enfoque surge cuando se quiere aplicar la probabilidad a situaciones del mundo físico o natural donde nos es posible aplicar el *principio de equiprobabilidad*. La noción intuitiva de probabilidad bajo este enfoque se basa en la suposición de que si en n observaciones de un experimento aleatorio, el evento A ocurre f veces, y si el valor de

n es muy grande, entonces la frecuencia relativa *f/n* se aproximará a la probabilidad p del evento A. La formalización de lo anterior está dada por la **ley de los grandes números,** la cual señala: *"Si f es el número de éxitos en n ensayos con probabilidad p de éxito en cada ensayo, entonces f/n es el número promedio de éxitos y su valor deber ser próximo a p".*

Es decir, cuando n crece, la probabilidad de que el número promedio de éxitos se desvíe de la probabilidad p en más de cualquier cantidad asignada previamente tiende a cero. Este enfoque de la probabilidad es muy útil cuando se dispone de información sobre la ocurrencia de un gran número de casos de un evento; presenta el inconveniente de decidir cuántos experimentos se necesitan para considerar que la frecuencia relativa se ha vuelto estable.

La aplicación de este enfoque en el cálculo de probabilidades requiere:

1. Que el experimento se repita en condiciones idénticas.
2. Que las repeticiones sean independientes una de la otra.
3. Que el número de repeticiones sea lo suficientemente grande como para lograr la estabilización de las frecuencias.

• Enfoque subjetivo de la probabilidad

En este enfoque, la probabilidad es una expresión del grado de creencia o percepción personal. Este punto de vista mantiene que la probabilidad mide la confianza que un individuo particular tiene sobre la verdad de una proposición en particular, por lo tanto, suele variar de una persona a otra.

Actividad de aprendizaje

Una pareja de recién casados planea tener tres hijos. Hágase el supuesto (bastante razonable, por cierto) de que es igualmente probable que un nacimiento sea Hombre (H) o Mujer (M). El espacio muestral está definido por los siguientes eventos: MMM, MMH, MHM, HMM, HMH, HHM, HHH, MHH.

- ¿Qué probabilidades tiene la pareja de que sus tres hijos sean hombres?
- ¿Qué probabilidades tiene la pareja de que sus tres hijos sean mujeres?
- ¿Qué probabilidades tiene la pareja de que en sus tres hijos haya hombres y mujeres?

Las preguntas anteriores pueden ser resueltas desde el enfoque basado en modelos y desde el enfoque basado en datos.

Enfoque basado en modelos (enfoque clásico)

En este enfoque la probabilidad de un evento se obtiene dividiendo casos favorables entre casos posibles, siempre que sean equiprobables. En el primer caso se tiene solo un resultado posible (HHH) entre ocho posibles, de tal forma la probabilidad de P(HHH)=1/8; de la misma manera P(MMM)=1/8 que es igual a 0.125. En el tercer caso se tienen

como eventos favorables MMH, MHM, HMM, HMH, HHM, MHH, por lo que la probabilidad que la pareja tenga entre sus hijos a hombres y mujeres es 6/8 que es igual a 0.75, lo cual es muy probable que ocurra.

Enfoque basado en datos (enfoque frecuencial)
En este enfoque la probabilidad de un evento se obtiene observando muchas familias de tres hijos y tomando sobre el género de los tres hijos. Supóngase que después de visitar 2000 familias de tres hijos se tiene que 260 tienen tres hijos hombre (HHH) entonces la frecuencia relativa es 260/2000=0.130, este valor debe ser cercano –tal vez igual- a la probabilidad del resultado.

En la práctica esta relación entre enfoque de modelos y enfoque de datos es muy utilizada en aplicaciones de la probabilidad. Se empieza observando un fenómeno y se obtienen datos que sugieren un determinado comportamiento, el cual se puede expresar mediante un modelo. Entre más datos se tienen, el modelo puede representar mejor el fenómeno en cuestión; decimos entonces que los datos nos ayudan a construir un modelo. En otras situaciones se puede utilizar el modelo para generar datos, es entonces que el modelo nos ayuda a generar datos, sin tener que recolectarlos. Por ejemplo, supongamos que después de analizar miles de familias de tres hijos, los resultados del enfoque de modelos difieren mucho del resultado que proporcionan los datos, podría sospecharse que el supuesto de equiprobabilidad de nacimiento de H y M no se cumple.

7.5 Propiedades básicas de la probabilidad

Consideremos que A y B son dos eventos que ocurren en un espacio muestral de probabilidad. Puede ser que nuestro interés sea calcular la probabilidad simple de uno u otro evento, pero también nos puede interesar la probabilidad de que A y B ocurran ligados de alguna forma; por ejemplo, que ocurra A o que ocurra B, o que ocurra A y B de manera conjunta. Ambos casos conducen a dos importantes propiedades o reglas básicas en el cálculo de probabilidades. En forma más precisa:
Si Ω es el espacio muestral y A y B dos de sus eventos, entonces:

1. $0 \leq P(A) \leq 1$, *donde* $P(A)$ se lee "la probabilidad del evento *A*".
2. $P(\varnothing) = 0$ y $P(\Omega) = 1$, donde Ø es el evento imposible y Ω el evento seguro.
3. Si A y B son eventos tales que $A \cap B = \varnothing$, entonces $P(A \cup B) = P(A) + P(B)$.

Las tres condiciones anteriores (axiomas) indican las propiedades básicas de las probabilidades, no obstante, no ofrecen indicaciones de cómo calcularlas. Para esto, hay al menos dos formas para determinar las probabilidades de eventos, se llaman *enfoque clásico* y *enfoque frecuencial* de probabilidad.

7.6 Variables aleatorias y distribuciones de probabilidad

En las secciones anteriores abordamos la idea de *probabilidad de eventos*; ahora abordaremos otra perspectiva, la de *probabilidad de variables aleatorias*. En las aplicaciones más frecuentes de la estadística, los datos son tomados de muestras aleatorias de una población o de experimentos aleatorizados; en estos casos las variables en cuestión (por ejemplo, medias y proporciones muestrales) varían de una muestra a otra, por lo que se denominan *variables aleatorias*. Estas variables pueden ser descritas por las mismas medidas de una variable estadística (por ejemplo: media y desviación estándar).

Las variables aleatorias se clasifican en discretas o continuas. Las *variables aleatorias* discretas solo toman valores enteros y proceden de procesos de conteo, mientras que las variables aleatorias continuas toman valores fraccionarios en un intervalo dado y están asociadas a procesos de medición. Las variables aleatorias se suelen representar con las últimas letras mayúsculas del alfabeto (por ejemplo, X,Y,Z) y los valores numéricos que toman con las respectivas letras minúsculas (x,y,z).

Las distribuciones de probabilidad se pueden expresar a través de tablas, gráficas y fórmulas matemáticas. Además, existe el recurso computacional que ayuda en el cálculo de probabilidades. *La distribución de probabilidad queda determinada cuando se conocen todos los valores que puede tomar una variable aleatoria y su correspondiente probabilidad*.

En cada categoría existen distribuciones específicas con nombre propio, en la categoría de distribuciones discretas consideraremos la *distribución binomial*, y en el grupo de las distribuciones continuas a la *distribución normal*. En los siguientes apartados nos adentraremos en el estudio de sus propiedades, métodos de cálculo y solución de problemas.

7.7 La transición de eventos aleatorios a variables aleatorias

Considera de nuevo el caso de la pareja que planea tener tres hijos. El espacio muestral está definido por los siguientes eventos: MMM, MMH, MHM, HMM, HMH, HHM, HHH, MHH. Ahora, dejemos de lado el razonamiento con eventos y pensemos en función de variables aleatorias, es decir transformemos los eventos en valores numéricos.

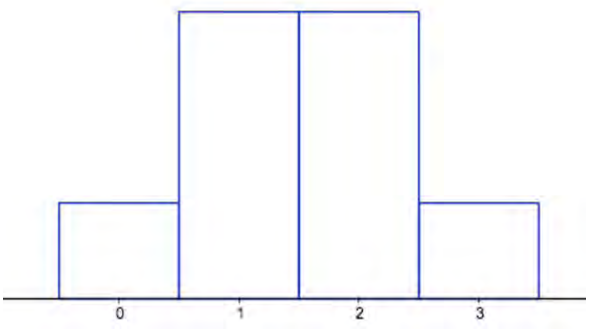
Defínase como X la variable aleatoria *número de hijos hombre (H) que el matrimonio puede tener en los tres hijos.* Fácilmente puede verse la posibilidad que X tome los valores 0, 1, 2 y 3.

Valor de la variable	Evento	Probabilidad
0	MMM	1/8=0.125
1	MHM, MMH, HMM	3/8=0.375
2	HHM, MHH, HMH	3/8=0.375
3	HHH	1/8=0.125

La tabla anterior muestra la relación que existe entre los eventos de un espacio muestral y los valores numéricos que puede tomar una variable aleatoria. En muchas situaciones (por ejemplo, en inferencia estadística) es más apropiado utilizar el enfoque de variables aleatorias que el enfoque de eventos. Con los resultados de la tabla se pueden responder preguntas como:

- ¿Qué valores numéricos puede tomar la variable aleatoria?
- ¿Qué probabilidad tiene cada uno de los valores de la variable aleatoria?
- ¿Qué valor de la variable tiene mayor o menor probabilidad de ocurrir?
- ¿A cuánto equivale la suma de probabilidades de todos los valores de una variable aleatoria?

Una representación gráfica de la distribución de probabilidad de X puede ser obtenida con el software *Geogebra*, y a través de ella podemos tener una mejor visualización del comportamiento de la variable aleatoria.



k	P(X = k)
0	0.125
1	0.375
2	0.375
3	0.125

Distribución de probabilidad de la variable X=numero de hijos hombre en tres nacimientos

La distribución de probabilidad muestra que es muy probable (0.875) que el matrimonio tenga al menos un hijo hombre (valores 1, 2 y 3). La probabilidad de tener tres hijos de un solo género (H o M) es 0.250.

7.8 Distribución de probabilidad binomial

La distribución binomial tiene mucha aplicación en muestreo de poblaciones en las que los elementos tienen variables que pueden tomar valores binarios o dicotómicos. Veamos los siguientes casos:

- Se selecciona una muestra de personas para que respondan una encuesta con una serie de preguntas que tienen dos opciones de respuesta (de acuerdo, en desacuerdo).
- Una prueba de calidad en una industria consiste de una muestra de componentes que tienen dos características (defectuoso, no defectuoso).
- Una revisión a una muestra de pacientes en un hospital sobre el brote de una epidemia tiene dos posibles resultados en cada paciente (sano, enfermo)

En cada una de las situaciones anteriores se ha seleccionado una muestra aleatoria de n objetos de una población finita de tamaño N. En cada objeto muestreado observamos si la variable presenta la característica de nuestro interés (por ejemplo: el paciente está sano, el componente está defectuoso). Designamos mediante x el número de objetos muestreados que tienen tal característica.

Aun cuando las tres situaciones anteriores proceden de contextos distintos, son esencialmente idénticas. En la búsqueda de un modelo general para representar situaciones que tengan la misma estructura que las mencionadas anteriormente, recurriremos a una urna con bolas de dos colores (ver figura).

La población es análoga a las N bolas en la urna. Supongamos que la variable de interés es el color de las bolas, y el valor que interesa es el color rojo (le llamaremos éxito). Seleccionamos n bolas de la urna y contamos las x bolas de color rojo que aparecen en la muestra. Puede resultar que ninguna bola sea roja o que todas las bolas sean rojas, en el primer caso x=0, y en el segundo x=n. En una nueva selección podríamos obtener un numero diferente de bolas rojas en la muestra, por lo cual se considera una variable (X). En muchas situaciones el tamaño de muestra n es muy pequeño respecto del tamaño de la población N; bajo estas condiciones la distribución de probabilidad de la variable X (número de bolas rojas en la muestra) es conocida como **distribución binomial.**

Otro modelo físico al que se puede recurrir para conceptualizar la idea central que subyace a la distribución binomial es el lanzamiento repetido de una moneda. En cada lanzamiento la moneda tiene dos resultados posibles, uno de ellos puede ser la característica de interés (por ejemplo: águila) al que llamaremos éxito. La variable (X) puede ser el número de águilas en una serie de n lanzamientos. Interesa conocer la cantidad (x) de **águilas** que suceden en las n repeticiones. Considerando que los resultados son independientes uno de otro y la probabilidad de obtener **águila** se mantiene constante de un lanzamiento a otro, la distribución de probabilidad de X es una **distribución binomial.**

En general podemos imaginar un proceso binomial como una secuencia de resultados binarios obtenidos de un proceso aleatorio, en la cual interesa la frecuencia

de un resultado en particular. Tanto el modelo de urna como el modelo de la moneda, nos pueden ser **útiles** para conceptualizar los elementos que intervienen en una distribución binomial.

• Condiciones para la distribución binomial

- El experimento aleatorio (por ejemplo: selección de una muestra de la población) consiste de n repeticiones idénticas.
- En cada repetición el experimento tiene dos resultados posibles: éxito (E) y fracaso (F).
- La probabilidad de éxito en cada repetición es igual a p y la probabilidad de fracaso es $1 - p$. Ambas se mantienen constantes de una repetición a otra.
- Las repeticiones del experimento son independientes.

La variable aleatoria de interés X, **es el número de éxitos observados en las *n* pruebas.** Si se cumplen las condiciones anteriores decimos que tiene distribución de probabilidad binomial, con parámetros n y p. Se acostumbra expresar simbólicamente como: $X \sim Binomial\ (n,p)$.

Ejemplo de un experimento binomial

El matrimonio que planea tener tres puede considerarse un experimento binomial, ya que cumple con todas las condiciones. Se tienen tres repeticiones idénticas (nacimiento de cada hijo), en cada nacimiento puede resultar hombre (H) o mujer (M), la probabilidad de éxito (consideremos a H como tal) es $p=0.5$ y su complemento $1-p=0.5$, se mantienen constantes de un nacimiento a otro. Se puede considerar que los nacimientos son independientes uno de otro. La variable X es el n*úmero de hijos hombre (H) obtenidos en los tres nacimientos. $X \sim Binomial\ (3,0.5)$*

• La fórmula de la distribución binomial

Señalamos anteriormente que un experimento binomial tiene resultados binarios mutuamente excluyentes, uno llamado éxito (E) y el otro llamado fracaso (F). Pensemos que el experimento se repite n veces y que se obtuvo el siguiente resultado.

E F F E E F E F . . . .F E E F

Por facilidad, acomodemos los resultados para que aparezcan juntos los éxitos y los fracasos. Consideremos que se presentaron x éxitos, por lo que se tendrán $n - x$ fracasos:

E E E E E E E . . . .E E F F F . . . .F F

x éxitos $\quad$ $n - x$ fracasos

La probabilidad de éxito es denotada p y la probabilidad de fracaso por $1 - p$, entonces la probabilidad de que suceda el resultado anterior se obtiene aplicando la regla del producto probabilidades, dado que los eventos se consideran independientes.

$$p\,p\,p\,\ldots p\,p\,\ldots(1-p)(1-p) = p^x(1-p)^{n-x}$$

La probabilidad anterior corresponde a una sola combinación de éxitos y fracasos, sin embargo se pueden presentar muchas combinaciones como la anterior. Entonces se debe multiplicar la probabilidad de un resultado por el total de combinaciones posibles. La expresión queda de la siguiente manera:

$$P(x) = \binom{n}{x} p^x(1-p)^{n-x} \quad x = 0, 1, 2, \ldots . n$$

$\binom{n}{x} = \frac{n!}{x!(n-x)!}$ representa las combinaciones posibles de éxitos en pruebas. A esta parte de la fórmula se le conoce como *coeficiente binomial*.

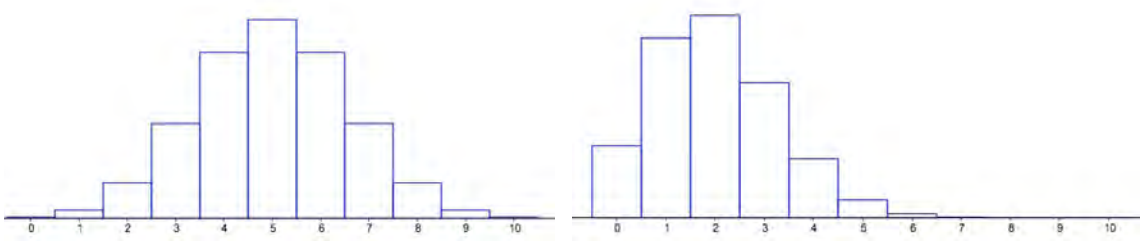
La fórmula de la distribución binomial depende de los parámetros n y p. Es decir, es del tamaño de muestra y de la probabilidad de tener éxito en cada repetición.

Para explicar los elementos que aparecen en la fórmula retomemos el caso de la familia que planea tener tres hijos. La probabilidad de tener un hijo, sea hombre o mujer es $\frac{1}{2}$ asi que la probabilidad de un evento en particular (por ejemplo HMH) es $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$, esto es la parte $p^x(1-p)^x$ de la fórmula.

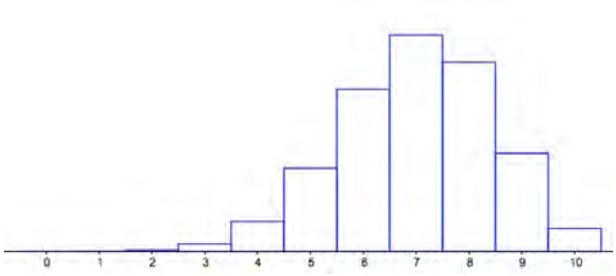
Supongamos que nos interesa la probabilidad de que sean dos hijos hombre (H), se tienen tres combinaciones de ello (HMH, HHM, MHH). Esta es la parte $\binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{3!}{2!(3-2)!} = 3.$ Entonces la probabilidad es igual a $3 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{3}{8}$

• Efecto de los parámetros en la distribución binomial

Los parámetros n y p determinan el comportamiento de la distribución, más precisamente, para cada par de valores n y p, existe una distribución de probabilidad. Utilicemos el software *Geogebra* para mostrar el efecto de variar el valor de p para un tamaño de muestra dado.



a) n=10, p=0.5 (forma simétrica) **b)** n=10, p=0.2 (forma sesgada a la derecha)



c) n=10, p=0.8 (forma sesgada a la izquierda)

Se observa lo siguiente:

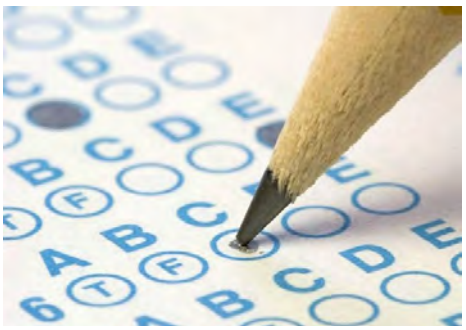
- Para valores de p=0.5, la distribución es simétrica.
- Para valores pequeños de p, la distribución es sesgada a la derecha.
- Para valores grandes de p, la distribución es sesgada a la izquierda.

