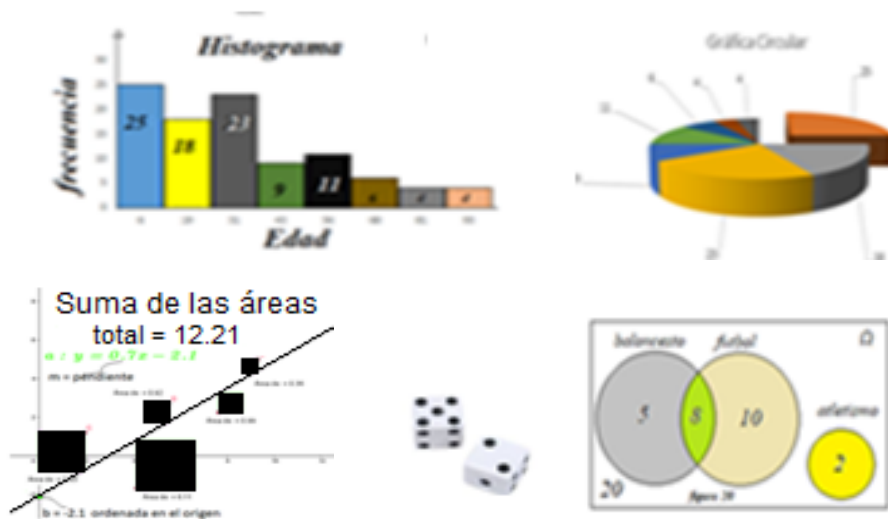




UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
COLEGIO DE CIENCIAS Y HUMANIDADES
Área de Matemáticas



Estadística y Probabilidad I

Cuaderno de Trabajo

PAE

Programa actualizado 2016

Hernández Hidalgo Concepción Julieta
Terrés Sandoval Armando
Valdez Monroy Julio César

Agosto 2019

Introducción

El contenido de este cuaderno de trabajo es una propuesta para el desarrollo de las actividades durante el curso del PAE (Programa de Apoyo al Egreso). Su finalidad es presentar de manera clara y sencilla cada uno de los temas que conforman la asignatura Estadística y Probabilidad I, del Programa de Estudio actualizado 2016 del Plan de Estudios del Colegio de Ciencias y Humanidades.

Este material permitirá al docente orientarse en la secuencia de los temas, brindándole herramientas y sugerencias para la impartición del curso, con la finalidad de contribuir al objetivo del PAE, que es conseguir la regularización académica del estudiante aunado al logro de aprendizajes significativos, bajo el enfoque didáctico que considera tres diferentes niveles de aprendizaje de la Estadística:

- Cultura estadística
- Razonamiento estadístico y
- Pensamiento estadístico

Destacando tres contribuciones al perfil de egreso propuesto en el Programa de Estudio actualizado 2016 de la asignatura:

- Conformación de su pensamiento estadístico sustentando la toma de decisiones sobre el comportamiento de diversos fenómenos, a partir de predicciones e inferencias estadísticas.
- Trascendencia a otras áreas del conocimiento, confiriéndole a su análisis un carácter de contribución a la validez científica.

Guía para su uso

En cada una de las tres Unidades que constituyen el curso, se presenta el tema central con sus respectivos subtemas revisando la parte teórica y conceptual. Posteriormente, se explican ejemplos, paso por paso, finalizando con ejercicios propuestos para reforzar los conocimientos adquiridos y desarrollar el pensamiento crítico al analizar y describir los resultados obtenidos.

En la parte final de este material se tiene una serie de referencias para que el estudiante consulte temas en los que desee profundizar; a los docentes les servirá para enriquecer o incrementar la complejidad de los ejercicios y contenidos abarcados en el cuaderno de trabajo.

Propósito

El cuaderno de trabajo Estadística y Probabilidad I, tiene como principal propósito ser un material accesible para el estudiante y de manejo sencillo para el profesor durante el desarrollo de las diez sesiones del curso PAE. En su diseño se consideraron las características de los alumnos asistentes, el tiempo disponible de cuatro horas en cada sesión y las actividades que se pueden realizar extra clase para reforzar los aprendizajes. El manejo del contenido propuesto queda abierto a la experiencia del docente en este tipo de cursos.

Estrategias de aprendizaje

El docente tendrá el papel de guiar durante las sesiones del curso a los estudiantes para promover su aprendizaje. Relacionará los conocimientos teóricos con los procedimientos necesarios para la solución de los problemas planteados, para el posterior análisis e interpretación de los resultados obtenidos.

Las actividades se podrán trabajar en equipo, a fin de promover valores como la tolerancia y respeto; además de desarrollar habilidades propias del trabajo colaborativo al enriquecer las ideas con la opinión de sus compañeros, para llegar a la solución de problemas mediante diversos puntos de vista que serán valorados para tomar la mejor decisión.

Formas de evaluación

Con la finalidad de que los temas revisados en cada sección se hayan comprendido, posterior a la parte teórica se presentan ejemplos resueltos que guían al estudiante, acompañado por el profesor, en el procedimiento de solución. Posteriormente se plantean ejercicios parcialmente resueltos que deberán ser completados por los alumnos de forma individual o por equipos. Al finalizar cada sección, se proponen ejercicios como evaluación final, cuyos resultados se revisarán en plenaria para resolver las dudas y consolidar los aprendizajes deseados.