Razonemos un poco más con las tres distribuciones anteriores. Si la probabilidad de éxito es pequeña, en una secuencia de repeticiones (caso b) es poco probable que se presenten muchos éxitos, por el contrario, si la probabilidad de éxito es grande (caso c), es más probable que se presenten muchos éxitos en una secuencia. Por ejemplo, obsérvese que para $x = 7$, en el caso b la probabilidad es casi nula, pero en el caso c, es el valor con mayor probabilidad de todos. En resumen, cuando el valor de una variable tiene poca probabilidad de ocurrir, es mucho menos probable que el valor suceda en forma repetida en una secuencia de resultados.

Actividad de aprendizaje

Un examen de opción múltiple consta de 6 preguntas. Cada pregunta proporciona tres respuestas de las cuales solo una es correcta. Supóngase que un estudiante decide responder completamente al azar. ¿Cuál es la probabilidad de que el estudiante apruebe el examen, si requiere contestar correctamente al menos 4 preguntas?

- ¿Se podría considerar que el examen cumple con las condiciones de un experimento binomial?
- El examen consta de n pruebas (cada pregunta se considera una prueba), cada pregunta tiene dos opciones (correcta o incorrecta), la probabilidad de éxito es mantiene constante (p=0.33, pues solo hay una opción correcta de tres), el alumno contesta al azar, por tanto, se pueden considerar independencia.
- ¿Cuál sería la variable aleatoria de interés?
- La variable de interés puede ser el número de respuestas correctas al contestar el examen, pero también puede ser el número de respuestas correctas. Optaremos por la primera opción.



Un estudiante obtuvo la probabilidad de obtener 6 respuestas correctas utilizando la fórmula de la

distribución binomial:

$$P(x) = \binom{n}{x} p^x (1-p)^x$$

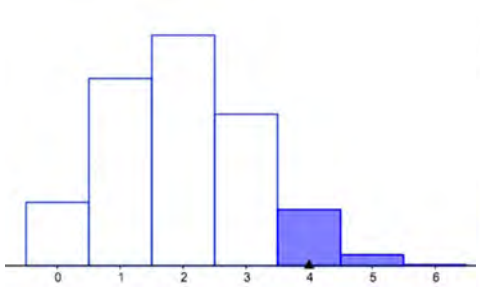
$$P(6) = \binom{6}{6} 0.33^6 (0.67)^0 = 0.0013$$

Utiliza la fórmula para completar la tabla con los valores y probabilidades que hacen falta.

Número de respuestas correctas *x*	Probabilidad *P(x)*
0	
1	
2	
3	
4	
5	
6	0.0013
Suma	

El cálculo de probabilidades con la fórmula, aún usando una calculadora resulta laborioso, y más aún cuando la variable toma muchos valores. Se recomienda utilizar la hoja de cálculo Excel o un software como *Geogebra*.

Utiliza el software *Geogebra* para verificar resultados que aparecen en la tabla. Se introducen los parámetros n=6, p=0.33 y resulta la siguiente distribución:



k	P(X = k)
0	0.0905
1	0.2673
2	0.3292
3	0.2162
4	0.0799
5	0.0157
6	0.0013

La tabla proporciona la probabilidad para cada valor de la variable, el área sombreada representa los valores (x=4, 5 y 6) que permiten que el alumno acredite el examen.

• Media y desviación estándar de la distribución binomial

Para el caso específico de la distribución binomial y cuando se conocen sus parámetros *n* y *p*, las fórmulas adoptan diferente presentación. Para el caso del valor esperado se tiene la siguiente expresión:

$$\mu = np$$

La desviación estándar se tiene calcula con la siguiente fórmula: $\sigma = \sqrt{np(1-p)}$

Para explicar los conceptos anteriores, considera de nuevo el examen de opción múltiple (n=6, p=0.33). El valor esperado de la variable número de respuestas correctas sería:

$$\mu = np = 6(0.33) = 2$$ respuestas correctas.

¿Qué significado tiene el valor esperado de una variable aleatoria?, es el valor que tiene mayor probabilidad de ocurrir. En este caso el valor esperado es 2 respuestas correctas, aunque pueden suceder los otros valores. Véase la distribución de probabilidad que la barra más alta corresponde a $x = 2$.

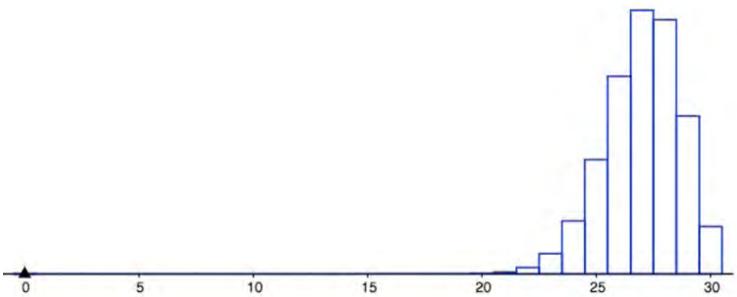
Por su parte, la desviación estándar es. $\sigma = \sqrt{np(1-p)} = \sqrt{6(0.33)(0.67)} = 1.15$

La desviación estándar establece la variabilidad que puede tener el valor esperado.

Actividad de aprendizaje

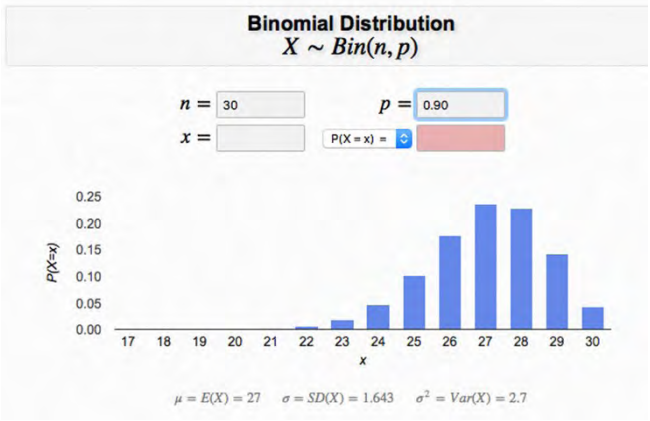
Según datos estadísticos de la Secretaría de Gobernación en México, en un balance realizado en el 2017 sobre el servicio de emergencias 911, se reporta que 9 de cada 10 llamadas son falsas emergencias, por lo que se ha decidido penalizar a las personas que incurran en esta práctica. Considérese variable aleatoria X el número de llamadas recibidas en un día en el servicio 911.

La variable X puede tomar dos valores: falsa o verdadera, se puede considerar que las llamadas son independientes, además consideraremos como éxito que la llamada sea falsa, por tanto, $p=0.90$. ¿Cuál sería la distribución de probabilidad de X si se toma una muestra aleatoria de 30 llamadas? Podemos recurrir al software *Geogebra* para construir la distribución binomial con parámetros *n* y *p*.



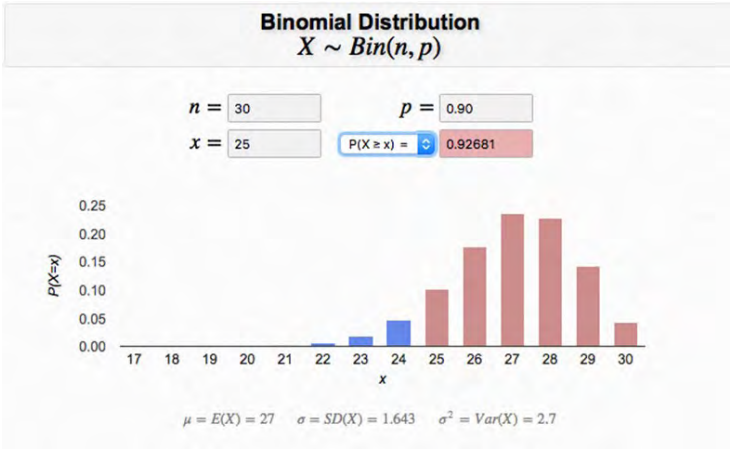
La distribución de probabilidad de X también puede ser construida utilizando el

siguiente applet: http://homepage.stat.uiowa.edu/~mbognar/applets/bin.html.



Hagamos un poco de razonamiento con la distribución de probabilidad. De entrada, se observa poco probable (no imposible) observar pocas llamadas falsas, pues la probabilidad de que entren este tipo de llamadas al 911 es muy alta. La probabilidad empieza a ser visible gráficamente a partir de x=22 llamadas, logrando su valor más alto en 27 llamadas, que es el valor esperado de la variable aleatoria (ver figura del applet).

Ahora supongamos que nos interesa conocer la probabilidad de que en la muestra de 30 llamadas al menos 25 sean falsas. Introduciendo los valores en el applet resulta lo siguiente:



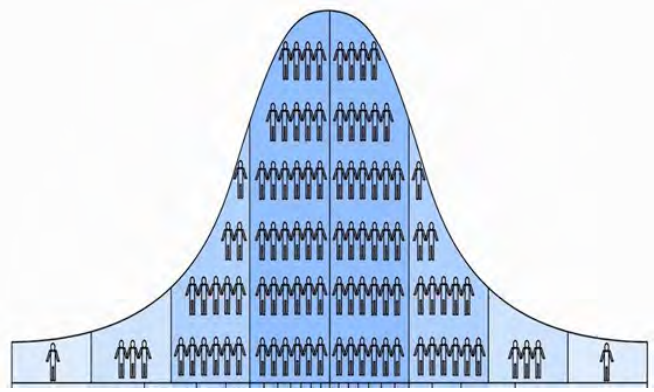
La probabilidad es de 0.9268. Obsérvese que el applet además de proporcionar la probabilidad para el valor o intervalo de valores de una variable permite visualizar la distribución con los valores de interés, y en la parte inferior proporciona el valor esperado $\mu=E(X)$ y la desviación estándar σ.

• Distribución de probabilidad normal

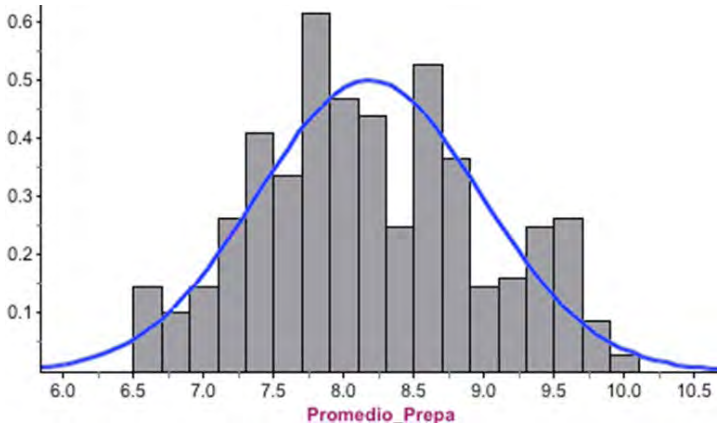
La distribución normal es una distribución para variables aleatorias continuas, también es conocida como distribución Gaussiana o campana de Gauss, en honor al matemático Karl Friedrich Gauss (1777-1855) quien en 1823 publicó su expresión matemática.

Muchos fenómenos de interés para el ser humano tienen un comportamiento como el que describe la distribución normal. Por ejemplo: diversos caracteres morfológicos de personas, animales y plantas, como la estatura, el peso y longitudes de algunas partes del cuerpo; caracteres psicológicos como el coeficiente intelectual;

los resultados de pruebas estandarizadas para evaluar conocimiento y habilidades de las personas como es el caso de Ceneval, entre otros más; además muchos métodos estadísticos se basan en la suposición de normalidad para generar estimaciones confiables.



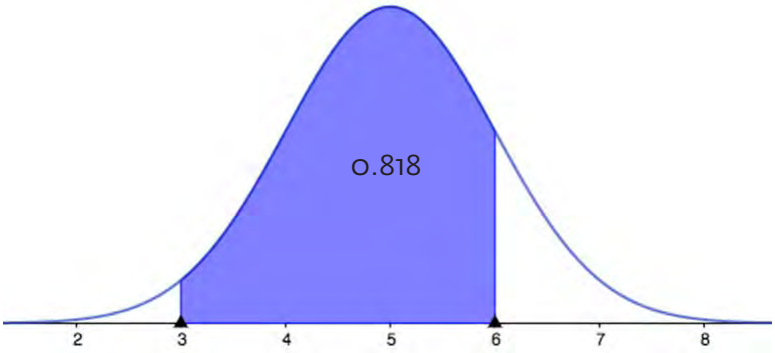
Para ejemplificar lo anterior, el siguiente histograma describe los promedios de 379 estudiantes al terminar su preparatoria. El histograma es aproximadamente simétrico y ambas colas descienden con suavidad a cierta distancia del centro, lo que quiere decir que la mayoría de los promedios se encuentran alrededor de 8, pocos estudiantes lograron altas y bajas calificaciones. La curva suave dibujada sobre el histograma es una buena descripción del aspecto general de los datos.



Promedios finales de 379 alumnos egresados de Preparatoria

Entre las ventajas de ajustar una distribución de datos a una curva como la anterior, está convertir los datos en probabilidades y hacer predicciones sobre las calificaciones de los estudiantes que concluyen la preparatoria. A la curva suavizada se le conoce como curva densidad de probabilidad, el área bajo la curva y entre cualquier intervalo de valores representa la proporción de datos situados en dicho intervalo.

La siguiente figura muestra la curva de densidad de una distribución simétrica, el área sombreada bajo la curva corresponde al intervalo que va de 3 a 6. Esta área es igual a 0.818. Esto significa que el 81.8% de las observaciones o datos está comprendidos entre dicho intervalo.



Una curva de densidad es una distribución idealizada de un conjunto de datos. Hemos visto en capítulos anteriores que una distribución de datos se puede describir por su media y su desviación estándar, estas medidas son los parámetros que definen el comportamiento de la distribución normal.

• Fórmula de la distribución de probabilidad normal

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

μ: es la media de la variable.
σ: es la desviación estándar de la variable.
x: representa un valor particular de la variable.

La distribución normal depende de los parámetros μ y σ, que representan la media y la desviación estándar respectivamente.

Cuando una variable aleatoria tiene distribución normal, se acostumbra simbolizarla así: $X \sim Normal(\mu,\sigma)$ o simplemente $X \sim N(\mu,\sigma)$**.** Por ejemplo, una variable aleatoria que tiene distribución normal con media $\mu=100$ y $\sigma=5$, se puede representar como $X \sim N(100,5)$. La gráfica correspondiente se muestra a continuación:

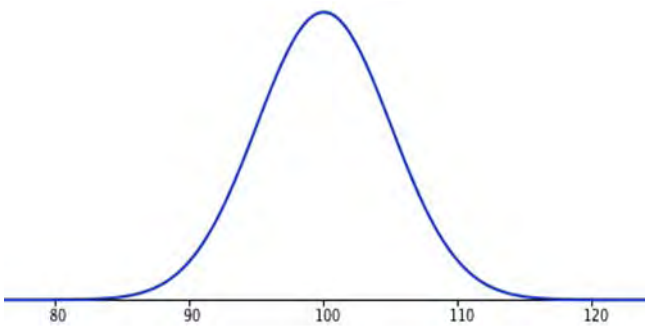


Figura: Distribución normal con $\mu = 100$ y $\sigma = 5$

Obsérvese que la media se encuentra en el centro de la distribución, y que el área de la distribución se extiende aproximadamente desde 85 hasta 115, es decir 3 deviaciones estándar antes y después de la media respectivamente. En lecciones posteriores verás que esta es una importante propiedad de la distribución normal.

Como un ejemplo notable por su importancia histórica consideremos los datos

analizados por Quetelet (1796-1874), científico belga con fuertes inclinaciones hacia la estadística, y que, interesado en la teoría del *hombre medio*, fue el primero en aplicar la distribución normal a datos humanos. Analizando datos sobre las medidas del tórax de más de 5000 soldados escoceses, observó que había mucha variabilidad en sus longitudes y concluyó que las medidas tenían una distribución muy aproximada a la normal con media del tórax de 39.8 pulgadas y desviación estándar de 2.05 pulgadas, esto es $X \sim N(39.8, 2.05)$.

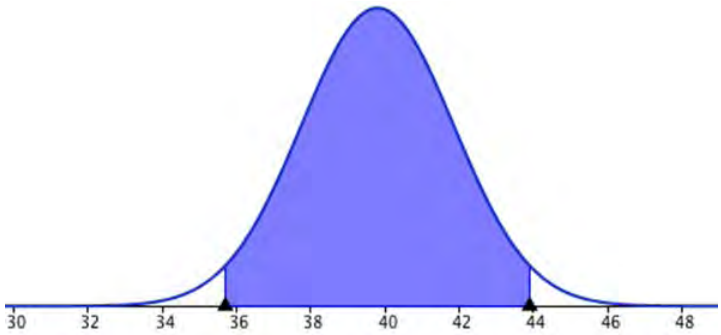
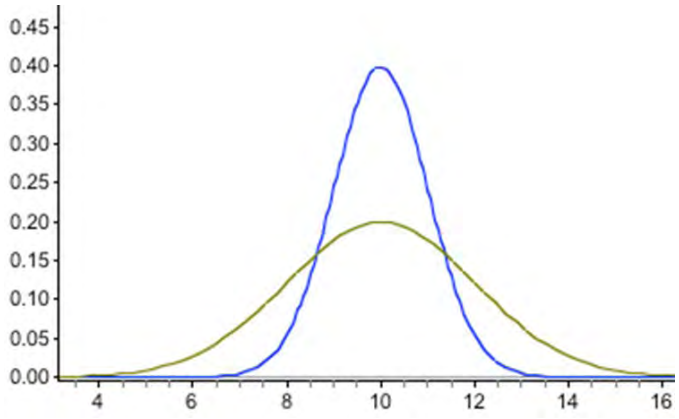


Figura: Distribución de las medidas de tórax de 5000 soldados escoceses

• Efecto de los parámetros en la distribución normal

La media μ define el centro y la desviación estándar σ la variabilidad o dispersión de la distribución normal, de tal forma, para cada par de valores (μ,σ) se tiene una distribución normal diferentes. En la siguiente figura se tienen las distribuciones N (10,1) y N (10,2). Obsérvese que ambas tienen la misma media, pero diferente desviación estándar. A mayor desviación la distribución se amplía más, pero disminuye su altura, conservando siempre su área.



Propiedades de la distribución de probabilidad normal

Los parámetros μ y σ determinan el comportamiento de la distribución normal, es decir: cuando varían los parámetros también varía la distribución. En este sentido, la expresión matemática de la distribucion normal representa una familia infinita de distribuciones, ante la infinidad de valores que pueden tomar los parámetros.

- La distribución normal es simétrica, por lo cual la mitad de las observaciones o datos están por debajo de la media y la otra mitad se encuentran por encima de la media.
- La media, la mediana y la moda de los datos de la distribución coinciden.
- La distribución se extiende en forma asintótica sobre el eje horizontal.

- Para cualquier distribución se pueden conocer las proporciones de datos o probabilidades, en función del número de desviaciones estándar, que se encuentran representadas en el eje horizontal. En general se cumple lo siguiente:

$$\mu \pm 1\sigma = 68\%, \quad \mu \pm 2\sigma = 95\%, \quad \mu \pm 3\sigma = 99.7\%.$$

A la última propiedad de la distribución normal también se le conoce como *regla empírica* o *regla* 68 - 95 - 99.7. Esto significa que entre 1 desviación estándar alrededor de la media se encuentran el 68% de los datos, a 2 desviaciones estándar se encuentran 95% de los datos, y entre tres desviaciones estándar se encuentran 99.7% de los datos.

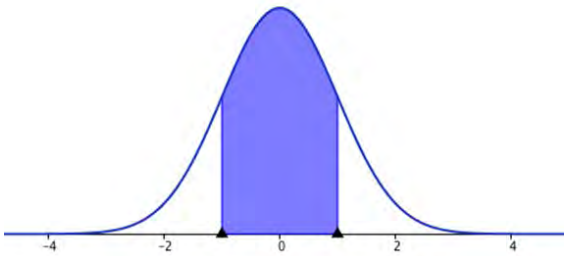


Figura: Distribución normal con área comprendida entre $\mu \pm 1\,\sigma$

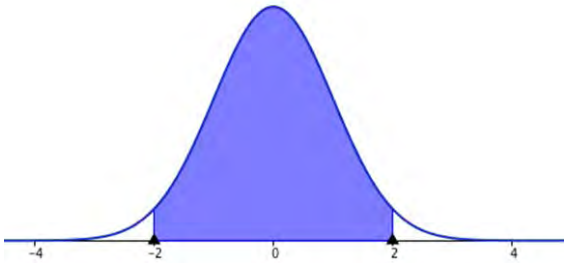


Figura: Distribución normal con área comprendida entre

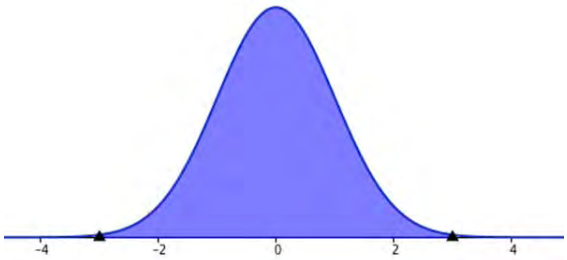


Figura: Distribución normal con área comprendida entre

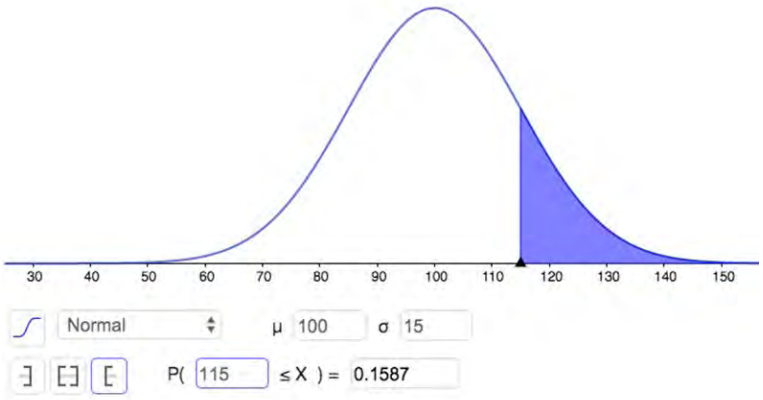
Actividad de aprendizaje

Uno de los métodos más conocidos para medir el coeficiente de inteligencia (CI) de las personas es la prueba Stanford-Binet. Los puntajes de la prueba tienen una distribución normal con media $\mu = 100$ desviación estándar $\sigma = 115$.

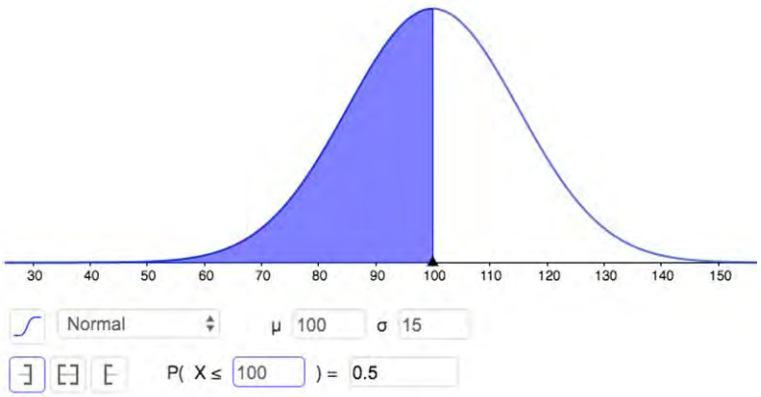
a) ¿Cuál será la probabilidad de que obtenga un CI mayor a 115?

b) ¿Cuál será la probabilidad de que obtenga un CI menor a 100?

Ingresamos a *Geogebra* para generar la distribución con los parámetros 100 y 15.



$P\,(X > 115) = 0.1587$

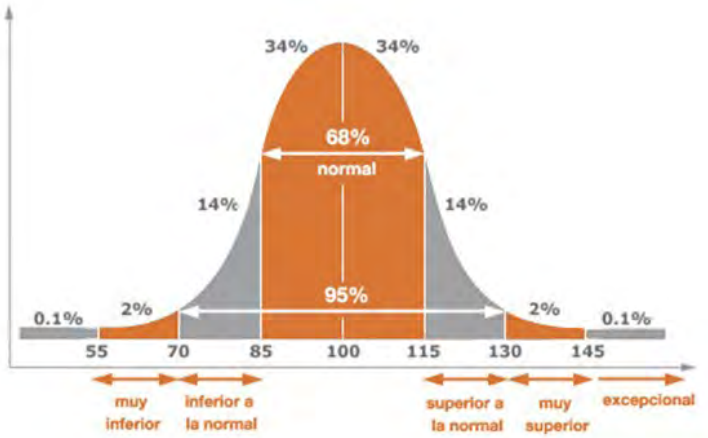


$P\,(X < 100) = 0.500$

Si en lugar de probabilidades, se utilizan proporciones, en el primer caso se puede decir que el 15.87% de la población tiene una CI mayor de 115; y que el 50% de la población tiene un 100 mayor o menor a 100.

En la siguiente figura se ha graficado la misma distribución para el CI. Utilizando las propiedades de la distribución normal se puede observar lo siguiente:

- El 68% de las personas tienen un CI entre 85 y 115.
- El 95% de las personas tiene un CI entre 70 y 130.
- El 14% de las personas tiene un CI superior a lo normal (entre 115 y 130).
- El 2.1% de las personas tiene un Ci muy superior o excepcional (mayor a 130).
- El 2% de las personas tiene un CI inferior a lo normal (entre 55 y 70).



Ejemplo tomado de https://bookdown.org/aquintela/EBE/ejemplos-de-la-distribucion-normal.html

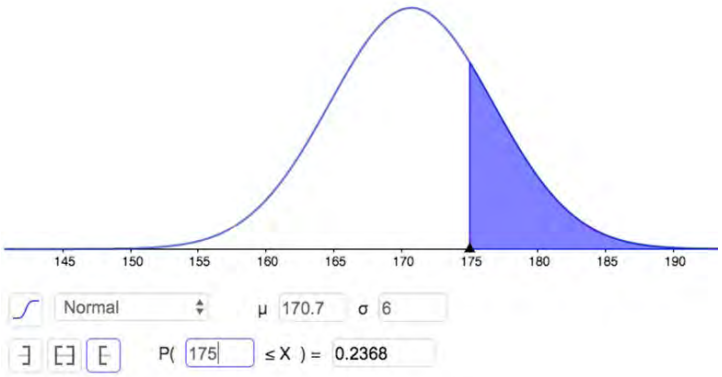
Actividad de aprendizaje

Un estudio con jóvenes estudiantes de 18 años realizado por la Universidad de Guadalajara reporta que las siguientes medidas de estatura y peso por hombres y mujeres.

	Estatura		Peso	
	Media	DE	Media	DE
Hombres	170.7	6.0	67.2	11.6
Mujeres	157.4	5.7	53.6	6.7

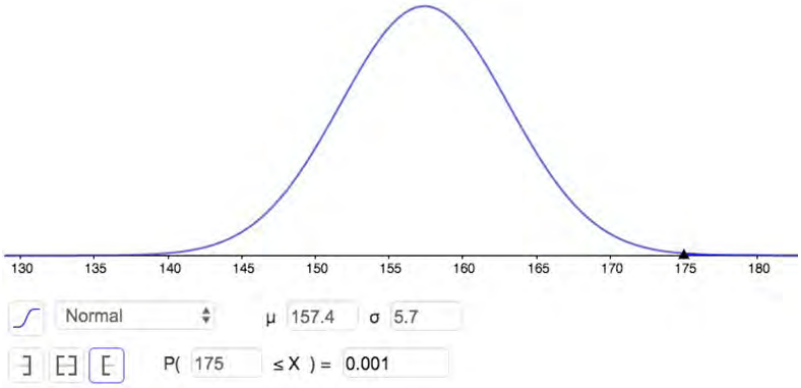
La estatura y el peso tienen una distribución aproximadamente normal. Considera que la muestra es representativa de los jóvenes de 18 años mexicanos. Responde las siguientes preguntas:

- ¿Qué proporción de jóvenes hombres de 18 años tendría una estatura mayor de 175 cm?



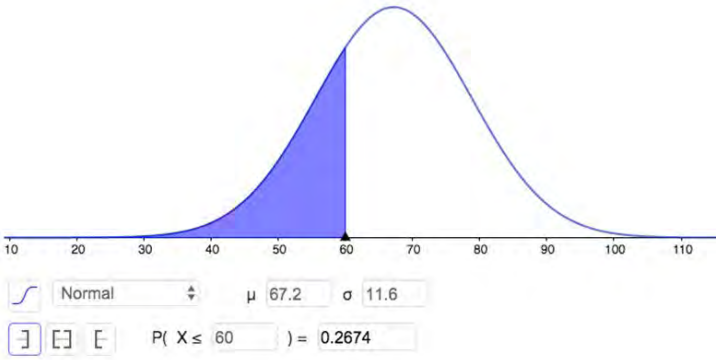
Normal (170.6, 6.0)

- ¿Qué proporción de jóvenes mujeres de 18 años tendría una estatura mayor de 175 cm?



Normal (157.4, 5.7)

- ¿Qué proporción de jóvenes hombres de 18 años tendría un peso menor de 60 kg?



Normal (67.2, 11.6)

- Verificar los resultados anteriores con el applet que aparece en la siguiente liga: https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html

Para tu reflexión

La sociedad actual requiere individuos que sean competentes para analizar, comprender e interpretar información de diversos hechos que suceden a su alrededor, los cuales con frecuencia involucran algún componente de incertidumbre (por ejemplo, eventos meteorológicos, eventos deportivos, juegos de azar, pronósticos financieros, seguros de vida o daños). Como consecuencia de ello, los temas de probabilidad han venido incrementando su presencia en el currículo escolar de matemáticas de muchos países, desde la escuela primaria hasta el nivel universitario.

Sin embargo, la enseñanza de la probabilidad es una tarea compleja, pues además de ser relativamente reciente en el currículo, tiene la característica de ser conceptualizada desde diversos enfoques (clásico, frecuencial, subjetivo). Adicionalmente, la investigación ha mostrado la facilidad con que las personas incurren en concepciones erróneas y sesgos cuando razonan sobre la incertidumbre y el azar.

Aunado a lo anterior, la enseñanza de la probabilidad ha estado centrada

principalmente en el enfoque clásico o axiomático, con uso a veces excesivo de técnicas combinatorias y conceptos formales que poco ayudan al desarrollo de un razonamiento probabilístico correcto, en tanto enfatizan casi de forma exclusiva en el uso de procedimientos y cálculo de probabilidades. Ante la problemática anterior, en este capítulo hemos mostrado algunos ejemplos sobre cómo hacer un mayor equilibrio entre el enfoque clásico, frecuencial de la probabilidad, para mejorar su comprensión.

Nota histórica

De acuerdo con algunos historiadores, la probabilidad surgió como una rama de las matemáticas con una correspondencia en la que intercambiaron discusiones sobre el tema Blaise Pascal y Pierre Fermat en 1654. Sin embargo, es importante precisar que desde siglos antes a Pascal y Fermat, se había pensado acerca de la probabilidad y varios problemas fueron tratados por abogados y matemáticos.

La probabilidad surge de dos raíces: los juegos de azar y el manejo de datos estadísticos relacionados con tablas de mortalidad. Desde luego, la raíz lúdica de la probabilidad ha tenido gran impacto en su enseñanza hasta nuestros días, cuando todavía observamos que muchos libros de texto y currículos hacen énfasis en los juegos de azar para explicar los principios de la probabilidad, particularmente en la probabilidad clásica basada en el modelo de equiprobabilidad que estableció Pierre Simón de Laplace.

Entre sus iniciadores se encuentra Girolamo Cardano, hombre apasionado de los juegos de azar que en 1539 publicó un trabajo para jugadores de azar titulado Liber de Ludo Aleae (El Estudio del Juego). Mención especial en el origen de la probabilidad tiene el problema de los puntos que consiste repartir las apuestas en un juego de azar en el que participan dos jugadores y que ha sido interrumpido antes de concluir. La solución del problema ocupó a muchos estudiosos de la época como Luca Facioli, Cardano y Tartaglia que por décadas trataron de solucionarlo de manera infructuosa.

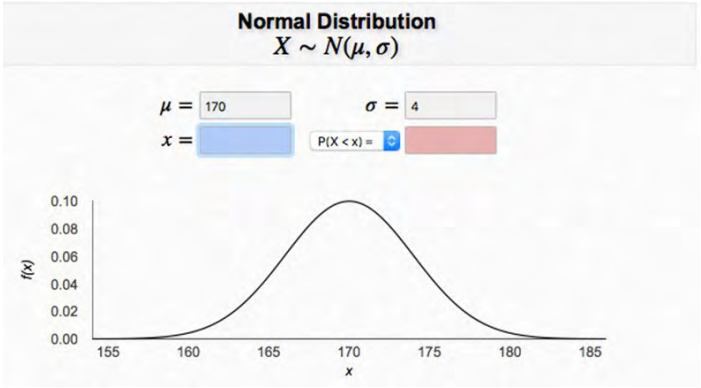
En 1654, Antoine Gombaud conocido como el Caballero de Meré, se encontró con el problema y lo planteó al matemático Blaise Pascal quien a su vez lo aborda junto con Pierre de Fermat a través de diversas cartas, y fueron quienes lo resolvieron de manera correcta. La solución al problema de los puntos es uno de los hechos que marca el inicio de la teoría de la probabilidad.

Evaluación del capítulo

1. Las estaturas de jóvenes mexicanos tienen una distribución aproximadamente normal con $\mu = 170\ cm$. y $\sigma = 4\ cm$. Utiliza el applet para responder las siguientes preguntas:

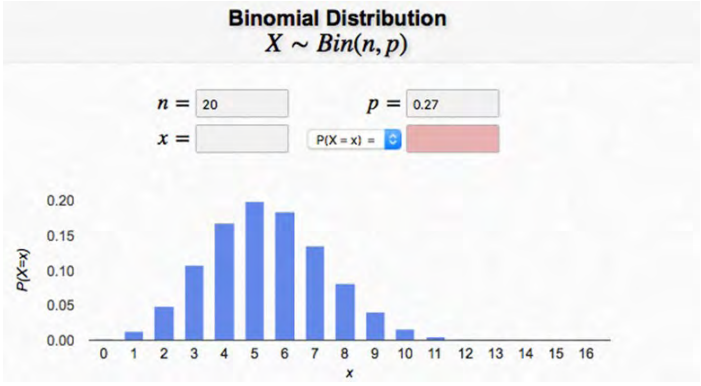
a) ¿Qué proporción de jóvenes tiene una estatura menor a 165 cm?

b) ¿Qué proporción de jóvenes tiene una estatura mayor a 180 cm?



https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html

2. Según el Censo de Población y Vivienda realizado por INEGI en 2010, el 27% de los mexicanos que emigran a otros países tienen una edad entre 25 y 34 años. Supongamos que se tiene una base de datos con todos los mexicanos que han emigrado y se selecciona una muestra aleatoria de 20 personas. Utiliza el applet para responder las siguientes preguntas:



https://homepage.divms.uiowa.edu/~mbognar/applets/bin.html

a) ¿Qué probabilidad existe de que en la muestra de 20 personas que han emigrado, todas sean de este grupo de edad?

b) ¿Qué probabilidad existe de que en la muestra de 20 personas que han emigrado, ninguna sea de este grupo de edad?

c) ¿Qué probabilidad existe de que en la muestra de 20 personas que han emigrado, al menos 10 sean de este grupo de edad?

d) ¿Cuál es el valor esperado de emigrantes de este grupo de edad en la muestra seleccionada?

3. Con base en los datos recopilados de los pacientes de Covid-19 y sus enfermedades previas, el IMSS construyó una calculadora de riesgo de complicación en caso de que un paciente adquiera la enfermedad. El riesgo se mide como una probabilidad, misma que aumenta conforme mayores factores de riesgo tiene la persona.

http://www.imss.gob.mx/covid-19/calculadora-complicaciones

4. ¿Qué enfoque de la probabilidad se está utilizando para conocer el nivel de riesgo de complicación?

a) Enfoque clásico
b) Enfoque subjetivo
c) Enfoque frecuencial

5. Un profesor está investigando cuánto tiempo tardan en resolver sus alumnos un problema de estadística para una prueba de admisión en la que el tiempo límite es de 1 hora. Después de aplicarla a una gran cantidad de alumnos obtuvo los siguientes resultados. Determina el valor esperado (promedio) de tiempo que tardarían los estudiantes en responder el problema.

X (tiempo en minutos)	5	6	7	8	9
P(X)	0.10	0.20	0.25	0.15	0.30

Bibliografía recomendada

- Conexiones entre Probabilidad Teórica y Probabilidad Frecuencial en un Ambiente de Modelación Computacional
  https://www.aiem.es/index.php/aiem/article/view/172
- Geogebra: una herramienta cognitiva para la enseñanza de la probabilidad
  https://docplayer.es/38490613-Geogebra-una-herramienta-cognitiva-para-la-ensenanza-de-la-probabilidad.html
- Razonamiento probabilístico en la vida cotidiana: un desafío educativo
  https://www.ugr.es/~batanero/pages/ARTICULOS/ConferenciaThales2006.pdf
- Comprensión que muestran profesores de secundaria acerca de los conceptos de probabilidad
  https://www.redalyc.org/pdf/405/40521127003.pdf
- Los inicios de la probabilidad
  https://revistasuma.es/IMG/pdf/55/007-020.pdf
- Significados de la probabilidad
  file:///Users/Administracion/Downloads/Dialnet-SignificadosDeLaProbabilidadEnLaEducacionSecundari-2096616.pdf
- Potencialidades y dificultades de la modelización de fenómenos aleatorios mediante simulación computacional en un curso universitario
  https://repensarlasmatematicas.files.wordpress.com/2012/08/inzunza2012-eipe.pdf
- La simulación como herramienta pedagógica en probabilidad: potencialidades y dificultades.
  https://www.youtube.com/watch?v=By2JisLRTus

Capitulo 8

Introducción a la Inferencia Estadística

Razonar desde una muestra de datos para hacer inferencias sobre una población es una difícil noción para la mayoría de los estudiantes

Richard Sheaffer

8.1 Introducción

La inferencia estadística constituye una de las áreas de mayor aplicación de la estadística, ya que mediante sus métodos se pueden obtener conclusiones significativas sobre una población, con base en la información que proporcionan los datos de una muestra aleatoria o un experimento aleatorizado. Los principales métodos de inferencia estadística son la *estimación de parámetros* y las *pruebas de hipótesis*, también llamadas *pruebas de significación*. La estimación de parámetros puede ser *puntual* o mediante *intervalos de confianza*, en ambos casos se tiene como propósito estimar el valor de un parámetro desconocido de una población (por ejemplo, una proporción o una media).

Por su parte, una prueba de hipótesis es un método que permite verificar una aseveración acerca del valor de un parámetro poblacional. Dado que los datos son proporcionados por una muestra, los resultados pueden estar sujetos a variaciones aleatorias. Una prueba de hipótesis permite decidir si pequeñas desviaciones observadas respecto al resultado que idealmente debería haber ocurrido, según nuestra hipótesis, son atribuibles al azar o efectivamente los resultados no se corresponden con la hipótesis que se ha planteado sobre el valor del parámetro.

En septiembre de 2019 Consulta Mitofsky realizó una encuesta para conocer la opinión de los mexicanos sobre el presidente Andrés Manuel López Obrador, obteniendo un 63.2% de aprobación de su gobierno. El periódico el Financiero, en octubre de 2019, en una encuesta sobre hechos violentos registrados en la ciudad de Culiacán, Sinaloa, planteó la pregunta: ¿en materia de seguridad pública, ¿qué es lo más importante que debe hacer el gobierno?, a la cual el 60% respondió: combatir el crimen organizado. Las

dos encuestas son ejemplos de aplicación de la inferencia estadística, pues en ambos casos los resultados fueron obtenidos a partir de la selección una muestra aleatoria de personas de una población.

Los resultados de experimentos aleatorizados son también frecuentes en reportes y noticias. Por ejemplo, la Comisión Federal de Comercio de los Estados Unidos (FTC) en septiembre de 2011 emitió una sanción a la empresa fabricante de ropa y calzado deportivo Reebok International Ltd., por declaraciones publicitarias engañosas e infundadas, al señalar que sus modelos de tenis EasyTone y RunTone daban tono y fuerza adicional a los músculos de las piernas y glúteos. La publicidad señalaba que los tenis reforzaban y tonificaban un 28% más los músculos de los glúteos y un 11% más los músculos de las piernas que los tenis habituales, al caminar o correr. Sin embargo, según la FTC, las afirmaciones no tenían ningún fundamento, ya que un estudio experimental aleatorizado, ciego y bien controlado, con al menos 6 semanas de duración, en el que se utilizaron herramientas de medición apropiadas, realizado por personas calificadas y con experiencia en experimentos estadísticos, confirmó que los resultados no concordaban con la publicidad de la empresa.

8.2 Elementos de una inferencia estadística

En un sentido convencional, una inferencia estadística es un enunciado sobre una población o un proceso, el cual es generado a partir de los datos de una muestra o experimento, y con un nivel explícito de confianza. De acuerdo con lo anterior, tres características clave forman parte de una inferencia estadística:

1. Un enunciado de *generalización* "más allá de los datos disponibles".
2. Uso de datos de una *muestra aleatoria* como evidencia para apoyar esta generalización.
3. Un *lenguaje probabilístico* que expresa una medida de la incertidumbre sobre la generalización.

Veamos las tres características o elementos de una inferencia estadística a través de los resultados de una encuesta de opinión. En mayo de 2019, la empresa Parametría realizó una encuesta en vivienda con alcance nacional para conocer la opinión de los mexicanos sobre los servicios que proporciona el IMSS. Seleccionó una muestra aleatoria sistemática de 800 personas mayores de 18 años, apoyándose para ello en el padrón de secciones electorales del Instituto Nacional Electoral (INE). **Se infiere que el 67% de la población mexicana califica Bien o Muy Bien** la atención de los médicos del IMSS. Se reporta una confiabilidad de la encuesta de 95% y un margen de error de muestreo ± 3.5%.

1. Enunciado de *generalización*, que va "más allá de los datos".

Se infiere que el 67% de la población mexicana mayor de 18 años califica de Bien a Muy Bien la atención de los médicos del IMSS. Se han generalizado a toda la población los resultados que se obtuvieron de una muestra de ella.

2. Uso de datos de una *muestra aleatoria* como evidencia para apoyar esta generalización. Los datos han sido obtenidos de las respuestas a preguntas planteadas a una muestra aleatoria de 800 personas.

3. Un *lenguaje probabilístico* que expresa la incertidumbre sobre la generalización. La incertidumbre sobre la generalización se expresa a través del nivel de confiabilidad (95%) y el margen de error por muestreo (±3.5%), lo cual genera un intervalo de confianza que va de 63.5% a 70.5%.

8.3 Poblaciones y muestras

Los investigadores y estadísticos utilizan muestras en lugar de poblaciones por cuestiones prácticas, entre las que destacan el tiempo y el costo; incluso porque en ocasiones es imposible estudiar una población completa. Los métodos estadísticos han logrado avanzar en precisión y confiabilidad, de tal forma que los investigadores se satisfacen con la selección de una muestra, aun cuando saben de antemano que la información de una muestra es parcial, y que los resultados obtenidos pueden no ser iguales a los resultados que se obtendrían si se estudiase la población de manera completa.

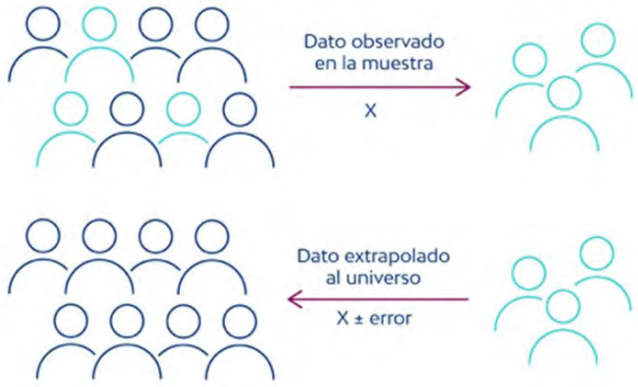
Los conceptos de población y muestra en apariencia son sencillos, pero veremos que la relación entre ellos es más compleja de lo que se piensa. A simple vista parece imposible que una muestra de 800 o 1,000 personas proporcione información lo suficientemente precisa sobre las características de una población de millones de personas; sin embargo, estos tamaños de muestra son bastante frecuentes en encuestas de opinión.

8.4 Parámetros y estadísticos

Comprender el significado de parámetro y estadístico, y la relación que guardan con los conceptos de población y muestra, es un buen inicio en el estudio de la inferencia estadística. Para fijar ideas tomemos como referencia la encuesta de opinión sobre los servicios del IMSS descrita anteriormente.

- *Población:* Todas las personas mayores de 18 años con credencial para votar.
- *Muestra*: 800 personas que fueron seleccionadas aleatoriamente por cierto método de muestreo.
- *Estadístico*: Proporción de la muestra (67%) que califica Bien o Muy Bien la atención de los médicos del IMSS.
- *Parámetro.* Proporción de la población que califica de Bien a Muy Bien la atención de los médicos del IMSS.

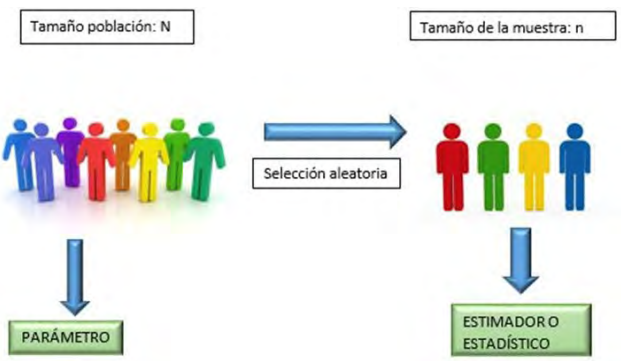
Obsérvese que el parámetro es desconocido, en la práctica estadística siempre lo es, pues es precisamente lo que se desea conocer. Lo que se ha hecho es una *estimación* de su valor a partir de los datos de la muestra; por ello los resultados van acompañados de un margen de error y una medida de confiabilidad.



Para simbolizar un estadístico se utilizan letras minúsculas, mientras que un parámetro se simboliza con letras mayúsculas. En el caso de la proporción de personas que califica Bien o Muy Bien la atención de los médicos del IMSS se puede simbolizar con , y el la proporción poblacional con En el caso de medias aritméticas se utiliza el símbolo para representar la media muestral y el símbolo para representar la media poblacional.

Resumiendo:

Los estadísticos son medidas numéricas (por ejemplo: proporciones y promedios) que se calculan sobre los datos de una muestra. Los parámetros son medidas numéricas que describen características de los elementos de una población (por ejemplo: proporciones y promedios). En este sentido, *los estadísticos son estimadores de los parámetros poblacionales*. En el caso de la encuesta del IMSS, la proporción de personas en la muestra que califica Bien o Muy Bien la atención de los médicos, es una estimación de la proporción de personas en la población que califica Bien o Muy Bien la atención de los médicos.



Ideas importantes

- Las poblaciones se pueden caracterizar por valores numéricos denominados parámetros (por ejemplo: media aritmética, proporción).

- De una población se pueden extraer muestras aleatorias de un tamaño dado, las cuales se puede caracterizar por valores numéricos llamados estadísticos (por ejemplo: media aritmética, proporción).
- Un estadístico puede constituirse en estimador del parámetro de una población. Por ejemplo: la media muestral es un estimador de la media poblacional.
- Una estimación de un parámetro es una aproximación de su valor y que contiene una cantidad de incertidumbre.

Actividad de aprendizaje

En las encuestas que se mencionan en la introducción de este capítulo, sobre la aprobación del presidente Andrés Manuel López Obrador y sobre Seguridad Pública y Violencia:

- Identifica el parámetro poblacional que se desea estimar en cada caso.
- ¿Cuál es la población objetivo en cada caso?
- Investiga en internet el tamaño de muestra utilizado en cada encuesta.

8.5 Variabilidad muestral

Señalamos anteriormente que la proporción muestral es un estimador de la proporción poblacional , y que la media muestral permite estimar la media poblacional . En el caso de la encuesta para evaluar los servicios médicos del IMSS, una muestra de 800 personas mayores de 18 años señala que 67% de ellas califican Bien o Muy Bien la atención de los médicos; si seleccionamos otra muestra, muy probablemente generará resultados diferentes, pues no incluye a las mismas personas. A esta característica del muestreo aleatorio se le conoce como *variabilidad muestral*. Se dice entonces que la proporción muestral *p* es una *variable aleatoria*, un concepto de teoría de la probabilidad que fue visto en el capítulo anterior.

Un conflicto que a muchas personas les genera la interpretación de los resultados de una muestra, tiene que ver con las siguientes preguntas: ¿cómo los resultados de una muestra pueden utilizarse para estimar los valores de una población?, si de una muestra a otra los resultados varían, ¿qué tan confiables pueden ser los resultados de una muestra para hacer una estimación sobre una población, sobre todo cuando la muestra es tan pequeña respecto al tamaño de la población?

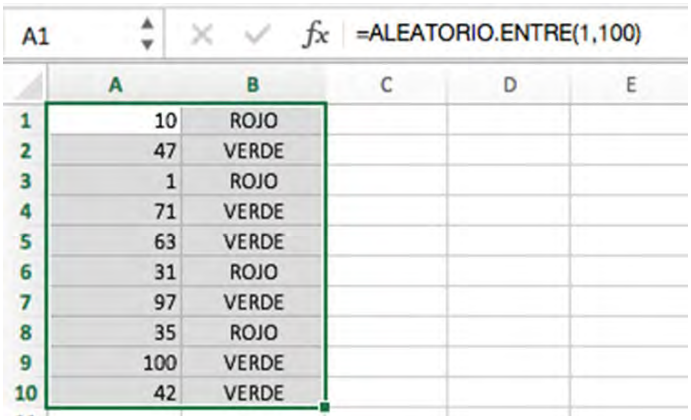
Existen principios y métodos estadísticos que permiten controlar y predecir la variabilidad muestral, de tal manera que una estimación sea confiable y con un margen de error aceptable. Veamos algunas ideas intuitivas a través de la simulación de muestras de una población, utilizando herramientas de software que se ejecutan en una página web de libre acceso, los cuales se conocen como *applets*.

Ideas básicas sobre simulación del muestreo

La simulación es una actividad mediante la cual se pueden extraer conclusiones acerca del comportamiento de un sistema, estudiando el comportamiento de un modelo cuyas relaciones de causa y efecto son las mismas (o similares) a las del sistema original. Para realizar la simulación necesitamos conocer ciertas características del sistema, que en nuestro caso sería la población y el tamaño de la muestra a seleccionar.

Para fijar ideas supongamos que en una población el 60% está de acuerdo con la propuesta A y el 40% con la propuesta B sobre una reforma a una ley. Podemos pensar en un modelo de urna en la que se colocan 100 canicas: 40 rojas que representan a los de la propuesta A y 60 verdes que representan a los de la propuesta B. Sin ver el contenido de la urna, se selecciona una muestra aleatoria de 10 canicas, se observa el color de cada una y se registra el resultado; un posible resultado podría ser 5 rojas y 5 verdes, otro podría ser 3 rojas y 7 verdes. Se regresan las canicas a la urna y podemos hacer una nueva selección. El proceso se repite tantas veces como se desee y se toma nota de las proporciones obtenidas que dan cuenta de la variabilidad muestral.

El proceso anterior es en sí una simulación del sistema, pero una simulación física que puede consumir mucho tiempo en un aula de clase. Podemos optar por una simulación en la computadora. Este tipo de simulación se fundamenta en la función *aleatorio* que está disponible en cualquier software, la cual genera números en forma aleatoria de acuerdo con una distribución de probabilidad previamente establecida. Para el caso que estamos explicando, partimos de la idea de que todos los elementos de la población (todas las canicas de la urna) deben tener la misma probabilidad de ser elegidos. La función *aleatorio* lo que hace es tomar al azar 10 números (pues la muestra es de tamaño 10) del 1 al 100, y ellos representan la muestra aleatoria. Los números menores de 40 son rojos (propuesta A) y los mayores a 40 son verdes (propuesta B). La siguiente figura muestra el proceso anterior en la hoja de *Excel*.



Función *aleatorio entre* (1,100) de *Excel* generando 10 números aleatorios

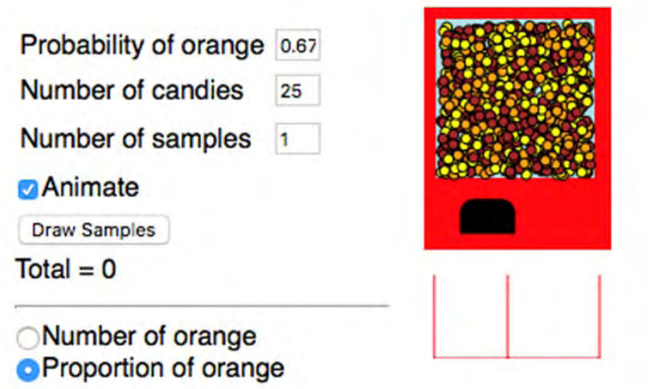
Los applets que utilizaremos para explicar varios conceptos en los siguientes apartados son muy fáciles de utilizar, pues no requieren introducir una función como lo hicimos con *Excel* debido a que ya la tienen programada, solo se requiere introducir el parámetro de la población, el tamaño de muestra y la cantidad de muestras que

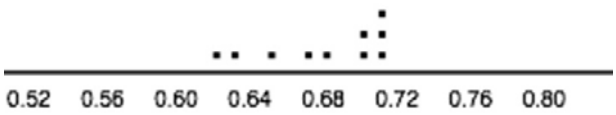
se desean seleccionar. Además, los resultados los presentan en forma gráfica para facilitar su visualización. Veamos algunos ejemplos.

Ejemplo de simulación del muestreo

En el sitio http://www.rossmanchance.com/applets/OneProp/OneProp.htm?candy=1 aparece un dispositivo virtual (applet) que permite simular el muestreo de una población.

Utilicemos de nuevo la encuesta de IMSS y consideremos que el 67% de la población mayor de 18 años califica Bien o Muy Bien la atención de los médicos, es decir P = 0.67. Consideremos la analogía entre la máquina y su contenido con la población con todos sus elementos. Consideremos que las bolas color naranja representan a los que califican Bien o Muy Bien la atención de los médicos; por lo tanto, la probabilidad de obtener una bola color naranja es 0.67. Seleccionamos una muestra de 25 personas y presionamos el botón *Draw Samples* (*Extraer Muestras*) 10 veces para extraer 10 muestras. Verifica que la configuración quede tal como aparece en la figura. Los resultados se muestran en la siguiente gráfica:





Proporciones muestrales obtenidas de la población (P=0.67, n=25)

Un hecho que es observable a simple vista es la variabilidad muestral. La diferencia de cada proporción muestral, respecto al parámetro P=67%, se conoce como *error de muestreo*.

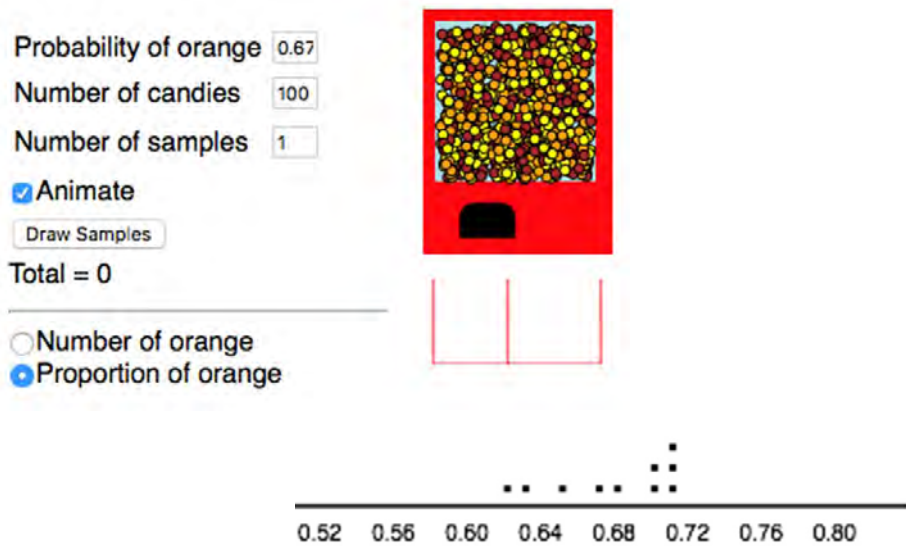
Muestra	Valor de p	Valor de P	Error de muestreo |p-P|
1	48%	67%	19%
2	64%	67%	3%
3	68%	67%	1%
4	68%	67%	1%
5	72%	67%	5%
6	72%	67%	5%

7	72%	67%	5%
8	72%	67%	5%
9	76%	67%	9%
10	80%	67%	13%

En la práctica estadística se selecciona una sola muestra de la población, así que, si hubiésemos seleccionado la primera muestra, habríamos inferido que la proporción de personas que califica Bien o Muy Bien la atención de los médicos es de 48%, lo cual conlleva un error muestral de 19%, pero si hubiésemos seleccionado la segunda muestra, el error hubiera sido de 3%, que es un error aceptable en la teoría de las encuestas.

Tamaño de muestra y variabilidad muestral

Regresemos al dispositivo virtual y cambiemos el tamaño de muestra a 100 personas. Seleccionemos de nuevo 10 muestras de una por una (ver figura).



Proporciones muestrales obtenidas de la población (P=0.67, n=100)

Muestra	Valor de p	Valor de P	Error de muestreo
1	62%	67%	5%
2	63%	67%	4%
3	65%	67%	2%
4	67%	67%	0%
5	68%	67%	1%
6	70%	67%	3%

7	70%	67%	3%
8	71%	67%	4%
9	71%	67%	4%
10	71%	67%	4%

Obsérvese que al incrementar el tamaño de muestra de 25 a 100, el error de muestreo ha disminuido. Ahora, si se promedian todos los valores de se obtiene una media de 67.8%, el cual es muy cercano al valor de P=67%. Este hecho significa que, a pesar de la variabilidad muestral, el promedio de las proporciones muestrales está muy cercano al valor del parámetro que se desea estimar.

Ideas importantes

- Si se seleccionan muestras de un tamaño dado de una misma población, los resultados suelen variar de una muestra a otra. Esta propiedad del muestreo es conocida como variabilidad muestral.
- Cuando se utilizan muestras para hacer inferencias sobre parámetros de una población, se genera inevitablemente un error conocido como error de muestreo.
- El error de muestreo está relacionado con el tamaño de muestra. A medida que se incrementa el tamaño de muestra se puede reducir el error de muestreo.
- En el caso que fuera posible seleccionar muchas muestras de un tamaño dado en forma repetida de la misma población, las estimaciones no deben estar muy lejanas de verdadero valor del parámetro poblacional. De hecho, para la totalidad de muestras posibles de seleccionar, la estimación debe ser igual al parámetro poblacional.

8.6 Distribuciones muestrales

Una distribución muestral representa el valor que puede tomar un estadístico (por ejemplo, la media) en cada una de las muestras aleatorias de un tamaño dado, que son posibles de seleccionar de una misma población. En este sentido, la distribución muestral permite conocer todas las posibilidades que tiene el estadístico de ocurrir, si el muestreo es repetido en las mismas condiciones.

Es importante reconocer que la muestra que se selecciona es solo una del gran conjunto de muestras que podrían ser extraídas de una población, y conocer la distribución muestral para un estadístico particular permite responder la pregunta esencial que caracteriza a la inferencia estadística (recurrente): *¿con qué frecuencia este método daría una respuesta correcta si es utilizado muchas veces en las mismas condiciones?*

Existen dos enfoques para construir la distribución muestral de un estadístico: enfoque teórico que utiliza conceptos de teoría de probabilidad y estadística

matemática; y el enfoque empírico que consiste en simular la selección de muestras de una población. El primer enfoque genera una distribución teórica de probabilidades para el muestreo; el segundo genera una distribución empírica, la cual puede ser una buena aproximación de la distribución teórica en la medida que se incrementa el número de muestras extraídas.

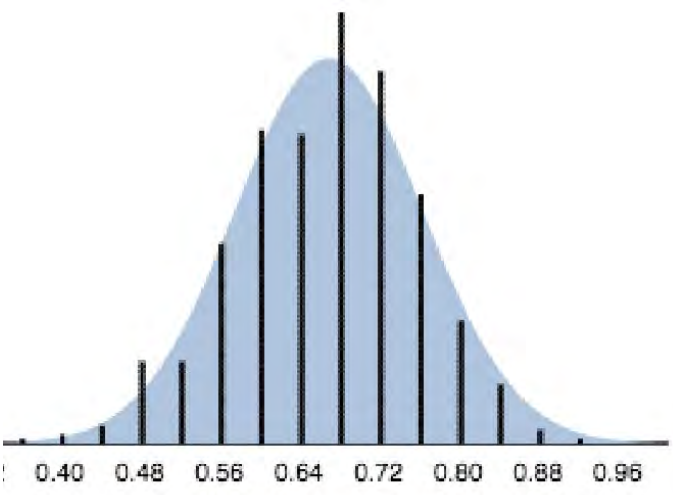
El primer enfoque ha predominado por mucho tiempo en el currículo y los libros de texto, sin embargo, en los años recientes ha cobrado especial importancia el enfoque de simulación basado en frecuencias, dado que requiere menos antecedentes matemáticos y de probabilidad. Este enfoque es conocido como *enfoque informal* a la inferencia estadística.

El acercamiento a los conceptos de inferencia desde una perspectiva informal requiere de herramientas tecnológicas con amplio potencial de representaciones visuales dinámicas para generar imágenes correctas de estos conceptos. En este libro utilizamos el enfoque empírico apoyados en software para simulación del muestreo, como es el caso de *Excel* y algunos applets diseñados por investigadores en didáctica de la estadística.

• Distribuciones muestrales empíricas

En la sección anterior identificamos que una característica intrínseca del muestreo es su variabilidad; la distribución muestral expresa dicha variabilidad y puede ser visualizada gráficamente, con ello es posible determinar rangos de valores que puede tomar un estadístico y su probabilidad de ocurrencia. Ello es sumamente importante para determinar la confiabilidad y precisión de una inferencia sobre un parámetro poblacional. La pregunta que está en el centro de esta discusión cuando se interpretan los resultados de un muestreo es: ¿qué tan confiable puede ser el resultado de una muestra para estimar un parámetro poblacional? El concepto que nos ayuda a responderla es la distribución muestral del estimador.

Vamos a recurrir de nuevo a la encuesta del IMSS, pero seleccionaremos esta vez un mayor número de muestras a las que obtuvimos cuando discutimos la variabilidad muestral. La gráfica que se ha obtenido es una distribución muestral empírica para la proporción de personas que opina Bien o Muy Bien sobre la atención de los médicos, considerando un total de 500 muestras de tamaño 25 cada una.



Distribución muestral de la proporción de personas que opina Bien o Muy Bien de la atención de los médicos en el IMSS.

Obsérvese que la distribución muestral tiene forma acampanada, y que la mayoría de los valores muestrales (valores de se ubican en un intervalo central de la distribución, alrededor del 0.67 que es el valor del parámetro *P*. Hay pocos valores muestrales que se encuentran alejados del centro (colas de la distribución).

Ideas importantes

- Los valores representados en una distribución muestral son proporciones calculadas (u otros estadísticos, por ejemplo, medias) en cada una de las muestras.
- La distribución muestral de algunos estadísticos (por ejemplo, la media y proporción) tiene una forma acampanada con centro en el parámetro de la población, sobre todo para tamaños de muestra superiores a 30.
- La mayoría de los valores muestrales se ubican en un intervalo central alrededor del centro de la distribución.
- Existen valores muestrales alejados del centro de la distribución (colas) pero con poca probabilidad de ocurrir.
- La desviación estándar -conocida también como error estándar- de la distribución muestral de una proporción está determinada por la expresión $\sqrt{\frac{p(1-p)}{n}}$, donde *p* es la proporción muestral y *n* es el tamaño de muestra.

Las ideas anteriores se deducen de un teorema muy importante en estadística conocido como *teorema del límite central*.

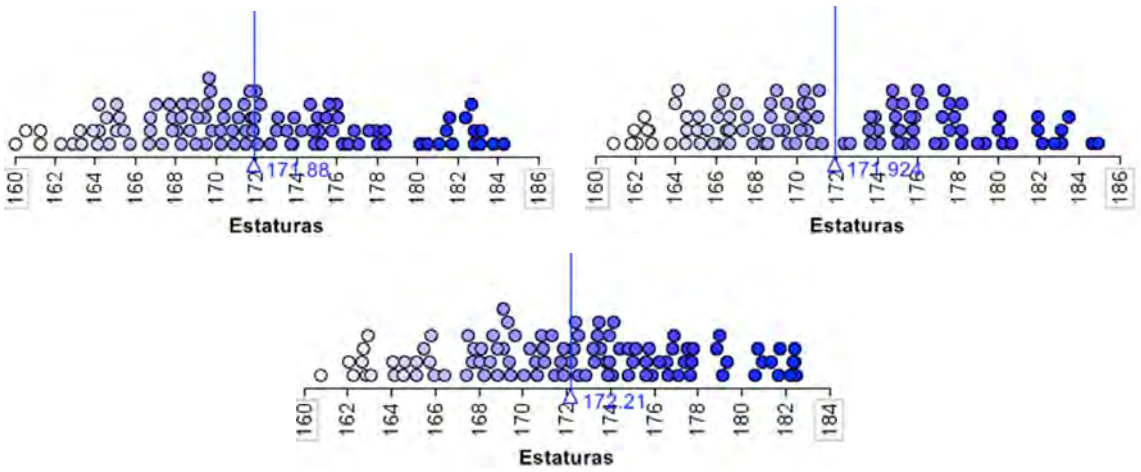
8.7 Esquema para la construcción e interpretación de una distribución muestral empírica

1. Concebir el muestreo como un proceso aleatorio: esto es, seleccionar una muestra de la población, registrar el dato de cada elemento de la muestra y calcular el estadístico en cuestión para estimar el correspondiente parámetro de la población.
2. Imaginar muestras de un mismo tamaño tomadas repetidamente y registrar el valor del estadístico en cada una de ellas.
3. Comprender que este proceso producirá una colección de resultados que serán en su mayoría diferentes del parámetro poblacional que deseamos estimar (la distribución muestral).
4. Comprender que debido al proceso de selección aleatoria hay variabilidad en los resultados, pero en una gran cantidad de repeticiones la distribución de los resultados llegará a ser estable y centrada en el verdadero valor del parámetro.

Expliquemos lo anterior mediante la siguiente situación: supongamos que estamos interesados en conocer la estatura promedio de los estudiantes de una universidad, para lo cual seleccionamos una muestra aleatoria de 100 estudiantes.

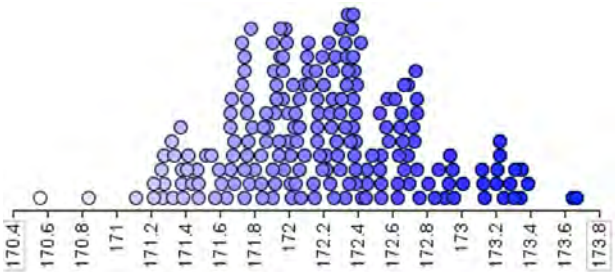
1. Concebir el proceso de muestreo aleatorio a partir de una población. Seleccionar 100 estudiantes de la población, obtener la estatura de cada uno de ellos y calcular la estatura media de la muestra.

2. Imaginar que se pueden tomar muestras repetidas de tamaño 100, registrar la estatura media para cada muestra.



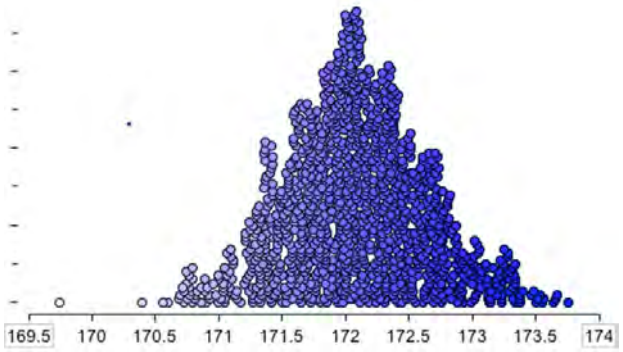
Las gráficas representan tres muestras de 100 estudiantes seleccionados de la población en forma aleatoria. Obsérvese para cada muestra la estatura media: 171.88, 171.92 y 172.21.

3. Comprender que este proceso producirá una colección de medias muestrales, y que la mayoría de las medias será diferente de la estatura media de la población.



Distribución empírica de medias muestrales (200 muestras, n=100)

4. Comprender que, debido a la selección aleatoria de las muestras, existe variabilidad en los resultados.



Distribución empírica de medias muestrales (1,200 muestras, n=100)

La distribución anterior, muestra que la media de las estaturas varía de 169.7 cm. a 173.6 cm. con un valor medio de 172 cm. y una forma acampanada muy aproximada a una distribución normal.

Ideas importantes

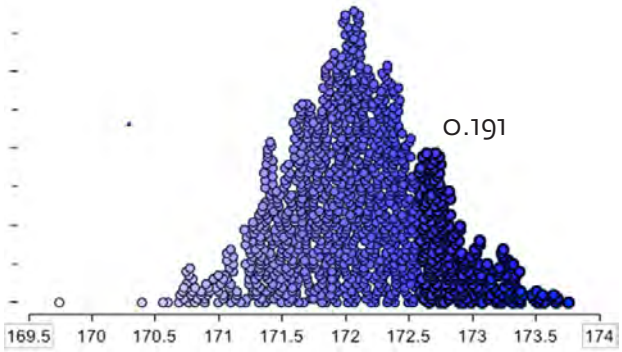
En el proceso de construcción de una distribución muestral aparecen tres distribuciones que es importante distinguir:

1. *La distribución de la población, que en la práctica estadística es desconocida.*
   La distribución de la población está representada por las estaturas de todos los estudiantes de la universidad en cuestión. Puede estar conformada por miles de datos, uno por cada estudiante de la universidad. Si la universidad tiene 50,000 estudiantes, entonces se tienen 50,000 datos.
2. *La distribución de cada muestra que se selecciona de la población.*
   La distribución de la muestra está conformada por todas las estaturas de los estudiantes seleccionados en la muestra. En este caso, la muestra fue de 100 estudiantes; se tienen por lo tanto 100 datos.
3. *La distribución muestral que se obtuvo con las medias de las muestras seleccionadas.*
   La distribución muestral está conformada por las medias de las estaturas calculadas en cada una de las muestras seleccionadas. En este caso utilizamos 1,200 muestras, pero podrían ser muchas más.

Interpretación de una distribución muestral

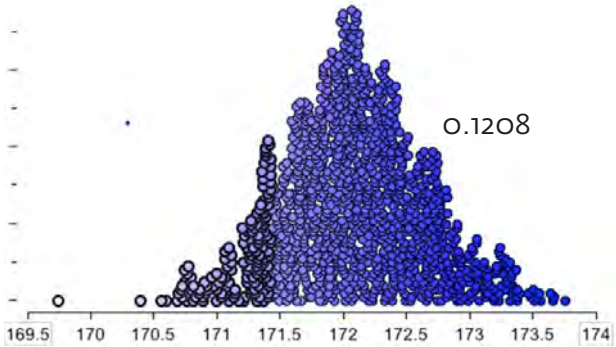
La distribución muestral que hemos obtenido empíricamente en la sección anterior contiene los resultados que se hubieran obtenido si se hubieran seleccionado 1,200 muestras de la población. En la práctica estadística se dispone de solo una muestra, pero conocer la distribución de todas las muestras (o al menos de una gran cantidad) es de suma importancia para hacer una inferencia.

En la siguiente figura nos preguntamos sobre la probabilidad de que una muestra de estudiantes tenga una estatura media superior a 172.59 cm. (véase el área sombreada). Haciendo los cálculos con un software obtenemos que $P(\bar{x} \geq 172.59) = 0.191$



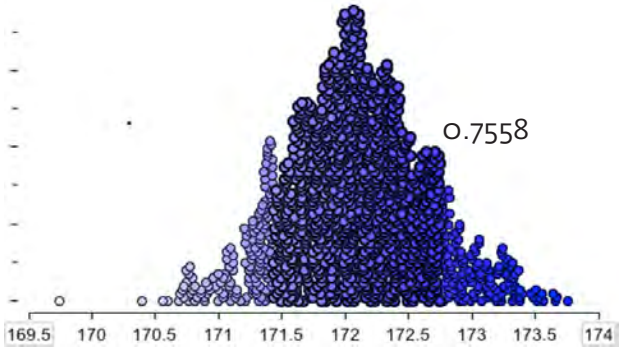
Distribución empírica de medias muestrales (1,200 muestras, n=100)

En la siguiente distribución nos preguntamos sobre la probabilidad de que una muestra de estudiantes tenga una estatura media inferior a 172.59 cm. (véase el área sombreada). El resultado sería: $P(\bar{x} \leq 171.43) = 0.1208$



Distribución empírica de medias muestrales (1,200 muestras, n=100)

Ahora, en la siguiente distribución nos preguntamos sobre la probabilidad de que una muestra de estudiantes tenga una estatura media inferior a 172.75 cm. pero superior a 171.43 (véase el área sombreada).



8.8 Introducción a los métodos de inferencia estadística

El enfoque tradicional para estudiar la inferencia estadística requiere de una buena comprensión de teoría de la probabilidad y estadística matemática. Sin embargo, con el desarrollo de la tecnología computacional, ha surgido un enfoque alternativo basado en simulación del muestreo en forma repetida para generar distribuciones muestrales empíricas.

A este enfoque se le conoce comúnmente como *enfoque informal*, ha cobrado gran auge en los años recientes como alternativa didáctica para la comprensión de los fundamentos de la inferencia estadística. En este libro utilizamos este enfoque apoyándonos en herramientas tecnológicas interactivas, dinámicas, con capacidad de visualización de los datos y sus representaciones.

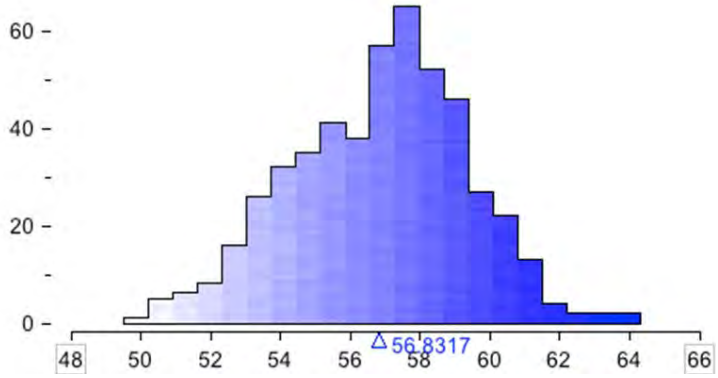
8.9 Estimación de parámetros poblacionales

La estimación de parámetros tiene como base conceptual la distribución muestral del estadístico en cuestión. Desde el enfoque informal, la distribución muestral que vamos a utilizar es empírica, es decir, es construida con los valores del estadístico que resultan en cada muestra seleccionada (o simulada) de la población.

Para introducir el tema consideramos el caso de una encuesta, por ser un ejemplo prototípico de estimación de parámetros. En febrero de 2020, en plena crisis del coronavirus, el periódico el Financiero publicó una encuesta nacional realizada por vía telefónica a una muestra probabilística de 410 mexicanos. La confiabilidad reportada de la encuesta es de 95% y el margen de error de las estimaciones es de ±4.8%. Ante la pregunta ¿a usted qué tanto le preocupa el tema del coronavirus?, el **57% de los encuestados respondió que les preocupa mucho.**

El 57% es una estimación puntual de la proporción de mexicanos que les preocupa mucho el tema del coronavirus. Si se selecciona otra muestra de igual tamaño en las mismas condiciones, ¿qué tanto podría variar el resultado? La respuesta, como vimos en los apartados anteriores, la proporciona la distribución muestral de la proporción.

Supongamos (y es bastante razonable hacerlo) que el verdadero porcentaje de la población que le preocupa mucho el coronavirus es de 57% (esto es P=0.57). Una simulación de 500 muestras de tamaño 410 de una población (como la que utilizó la encuestadora) con P=0.57 nos conduce a la siguiente distribución muestral en su forma empírica:



Distribución muestral empírica de la proporción **p** (preocupa mucho el coronavirus) (500 muestras, n=410)

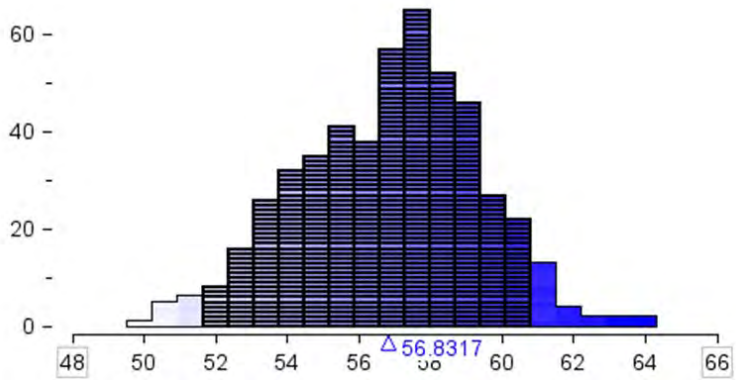
Podemos ver que la proporción muestral *p* (al menos en esas 500 muestras) varía aproximadamente de 50% a 64%. En ese intervalo se encuentra el valor de P que hemos supuesto (P=0.57). Es decir, si repetimos el muestreo muchas veces tenemos mucha confianza en que P se encuentre entre dichos límites; a pesar de ello, puede no ser muy satisfactorio, ya que entre el límite inferior y el límite superior hay una diferencia de 14%, por lo que el intervalo es amplio e impreciso. Posteriormente veremos cómo se puede reducir la amplitud del intervalo para aumentar su precisión.

Otra observación importante consiste en que los valores de *p* se concentran alrededor del promedio 56.83 (podría tomarse como 57, que es justamente el valor del parámetro poblacional P). Esta es una importante propiedad de las distribuciones muestrales: *la media de la distribución muestral está centrada en el parámetro poblacional*. Dicho de otra manera, la media de la distribución muestral es igual al parámetro cuyo valor queremos conocer.

Otra característica importante que se puede visualizar es que la distribución muestral *tiene forma acampanada muy aproximada a la distribución normal*, por tanto, podemos usar las propiedades de la distribución normal para calcular probabilidades de que ocurran ciertas proporciones muestrales y, por ende, de que el parámetro poblacional que se desea estimar se encuentre entre ellas.

Confianza estadística

La *regla empírica* de la distribución normal (vista en el capítulo de probabilidad) establece que el 95% de las muestras se encuentra a desviaciones del centro de la distribución muestral (área sombreada). Esta propiedad es muy utilizada en la estimación de parámetros, pues 95% es un nivel de confianza bastante razonable para una estimación. Usaremos este caso de la regla para explicar los conceptos que se involucran en una estimación por intervalos de confianza.



Distribución con 500 proporciones muestrales **p** (preocupa mucho el coronavirus) (n = 410)

El intervalo central sombreado en la distribución muestral contiene el 95% de las muestras (475 de 500), y está delimitado por los valores 52% y 61%. Se puede interpretar como sigue: se tiene una confianza estadística de 95% que la proporción P de mexicanos a quienes les preocupa mucho el coronavirus se encuentre entre 52% y 61%. Obsérvese que nos referimos al parámetro P, cuyo valor queremos inferir a partir de los valores muestrales *p*.

Recuerda que la distribución muestral de *p* se obtuvo de manera empírica para 500 muestras, si incrementamos el número de muestras los resultados pueden variar ligeramente y la distribución tendrá una forma aún más cercana a la distribución normal.

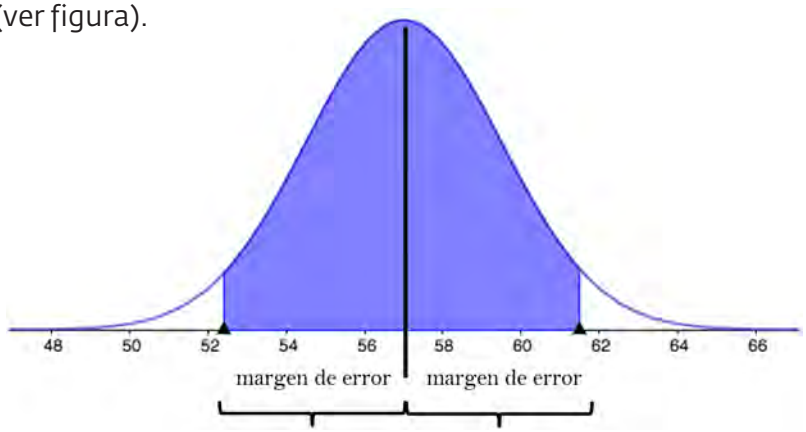
En la práctica no podemos saber si la muestra con la que estamos trabajando es una de las 95 muestras de cada 100 para las cuales el intervalo contiene a P, o si es una de las 5 de cada 100 muestras que no contienen a P. Sin embargo, es satisfactorio saber que el método para estimar P acierta en el 95% de los casos. Este es el significado de la **confianza estadística.**

Error de muestreo o error de estimación

El otro componente que forma parte de una estimación es el *error de muestreo* o *error de estimación*. Se define como la diferencia entre el valor obtenido en la muestra (estimación puntual) y el parámetro que se desea estimar. Si *p* es la proporción muestral y P es la proporción poblacional, el error de muestreo es para una muestra dada.

En un muestreo repetido el error variará, pudiendo ser menor en unos casos que en otros; interesa entonces conocer cuál es el máximo error de muestreo que se puede presentar, para ello los estadísticos han ideado el concepto de *margen de error*. El margen de error es una medida de precisión que indica qué tan lejos puede estar un resultado muestral del parámetro que se desea estimar, como consecuencia de la variabilidad intrínseca del muestreo aleatorio.

El margen de error se puede visualizar como la distancia que existe entre el centro de la distribución muestral y cualquiera de los límites del intervalo, pues es la distancia máxima a la que se puede encontrar una muestra sin salirse del intervalo de estimación (ver figura).



El margen de error para un nivel de confianza de 95% es igual a 2 desviaciones estándar (propiedad derivada de la regla empírica de la distribución normal). Calculemos entonces la desviación estándar de la distribución muestral y multipliquemos por 2 el resultado. En la siguiente sección nos ocuparemos de los cálculos y detalles procedimentales.

Construcción de un intervalo de confianza para un parámetro poblacional

En una estimación por intervalo se proporciona un límite inferior y un límite superior, entre los cuales se espera se encuentre el parámetro poblacional. El intervalo se determina sumando y restando el margen de error de la estimación puntual obtenida de los datos de la muestra. Por ejemplo, si la estimación puntual de P es 0.57 (57%) y se tiene un margen de error de 5%. El intervalo queda formado por (0.52, 0.62), ya que el error se puede presentar en uno o en otro sentido.

Esta imagen del intervalo de confianza ayuda a comprender qué tan lejos puede estar el resultado de la muestra del parámetro poblacional. Sin embargo, en reportes de encuestas y otros estudios es común que se reporte la estimación en forma puntual, y en la descripción de la metodología del estudio, se reporta por el margen de error y otros indicadores como la confiabilidad, tamaño de muestra, método de muestreo.

En el caso del IMSS la estimación puntual de la proporción poblacional de personas que califica Bien o Muy Bien la atención de los médicos es 67%, pero debe tenerse cuidado al interpretar dicho valor, pues si el margen de error es de 3.5%, el verdadero valor de la proporción poblacional puede estar entre 63.5% y 70.5%. Entonces, la forma general de un intervalo de confianza está determinada por:

Estimación puntual±Margen de error

En suma, *un intervalo de confianza es un rango de valores calculado a partir de los datos de una muestra, entre los cuales se estima que podría estar el valor de un parámetro de la población (por ejemplo: la media o la proporción).*

Elementos de un intervalo de confianza

1. La estimación puntual del parámetro.
2. El nivel de confianza.
3. El margen de error.

Estos elementos pueden ser analizados en los resultados de cualquier encuesta. En el caso de la encuesta del coronavirus se utilizó un tamaño de muestra de 410 personas para generar una estimación puntual de 57% de la población que le preocupa mucho el coronavirus. El nivel de confianza utilizado para realizar los cálculos fue de 95% y el margen de error (máximo error de muestreo esperado) fue de 4.8%.

Una vez que hemos comentado los aspectos generales y conceptos que integran un intervalo de confianza, procederemos a explicar los procedimientos de cálculo para el caso de la proporción, uno de los estadísticos más utilizados en las encuestas.

Procedimiento para calcular un intervalo de confianza

1. Se selecciona una muestra aleatoria de una población.
2. Se calcula el valor muestral de interés (por ejemplo, una proporción) mismo que constituye una estimación puntual del valor del parámetro que se desea conocer.
3. Se elige un cierto nivel de confianza para la estimación (95% es un valor muy utilizado)
4. Se calcula el margen de error,
5. Se determina el intervalo sumando y restando el margen de error a la estimación puntual.

Intervalo de confianza de 95% para una proporción

- Forma general del intervalo

Estimación puntual ± margen de error

- Forma particular del intervalo para el caso de la proporción

$$p \pm 2\sqrt{\frac{p(1-p)}{n}}$$

donde:

p Proporción muestral (estimación puntual)

$\sqrt{\frac{p(1-p)}{n}}$ Desviación estándar de la proporción muestral

$2\sqrt{\frac{p(1-p)}{n}}$ Margen de error

Obsérvese la relación que tiene el margen de error con el tamaño de muestra n, el valor de p y el nivel de confianza definido por el número 2.

- El nivel de confianza (expresado por el número 2) tiene un efecto directamente proporcional en el margen de error, por lo que aumentar este valor también aumenta el margen de error.
- El tamaño de muestra aparece en el denominador, por lo que su efecto es inversamente proporcional al margen de error, es decir, a mayor tamaño de muestra disminuye el margen de error.
- El valor de p puede varíar entre 0 y 1 según la muestra seleccionada.

Apliquemos lo anterior al caso de la encuesta sobre el coronavirus: *p=0.57 y n=410.*

$$\text{Margen de error (m)} = 2\sqrt{\frac{0.57(0.43)}{410}} = 4.9\%$$

Señalamos anteriormente que la estimación de un parámetro puede presentarse de dos formas. Veamos:

1. Al 57% de los mexicanos les preocupa mucho el coronavirus. El estudio tiene una confianza estadística de 95% y un margen de error de 4.9%.
2. El porcentaje de mexicanos que les preocupa mucho el coronavirus está entre 52.1% y 61.9%. El estudio tiene una confianza estadística de 95%.

Tamaño de muestra para estimar una proporción

El tamaño de la muestra que se debe utilizar para estimar un parámetro poblacional es una de las principales preguntas en la fase de la recopilación de los datos. En los párrafos anteriores hemos visto que en cálculo del margen de error se entrelazan el nivel de confiabilidad, el tamaño de la muestra y el valor de la proporción muestral. De manera más precisa:

$$\text{Margen de error (m)} = 2\sqrt{\frac{p(1-p)}{n}}$$

Despejando el valor del tamaño de muestra (n), la expresión queda de la siguiente manera:

$$n = \frac{4p(1-p)}{m^2}$$

El tamaño de muestra *n* depende de los valores de *m* y *p*. El valor de *m* (margen de error) se establece a criterio del investigador (valores de 3% a 5% son muy aceptables), el valor de *p* es desconocido en esta etapa del estudio, pero se sabe que puede variar entre 0 y 1; de hecho, el valor de p=0.5 maximiza el valor del tamaño de muestra, por lo que se utiliza para realizar el cálculo. Sustituyendo los valores p=0.5 y m=0.03, la expresión queda de la siguiente manera:

$$n = \frac{4\times0.5(1-0.5)}{(0.03)^2} = 1,111$$

Ahora consideremos de nuevo la encuesta del coronavirus. Sustituyamos p=0.5 y m=0.049 (margen de error reportado por la encuestadora):

$$n = \frac{4 \times 0.5(1-0.5)}{(0.049)^2} = 416$$

La encuestadora utilizó una muestra de 410 personas para el estudio, lo cual es muy cercano al tamaño de muestra que hemos calculado.

Relación entre tamaño de muestra y margen de error

Si utilizamos un software para realizar cálculos para diferentes tamaños de muestra, considerando un nivel de confianza de 95%, se obtiene la siguiente relación entre el

tamaño de muestra y el margen de error:

Del análisis de la gráfica se observa lo siguiente:

- Conforme aumenta el tamaño de muestra el margen de error disminuye.
- Si el máximo margen de error que se considera aceptable es del 5%, este se puede obtener con un tamaño de muestra de 400 aproximadamente.
- En teoría de encuestas el margen de error más utilizado es de 3% a 4%, lo cual se logra con un tamaño de muestra entre 800 y 1200.

Actividad de aprendizaje

Ingresa al sitio web de la encuestadora Parametría y analiza los resultados de la encuesta que se encuentra en la siguiente liga: http://www.parametria.com.mx/carta_parametrica.php?cp=5140. Revisa la nota metodológica y utiliza la gráfica anterior para verificar si el margen de error es compatible con tamaño de muestra y la confiabilidad reportada.

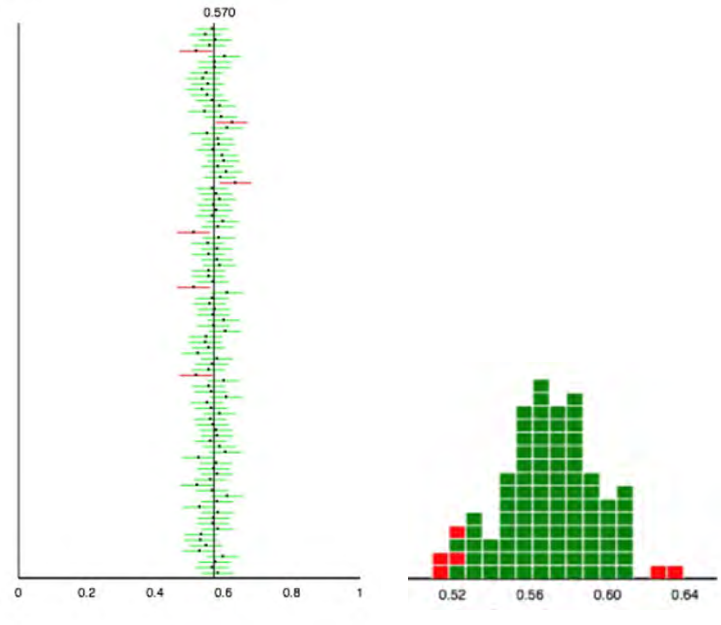
Relación entre tamaño de muestra y confiabilidad estadística

Una imagen visual que puede ayudarnos a comprender la confiabilidad de un intervalo de confianza se obtiene simulando una cierta cantidad de intervalos, y visualizando la cantidad de aquellos que contienen al parámetro. Siguiendo con la información de la encuesta del coronavirus, (nivel de confianza 95%, tamaño de muestra de 410 personas) introduzcamos los valores en el applet que aparece en http://www.rossmanchance.com/applets/ConfSim.html y simulemos 100 muestras con sus respectivos intervalos.

Los intervalos son los segmentos de color verde y rojo, el punto que está en el medio de cada intervalo representa la estimación puntual obtenida de cada muestra; sumando y restando el margen de error se obtiene el ancho de cada intervalo; la línea vertical representa el parámetro P=0.57. Se observa que, de los 100 intervalos simulados, 94 capturan al parámetro (color verde), ya que intersectan la línea y 6

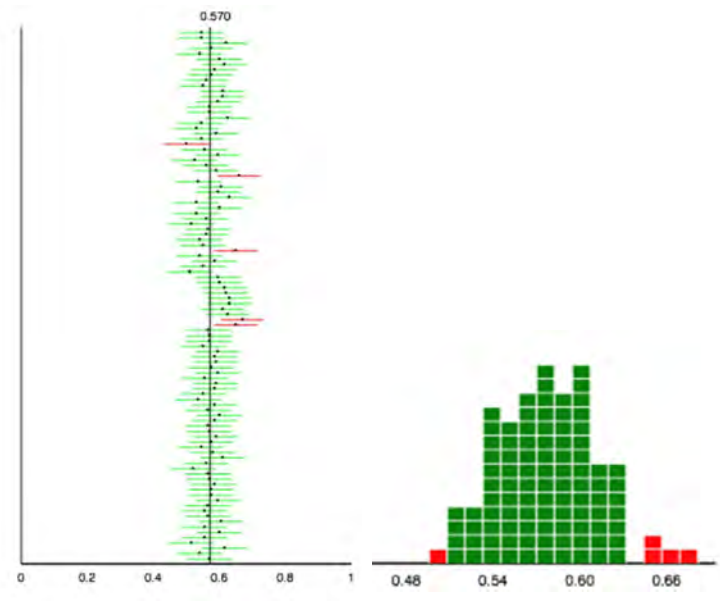
intervalos no lo capturan (color rojo).

Otra información relevante que se obtiene de la simulación de intervalos es la gráfica que aparece en la parte inferior derecha, la cual representa a la distribución muestral de las proporciones obtenidas en cada muestra. Los cuadros rojos en las colas de las distribuciones representan las muestras que por estar alejadas generaron intervalos que no capturan al parámetro.



Simulación de 100 intervalos de confianza de 95% para muestras de tamaño 410

Con el propósito de ver el efecto del tamaño de muestra en los intervalos, utilicemos ahora muestras de tamaño 200, en lugar de 410 como lo hicimos anteriormente.



Simulación de 100 intervalos de confianza de 95% para muestras de tamaño 200

Obsérvese que los intervalos ahora son más amplios que cuando la muestra era de 410. La cantidad de intervalos capturados fue de 95 de 100. En el primer caso se capturaron 94% de los intervalos y en el segundo el 95% de los intervalos. Este porcentaje de intervalos capturados tiene relación directa con el nivel de confiabilidad de 95% que se estableció previamente. El nivel de confiabilidad no depende del tamaño de muestra, sino que es un indicador que se define previamente antes de calcular un intervalo. Los valores más frecuentes que se utilizan en el cálculo de intervalos son 90%, 95% y 99%, pero por mucho el 95% es el valor más utilizado.

Algunas concepciones erróneas sobre los intervalos de confianza

1. *El nivel de confianza es la probabilidad de que el intervalo contenga al parámetro.*

   Esta confusión es muy frecuente, pues el nivel de confianza se expresa como una probabilidad, pero no de esa manera. El nivel de confianza representa la frecuencia con la cual, en un muestreo de la misma población y repetido en condiciones idénticas, el intervalo construido contiene al parámetro. Una idea correcta del nivel de confianza se puede desarrollar con la simulación de muchos intervalos y ver cuántos de ellos contienen el parámetro.

3. *Si la confiabilidad de un intervalo de confianza es de 95%, entonces el margen de error es de 5%.*

   Muchas personas tienen la concepción equivocada que el nivel de confianza de un intervalo y el margen de error se complementan para sumar el 100%. El margen de error indica que tan alejado puede estar la estimación puntual del valor del parámetro poblacional. Valores comunes del margen de error en encuestas oscilan entre 3 y 5%. Por su parte, el nivel de confianza o confianza estadística representa el porcentaje de intervalos que contendrían al parámetro, en caso que el muestreo se repitiera en las mismas condiciones muchas veces. El valor más utilizado en la confiabilidad de encuestas es de 95%.

4. *Un intervalo de confianza es más amplio conforme aumenta el tamaño de muestra*

   Al contrario, la amplitud de un intervalo de confianza disminuye al incrementar el tamaño de muestra. Lo mismo sucede con el margen de error pues está en estrecha relación con la amplitud del intervalo. Recuerda que el intervalo se forma restando y sumando el margen de error de la estimación puntual.

5. *Un intervalo con 90% de confianza es más amplio que un intervalo con 95% de confianza (para los mismos datos).*

   Esta es una confusión que tienen muchas personas. Al contrario, 95% de confianza implica mayor amplitud en el intervalo. Ello también es cierto para el margen de error. Para un mismo conjunto de datos, si se quiere mayor confianza, se aumenta el margen de error.

8.10 Las pruebas de significación

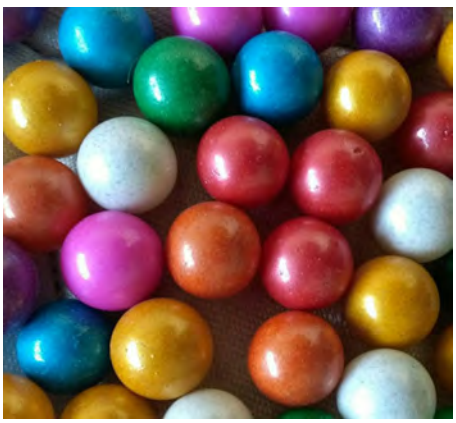
En la historia de las pruebas de hipótesis (ver nota histórica al final del capítulo) se pueden identificar dos enfoques en su conceptualización, el enfoque de Fisher y el enfoque de Neyman-Pearson. Fisher les llamó pru*ebas de significación* y Neyman-Pearson le llamaron *prueba de hipótesis*. Nuestro enfoque será en las pruebas de significación de Fisher, por lo que haremos referencia continua a ellas en el capítulo.

Las pruebas de significación (PS) evalúan si los datos de una muestra aleatoria seleccionada de una población apoyan o rechazan una hipótesis sobre el valor de un parámetro poblacional (por ejemplo, una proporción o una media). En la práctica el propósito de una PS es responder la pregunta: *¿los resultados de la muestra han sucedido por azar o constituyen evidencia a favor de la hipótesis sobre el valor del parámetro?*

Modelo de urna y simulación de muestras para explicar el razonamiento de las PS

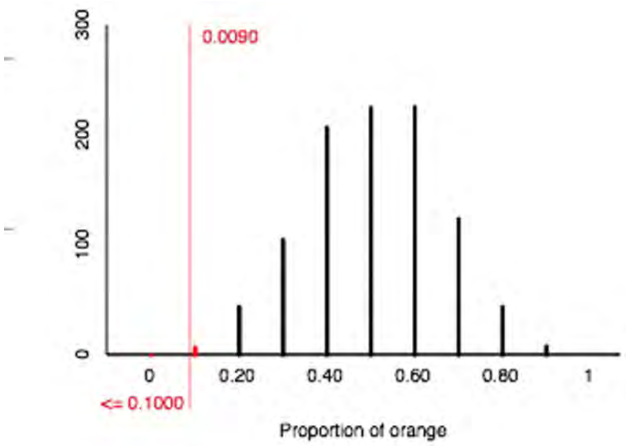
Supóngase que en una población el 50% de las personas opinan a favor del uso de la marihuana para fines medicinales. La población se puede representar de manera concreta por una gran urna repleta de canicas mezcladas en forma homogénea, en la cual el 50% son de color naranja (a favor de la legalización de la marihuana).

Un estudiante selecciona 10 canicas de manera aleatoria y obtiene 1 canica color naranja; el resultado lo hace dudar sobre si verdaderamente en la urna existe la proporción de 50% de canicas naranja. Esto es, en la muestra de 10 personas, solo una está a favor de la legalización, cuando en la población lo está el 50% de las personas.



Con los datos de la muestra, ¿puede el estudiante refutar la *hipótesis que hay una proporción de 50% de canicas naranja?* Para responder la pregunta optaremos por simular la selección de muchas muestras y observar la frecuencia con la que sucede dicho resultado. Es decir, se requiere conocer la distribución muestral de la proporción de canicas color naranja.

Partiremos de la hipótesis (a la que llamaremos hipótesis nula) que la proporción de canicas naranja en la urna es igual a 50% (P=0.50). Simulamos 1,000 muestras de tamaño 10 en un software y utilizaremos como resultado de interés la *proporción de canicas naranja en cada muestra*. El software registra la proporción de canicas naranja en cada muestra y reporta la siguiente distribución.



Distribución muestral de la proporción de canicas naranja (1,000 muestras de tamaño 10 cada una)
http://www.rossmanchance.com/applets/OneProp/OneProp.htm?candy=1

De la gráfica se observa que muestras con 1 canica naranja o incluso menos, ocurren muy pocas veces (9 de cada 1000 veces), traducido a probabilidades esto es 0.009. Podemos decir entonces, que bajo la hipótesis que el 50% de las canicas son naranja, la probabilidad de obtener una muestra de tamaño 10 que incluya una canica naranja o menos es de 0.009. A este valor de la probabilidad se le conoce técnicamente como **p-valor** y ocupa un rol muy importante en la decisión de rechazar o aceptar la hipótesis. Hay dos posibles explicaciones del resultado obtenido:

1. La hipótesis nula es cierta (P=0.50) y la muestra obtenida se ubica en una cola de la distribución (ocurrió por azar y no porque haya menos canicas color naranja en la población).
2. La hipótesis nula no es cierta y su valor es menor al especificado (P<0.50), por eso la muestra resultó con pocas canicas naranja (ocurrió porque en realidad hay menos canicas color naranja en la población).

¿Cómo decidir entre las dos posibilidades? La metodología de las PS utiliza criterios probabilísticos para la decisión. El aspecto central de la decisión radica en definir un valor de probabilidad a partir del cual se considera que la *muestra es inusual*, y no producto del azar sino de otros factores que tienen efecto en los resultados. A dicha probabilidad se le conoce como **significación estadística.**

En la práctica estadística se ha adoptado la convención de utilizar los valores de y . A dichos valores se le conoce como **niveles de significación estadística.** Cuando una muestra tiene un *p valor* menor al nivel de significación estadística elegido (, se considera una muestra inusual que contradice la hipótesis nula; se dice entonces que los resultados son **estadísticamente significativos** al nivel de significancia .

En suma, un resultado se considera estadísticamente significativo al nivel cuando es muy poco probable que suceda por azar. Cuanto más pequeño sea el *p valor* más significativo es el resultado, por ende, más fuerte es la evidencia de que

los resultados no se deben al azar sino al efecto de otros factores. Sin embargo, es importante enfatizar sobre el error que se puede cometer al rechazar la hipótesis nula, pues a pesar de la poca probabilidad de un resultado muestral, este puede ocurrir aun siendo cierta la hipótesis; por ello el valor del nivel de significación suele ser muy pequeño para minimizar el riesgo de cometer este error.

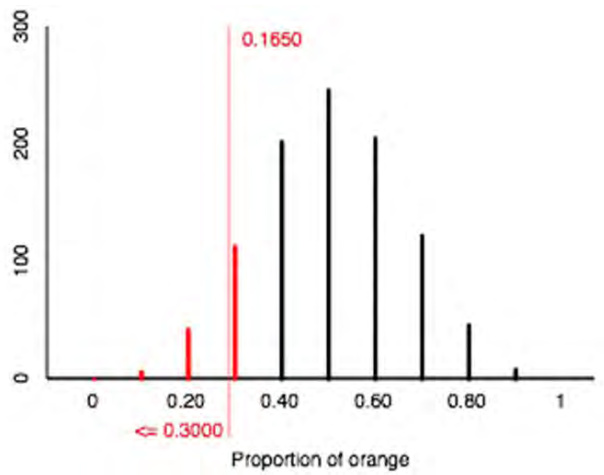
De regreso al problema de la legalización de la marihuana

Ahora que hemos discutido sobre la lógica y el razonamiento de las PS, volvamos al contexto de la urna con 50% de canicas naranja que utilizamos como analogía de una población en la que el 50% de las personas está a favor de la legalización de la mariguana. Los resultados muestrales tienen muy poca probabilidad de ocurrir (p=0.009) bajo la hipótesis de que el 50% de las personas está a favor de la legalización, lo que pone en duda la validez de la hipótesis.

Si elegimos un nivel de significación , el valor de *p* es menor a por lo que la hipótesis debe ser rechazada. Incluso si somos exigentes y elegimos un nivel de significación más pequeño , el valor *p* sigue siendo menor a , por lo que la hipótesis también es rechazada. Se dice entonces que los resultados son estadísticamente significativos y por lo tanto hay evidencia estadística para rechazar la hipótesis nula (P=0.50) a un nivel de significancia del 5% para el primer caso, o del 1% para el segundo caso.

Una extensión del problema

Consideremos las mismas condiciones del problema de la marihuana, solo que ahora la muestra seleccionada obtuvo 3 canicas naranja, es decir, en la muestra de 10 personas 3 se manifiestan a favor de la legalización. Bajo la hipótesis de la proporción de personas que están a favor de la legalización (P=0.50), simulamos 1,000 muestras de tamaño 10 y los resultados se muestran en la siguiente distribución:



De la gráfica se observa que muestras con 3 personas que están a favor o menos, ocurren 165 veces de 1 000; es decir el p valor es igual a 0.165. ¿es este resultado inusual bajo la hipótesis que el 50% de las personas están a favor de la legalización?; desde luego que no, pues es mayor al nivel de significación que hemos elegido para realizar la prueba. En conclusión, la hipótesis P=50% no se rechaza, y se acepta que el resultado de la muestra de 10 personas con 3 de ellas a favor es un resultado plausible bajo la hipótesis; su distancia respecto del parámetro de debe a la variabilidad muestral y no a que en la población haya menos del 50% de personas a favor de la legalización.

Elementos y proceso prueba de una PS

• Hipótesis nula:

Por lo general es una afirmación en forma de igualdad o no diferencia, que involucra el valor de un parámetro poblacional. El propósito de la prueba es recabar evidencia a través de los datos de una muestra en contra de la hipótesis nula. Se abrevia mediante . En el problema de la legalización de la mariguana se afirma que el 50% está a favor de ella, por lo que la hipótesis nula queda de la siguiente manera:

$$H_0: P=0.50$$

• Valor p

Este valor se utiliza como una medida de contundencia de la evidencia en contra de Representa la probabilidad de encontrar un resultado tan alejado o más del parámetro, como el valor observado en la muestra. Se simboliza como *p valor*. Cuanto menor el *p valor* es, menor probabilidad que los resultados sucedan solo por azar y por tanto es más fuerte la evidencia en contra de la hipótesis nula.

• Significación estadística (nivel de significación)

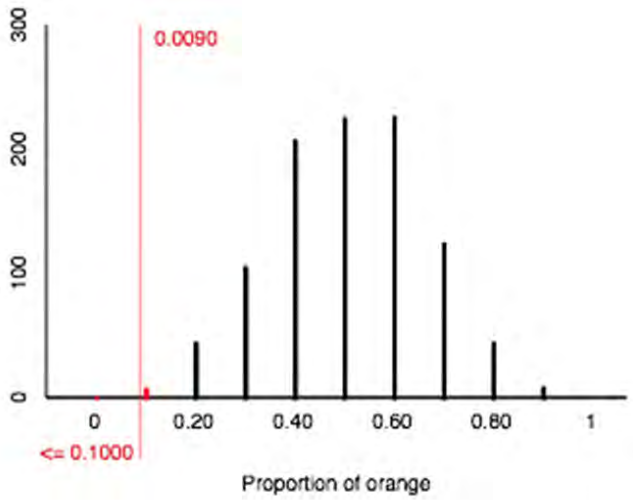
Se define como la probabilidad de rechazar la hipótesis nula siendo cierta. Es un valor que se define antes de realizar la prueba, para saber de antemano que nivel de significación tendrán los resultados si la hipótesis nula es rechazada. Se simboliza con . Valores muy frecuentes de son 0.05, 0.01 y 0.001.

• Criterio de decisión sobre la hipótesis nula

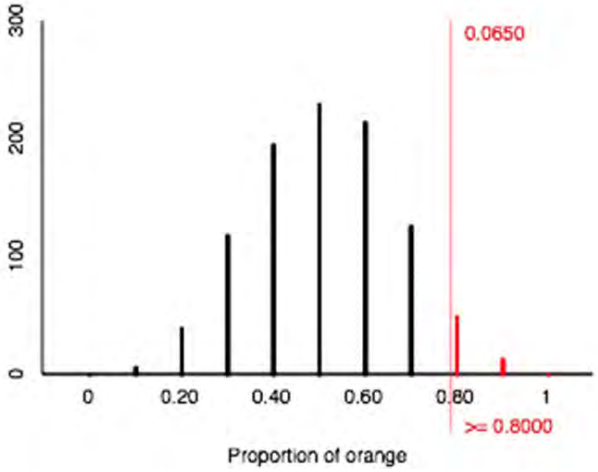
Si el *p valor* es menor al nivel de significación, la hipótesis nula se rechaza al nivel se significación .

Diversas configuraciones de una Prueba de Significación

El problema de la legalización de la mariguana que hemos discutido conduce a una PS de una cola, específicamente a una PS de cola izquierda. Esto se debe a que estamos interesados en desviaciones de la hipótesis nula en la dirección izquierda.



Sin embargo, supongamos que la sospecha es que más del 50% de la población está de acuerdo con la legalización, ahora estamos interesados en desviaciones de la hipótesis nula en la dirección derecha.



El otro caso que se puede presentar es cuando estamos interesados en desviaciones de la hipótesis nula en las dos direcciones, entonces se dice que la PS es de dos colas.

Ideas importantes

- Las PS se proponen evaluar la evidencia de los datos de una muestra en contra de una hipótesis (hipótesis nula) sobre un parámetro poblacional. Las PS se refieren a parámetros, nunca a estadísticos.
- La hipótesis nula se plantea en forma de igualdad o no diferencia con otros parámetros. Su rechazo implica que los resultados no se deben al azar sino a otras causas.
- La decisión sobre si el parámetro tiene determinado valor se toma con base en criterios probabilísticos, por lo que la decisión conlleva un componente de incertidumbre. Se puede rechazar la hipótesis nula siendo cierta y viceversa.
- Resultados alejados del valor del parámetro fijado en la hipótesis nula ponen en duda la validez de la hipótesis nula.
- La probabilidad de obtener un resultado o más alejado al valor obtenido se conoce como *p valor*. El *p valor* mide la fuerza de la evidencia estadística en contra de la hipótesis. Entre más pequeño es el *p valor* más fuerte es la evidencia en contra de la hipótesis.
- La probabilidad de rechazar la hipótesis nula siendo cierta se conoce como nivel se significación, se representa por la letra $\propto$.
- Un resultado es significativo si tiene poca probabilidad de ocurrir ($p<\propto$). Resultados significativos siempre implican el rechazo de la hipótesis nula.

Algunas concepciones erróneas sobre las pruebas de significación

- La hipótesis nula puede referirse tanto a la población como a la muestra.
- La hipótesis nula solo puede referirse al valor de un parámetro poblacional.
- El nivel de significación α es la probabilidad de que la hipótesis nula sea cierta dado que fue rechazada.
- A la inversa, el nivel de significación α es la probabilidad de rechazar la hipótesis nula, dado que fue cierta.
- Una prueba de significación es cómo una prueba de demostración matemática, por lo tanto, cuando no se rechaza la hipótesis nula significa que ha sido probada como cierta, y cuando se rechaza ha sido probada como falsa.

Una prueba de significación no es una prueba matemática que determina con toda certidumbre la validez o no de la hipótesis nula. La aceptación o rechazo de la hipótesis nula está sujeta a criterios probabilísticos, por lo que, a pesar de ser rechazada o aceptada la hipótesis nula, puede que sea falsa y viceversa.

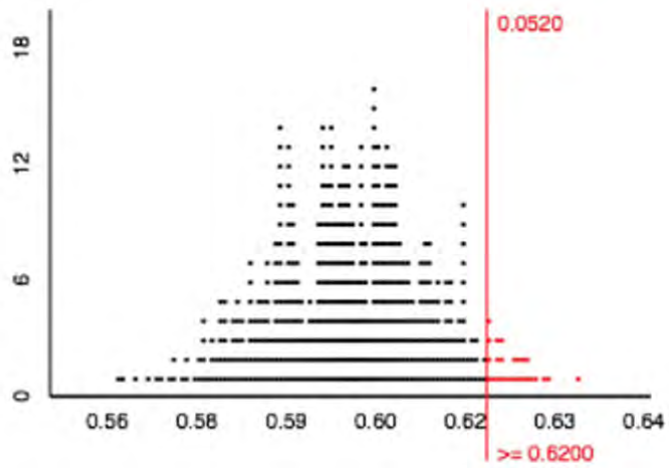
Actividad de aprendizaje

La encuestadora Parametría en 2015 realizó una encuesta a la que tituló ¿en quién confían los mexicanos? (http://www.parametria.com.mx/carta_parametrica.php?cp=4815). Se utilizó una muestra aleatoria de 1,600 mexicanos mayores de 18 años y, entre otros resultados, se obtuvo el siguiente: *seis de cada diez mexicanos (60%) dijeron tener mucha o algo de confianza en el ejército.* Consideremos este valor como el verdadero porcentaje poblacional (P=0.60).

Supongamos que desconfiamos del resultado de la encuesta y creemos que el porcentaje debe ser mayor, para lo cual hacemos nuestra propia encuesta considerando el mismo tamaño de muestra que utilizó la encuestadora y las mismas condiciones de muestreo. Los resultados de la muestra revelan que el *62% de los mexicanos tienen mucha o algo de confianza en el ejército*. ¿podría ser ésta una muestra inusual o en realidad más del 60% de los mexicanos tiene mucha o algo de confianza en el ejército?

Hagamos una prueba a un nivel de significación de 1% (a=0.01), tomando la hipótesis nula como . Simulamos 500 muestras de tamaño 1,600 con este parámetro para ver que ocurre si el muestreo se repitiera muchas veces.

http://www.rossmanchance.com/applets/OneProp/OneProp.htm?candy=1



La distribución muestral obtenida nos señala que, en las 500 muestras, 26 tuvieron una proporción igual o mayor a 0.62, es decir el *p valor* es igual a 0.052. Dado que el valor *p* es mayor al nivel de significación elegido, no rechazamos la hipótesis nula. Concluimos que no hay evidencia estadística para rechazar la hipótesis de que el 60% de los mexicanos tiene mucha o algo de confianza en el ejército.

- Considera el mismo valor muestral (p=0.62) pero con un nivel de significación de 5% (α=0.05). Realiza la prueba y discute los resultados.
- Considera que la muestra obtenida generó una p=0.63. Realiza la prueba y discute los resultados con α=0.05.

Nota histórica

Los orígenes de la inferencia estadística se remontan al siglo XVI, cuando los matemáticos empezaron a darse cuenta que muchos conceptos de probabilidad no podían separarse de la estadística y, como consecuencia, empezaron a considerar modelos probabilísticos para inferir propiedades de la observación de datos. Sin embargo, es hasta el siglo XX cuando la inferencia estadística empieza a tener su mayor desarrollo como disciplina científica, específicamente en las décadas de 1920 y 1930 con los trabajos desarrollados por Ronald Fisher (1890-1962), Jerzy Neyman (1894-1981) y Eagon Pearson (1895-1980) quienes, apoyados en los avances de la teoría de probabilidades y la estadística matemática, sentaron las bases teóricas y metodológicas.

El enfoque de Fisher, conocido como pruebas de significación, se caracteriza por definir únicamente una hipótesis (hipótesis nula) y a partir de la ella y con base en la distribución muestral del estadístico de prueba se estima la probabilidad de una muestra de datos para decidir sobre el rechazo o no rechazo de la hipótesis. Los datos solo permiten rechazar la hipótesis, pero no pueden confirmarla. Por su parte, el enfoque de Neyman-Pearson denominado prueba de hipótesis, se caracteriza por la adición de una hipótesis alternativa en contraposición con la hipótesis nula, lo que conduce a la definición de regiones (de rechazo y no rechazo) y errores asociados a la decisión sobre la hipótesis nula, denominados errores tipo I y tipo II.

En este sentido, en el enfoque de Neymar-Pearson una prueba de hipótesis es una regla de comportamiento inductivo que permite elegir entre una hipótesis nula y una hipótesis alternativa. Sin embargo, la integración de los dos modelos por parte de estadísticos, investigadores y autores de libros de texto se hizo práctica común desde sus orígenes. Es decir, al aplicar las pruebas de hipótesis comúnmente se utilizan elementos de los dos enfoques de forma ecléctica, dando lugar a un modelo híbrido que a su vez ha sido causa de controversias y críticas en su aplicación.

Para tu reflexión

Para comprender la lógica que subyace a las pruebas de hipótesis estadísticas es

importante partir del concepto de hipótesis de investigación. Una hipótesis de investigación es un enunciado que establece un investigador en el cual se ofrece una posible respuesta a una pregunta de investigación. Por ejemplo, un investigador educativo puede plantear la siguiente hipótesis: *el uso de software educativo con representaciones dinámicas de los datos acompañado del uso de datos reales, facilita la exploración y contribuye a una mejor comprensión del análisis descriptivo de datos*. Dicho enunciado como tal, tiene dos valores de "verdad": verdadero o falso. En este sentido, la investigación se realiza para determinar el valor de verdad que corresponde a la hipótesis de investigación; es decir, se realiza un estudio para obtener los datos que contengan la información necesaria para dar respuesta a la pregunta de investigación y decidir si la hipótesis de investigación se rechaza o no.

Por su parte, las hipótesis estadísticas son afirmaciones acerca de los parámetros, por lo que para probar la validez de una hipótesis de investigación es necesario plantearla en término de hipótesis estadísticas. De esta manera, y teniendo en cuenta el ejemplo del software educativo mencionado anteriormente, consideremos que se selecciona una muestra de estudiantes, a partir de la cual se forman dos grupos en forma aleatoria con condiciones iguales en todas las variables sobre el conocimiento del tema, para evitar que los resultados puedan deberse a otros factores que no son de interés; uno toma la clase de análisis descriptivo de datos de forma tradicional (grupo control) y otro toma la clase usando el software educativo (grupo experimental). Al final del curso se aplica un mismo cuestionario para determinar los puntajes promedio de cada grupo. La decisión involucra el planteamiento de hipótesis sobre los puntajes promedio de ambas poblaciones teóricas de estudiantes (los que recibieron la enseñanza tradicional y los que recibieron la enseñanza con software educativo).

En el enfoque de Fisher se plantea una sola hipótesis estadística (hipótesis nula), que suele ser expresada en términos de no diferencia. Para nuestro ejemplo podemos partir de que el aprendizaje con los dos métodos de enseñanza es igual de efectivo, lo que nos conduce a la siguiente hipótesis nula $\mu s = \mu t$ (μs significa promedio de la clase con software, μt significa promedio con clase tradicional). Es decir, de ser cierta la hipótesis, no debe haber diferencia entre los puntajes promedio en el cuestionario de ambas poblaciones de estudiantes, solo la que pudiera ocasionarse por la aleatoriedad del muestreo. Si el investigador encuentra una diferencia positiva entre los puntajes promedio de los dos grupos, esto es, $\mu s - \mu t > 0$ (el promedio con software es mayor al promedio en la clase tradicional), es un resultado que apoya su hipótesis de investigación. En caso que la diferencia sea negativa, es decir, $\mu s - \mu t < 0$ (el promedio con software es menor al promedio en la clase tradicional), el resultado contradice la hipótesis nula.

Evaluación del capítulo

1. Explica brevemente porqué el muestreo no probabilístico no permite generalizar los datos obtenidos en una muestra hacia toda la población.

2. A continuación, se presentan diversas afirmaciones sobre conceptos que se involucran en el muestreo y la inferencia estadística. Selecciona las que consideres correctas:

- Si se seleccionan muestras de un tamaño dado de una misma población, los resultados suelen variar de una muestra a otra.
- Cuando se utilizan muestras para hacer inferencias sobre parámetros de una población, se genera inevitablemente un error conocido como error de muestreo.
- El error de muestreo está relacionado con el tamaño de muestra. A medida que se incrementa el tamaño de muestra se puede incrementar el error de muestreo.
- En una distribución muestral se tienen todos los valores que puede tomar un estadístico (por ejemplo, una proporción) en cada una de las muestras que es posible seleccionar de una población.
- La distribución muestral de algunos estadísticos (por ejemplo, la media) tiene una forma acampanada con centro en el parámetro de la población, sobre todo para tamaños de muestra superiores a 30.

3. La siguiente gráfica permite visualizar la relación entre el tamaño de muestra y el margen de error que se obtiene en un intervalo de confianza de 95% para una proporción. Se desea realizar una encuesta para conocer la opinión de los mexicanos sobre la legalización de la marihuana ¿de qué tamaño debe ser la muestra aleatoria a seleccionar, si se establece cómo un margen de error del 5% para la estimación del porcentaje de mexicanos que está de acuerdo con la legalización?

a) La muestra debe ser de 400 personas
b) La muestra debe ser de 1000 personas
c) La muestra debe ser de 500 personas

4. Selecciona las afirmaciones sobre pruebas de significación (PS) que son correctas.

- Las PS se proponen evaluar la evidencia de los datos de una muestra a favor de la hipótesis nula sobre el valor de un parámetro poblacional.
- La hipótesis nula se plantea en forma de igualdad o no diferencia con otros parámetros. Su rechazo implica que los resultados no se deben al azar sino a otras causas.
- La decisión sobre si el parámetro tiene determinado valor se toma con base en criterios probabilísticos, por lo que la decisión conlleva un componente de incertidumbre. Se puede rechazar la hipótesis nula siendo cierta y viceversa.
- Resultados alejados del valor del parámetro fijado en la hipótesis nula ponen en duda la validez de la hipótesis nula.
- La probabilidad de obtener un resultado o más alejado al valor obtenido, se conoce como *p valor*.
- El *p valor* mide la fuerza de la evidencia estadística en contra de la hipótesis. Entre más pequeño es el *p valor* más fuerte es la evidencia en contra de la hipótesis.
- La probabilidad de rechazar la hipótesis nula siendo cierta se conoce como nivel se significación, se representa por la letra .
- Un resultado es significativo si tiene poca probabilidad de ocurrir (p<). Resultados significativos siempre implican el rechazo de la hipótesis nula.

5. Lee la siguiente nota que publicó el periódico El Financiero. Explica a qué se refiere el enunciado "Este porcentaje representa un cambio estadísticamente significativo respecto al 72.9% registrado en diciembre de 2019".

CIFRAS DE LA ENSU DE INEGI

En México, el 68.1% cree que vivir en su ciudad es inseguro

Ese dato registra una baja de 4.8 por ciento respecto del dato de diciembre de 2019

DAVID SAÚL VELA
dvela@elfinanciero.com.mx

En diciembre pasado el 68.1 por ciento de los mexicanos (población de 18 años y más) consideró que vivir en su ciudad es inseguro, revela la Encuesta Nacional de Seguridad Pública Urbana (ENSU), realizada por el INEGI.

"Este porcentaje representa un cambio estadísticamente significativo respecto al 72.9 por ciento registrado en diciembre de 2019, pero no es estadísticamente diferente al 67.8 por ciento de septiembre de 2020", dice el estudio.

México; Coatzacoalcos, Veracruz; Cancún, Quintana Roo; Cuernavaca, Morelos y San Luis Potosí capital, con 94.8, 89.9, 88.9, 88.1, 87.7 y 87 por ciento, respectivamente.

En contraste, las ciudades con menor percepción de inseguridad fueron San Pedro Garza García, Nuevo León; Los Cabos, Baja California; Mérida, Yucatán; Saltillo, Coahuila; La Paz, Baja California Sur; y San Nicolás de los Garza, Nuevo León, con 11.7, 17.3, 24.6, 30.9, 31.8 y 31.8 ciento, respectivamente.

De acuerdo con la edición 29 del estudio, hubo 14 ciudades que tuvieron cambios estadísticamente significativos respecto a septiembre de 2020, seis de ellas tuvieron reducciones y ocho incrementaron.

OTRO DATO. En el segundo semestre de 2020, 14.3% fue víctima de acoso y/o violencia sexual. El porcentaje en mujeres fue de 21.6.

ciento, mientras que para el caso de los hombres fue de 62.7 por ciento.

A nivel nacional, 28.1 por ciento de los hogares contó con al menos una víctima de robo y/o extorsión en el segundo semestre de 2020.

Las ciudades con mayor porcentaje de hogares con al menos una víctima fueron: Iztapalapa, Ciudad de México; Atizapán, Estado de México; Tláhuac, Ciudad de México; Cuautitlán Izcalli, Estado de México, y Magdalena Contreras, Ciudad de México, con 47.1, 43.2, 42, 40.6 y 40.4 por ciento, respectivamente.

En cuanto a víctimas de actos de corrupción por parte de autoridades de seguridad pública, de julio a diciembre de 2020 se estima que 12 por ciento de

6. Se realizó un estudio para evaluar el efecto del medicamento *Remdesivir* en adultos estadounidenses enfermos de Covid-19. En un experimento aleatorizado se asignaron al azar 1,062 pacientes a dos grupos. El primero llamado grupo experimental conformado por 541 pacientes a los que se les inyectó 100 mg diarios de *Remdesivir* durante 10 días. El segundo grupo denominado grupo Placebo, conformado por 521 pacientes, se les administró un placebo (una inyección que no tiene el medicamento, pero el paciente no lo sabe). La variable de medición fue el tiempo de recuperación

de la enfermedad, el cual fue reportado de la siguiente manera:

Grupo	Tiempo promedio de recuperación	Intervalo de confianza (CI) para el tiempo de recuperación
Remdesivir (541 pacientes)	10 días	CI 95% (9 a 11 días)
Placebo (521 pacientes)	15 días	CI 95% (13 a 18 días)

- Teniendo en cuenta que el experimento utilizó una muestra de 1,062 pacientes ¿consideras que el *Remdesivir* tuvo un efecto positivo en el tiempo de recuperación de los pacientes enfermos de Covid-19?
- Los resultados obtenidos para esta muestra de pacientes, ¿podrían ser extendidos a toda la población de pacientes adultos enfermos de Covid-19?
- La diferencia en los tiempos de recuperación entre quienes tomaron *Remdesivir* y los del grupo Placebo, ¿podría haber ocurrido por azar o casualidad? ¿o en su lugar se debió al efecto del *Remdesivir*? Utiliza la segunda y última columna para argumentar.

Bibliografía recomendada

- Statistics and Probability in High School. Carmen Batanero y Manfred Borovnick. Sense Publishers. 2016. Primer edición.
- Caracterización del razonamiento estadístico de estudiantes universitarios acerca de las pruebas de hipótesis http://www.scielo.org.mx/pdf/relime/v16n2/v16n2a3.pdf
- Aproximación informal al contraste de hipótesis https://www.ugr.es/~batanero/documentos/Aproximacion.pdf
- La catadora de té. https://es.wikipedia.org/wiki/La_catadora_de_té
- Uso de gráficas y nociones de inferencia en conferencia de prensa sobre Covid-19 https://www.pscp.tv/w/1rmxPAOrprmKN?t=6m53s
- Construcción de significados sobre distribuciones muestrales y conceptos previos a la inferencia estadística https://www.redalyc.org/pdf/405/40516761006.pdf

Pensamiento estadístico
para docentes de bachillerato
de Santiago Inzunza Cázares
se terminó de editar en
el mes de octubre de 2023
en el Colegio de Bachilleres
del Estado de Sinaloa.

En las últimas tres décadas, los contenidos de estadística y probabilidad han ocupado espacios en los programas de estudio de matemáticas en todos los niveles educativos como ninguna otra temática; esto por la importancia de la estadística como herramienta metodológica y transversal a las demás ciencias, pero además por la importancia que está teniendo la alfabetización y el pensamiento estadístico en la sociedad actual, caracterizada por muchos expertos, como la sociedad de la información y del conocimiento.

Entre las causas principales que han generado la expansión de la estadística como campo de estudio y aplicación, destaca la revolución de los datos, impulsada por el creciente desarrollo de las tecnologías de la información y de las comunicaciones. Ello a su vez ha generado un fenómeno conocido como cuantificación o datificación de la sociedad, caracterizado por la necesidad de expresar y representar el comportamiento de diversos fenómenos en términos cuantitativos mediante gráficas, tablas, porcentajes, promedios, correlaciones, modelos, variabilidad, y otras medidas que definen el comportamiento de los datos provenientes de muestras, poblaciones y experimentos aleatorizados.