Bibliografía

En esta sección el estudiante contará con un listado de referencias que le permitirá profundizar en los temas revisados. Estos recursos se encuentran disponibles físicamente en la biblioteca del plantel y/o en formato digital entre los recursos proporcionados por la UNAM en su sistema de Bibliotecas en línea. Al docente le servirá como apoyo en la mejora de su práctica, para enriquecer o incrementar la complejidad de los ejercicios que contiene este cuaderno de trabajo.

Índice

UNIDAD 1. Obtención, descripción e interpretación de información estadística

1.1.	Conceptos básicos _____	1
1.2.	Ciclo de investigación estadística _____	5
1.3.	Representación tabular y gráfica de una variable cualitativa _____	6
1.3.1.	Gráfico de barras _____	7
1.3.2.	Gráfico de sectores _____	8
1.4.	Representación tabular y gráfica de una variable cuantitativa _____	9
1.4.1.	Tabla de distribución de frecuencias _____	10
1.4.2.	Histograma _____	12
1.4.3.	Análisis de un histograma _____	13
1.4.4.	Polígono de frecuencias _____	15
1.4.5.	Polígono de frecuencias acumuladas _____	17
1.4.6.	Polígono de frecuencias relativas y relativas acumuladas _____	19
1.5.	Representación tabular y gráfica de una variable cuantitativa puntual _____	22
1.6.	Medidas de tendencia central para datos no agrupados _____	25
1.6.1	Moda _____	25
1.6.2.	Mediana _____	25
1.6.3	Media _____	25
1.6.4.	Comparación entre la media y la mediana _____	28
1.7.	Medidas de tendencia central para datos agrupados _____	29
1.7.1	Media _____	30
1.7.2.	Mediana _____	31
1.7.3.	Moda _____	32
1.8.	Medidas de posición y de dispersión _____	34
1.8.1.	Rango _____	35
1.8.2.	Cuartiles _____	37
1.8.3.	Diagrama de caja _____	38
1.8.4.	Rango intercuartil _____	39
1.8.5.	Varianza _____	40
1.8.6.	Desviación estándar _____	42
1.9.	Varianza y desviación estándar para datos agrupados _____	42
1.9.1.	Varianza _____	43
1.9.2.	Desviación estándar _____	44
1.9.3.	Propiedades de la desviación estándar _____	45
	Evaluación de la Unidad 1 _____	46

UNIDAD 2. Obtención, descripción e interpretación de información estadística para datos bivariados

2.1.	Introducción _____	51
2.2.	Asociación entre dos variables cualitativas _____	52
2.2.1.	Tablas de contingencia _____	52
2.2.2.	Tablas de frecuencias relativas _____	54
2.2.3.	Tablas de porcentajes por fila _____	55

Evaluación tema 1 _____ 56

2.3.	Variables cuantitativas _____	58
2.3.1.	Diagrama de dispersión _____	58
2.3.2.	Coeficiente de correlación _____	60
2.3.3.	Mínimos cuadrados _____	63
2.3.4.	Regresión lineal _____	67

Evaluación tema 2 _____ 68

UNIDAD 3. Probabilidad Azar: modelación y toma de decisiones.

3.1.	Fenómenos deterministas y aleatorios _____	71
3.2.	Espacio muestral y diferentes tipos de eventos _____	72
3.3.	Enfoques de probabilidad _____	74
3.4.	Cálculo de probabilidades de eventos simples y compuestos _____	76
3.5.	Probabilidad Condicional y eventos independientes _____	97

Propuesta de evaluación para la Unidad 3 _____ 101

Bibliografía _____ 103

UNIDAD 1. OBTENCIÓN, DESCRIPCIÓN E INTERPRETACIÓN DE INFORMACIÓN ESTADÍSTICA

Presentación

En esta unidad conocerás los principales conceptos de la estadística que serán utilizados a lo largo del cuaderno de trabajo. Aprenderás a organizar, describir y analizar un conjunto de datos mediante representaciones tabulares y gráficas, así como a través de resúmenes numéricos. Estas herramientas te permitirán conocer el comportamiento de diversos conjuntos de datos, lo que a su vez fomentará el desarrollo de tu pensamiento estadístico.

Propósito

Al terminar esta unidad, realizarás inferencias informales acerca del comportamiento de una característica de interés en una población definida dentro de su entorno, a partir del análisis de su tendencia, variabilidad y distribución, en una muestra obtenida de dicha población, para contribuir a la formación de su pensamiento estadístico.

1.1. Conceptos básicos

En términos simples, la estadística trata sobre los métodos de recolección, organización y análisis de datos para la toma de decisiones, en donde el contexto y la variabilidad juegan un papel fundamental. La Estadística puede ser estudiada desde dos perspectivas:

- **Estadística descriptiva.** Las afirmaciones o resultados que se obtienen de analizar un conjunto de datos se refieren sólo a este conjunto.
- **Estadística inferencial.** Las afirmaciones o resultados que se obtienen son acerca de la población de la cual proviene el conjunto de datos analizado.

En la actualidad, la estadística permea desde las grandes disciplinas (política, científica, social), hasta las actividades más cotidianas (ir de compras, ver televisión, leer el periódico, navegar en internet, etc.). De esta manera, se consideran tres las razones para incluir la enseñanza de la estadística en la escuela: su papel en el desarrollo del razonamiento crítico, como instrumento de análisis en otras disciplinas, y su rol en la planeación y toma de decisiones en otras áreas profesionales.

Para comenzar a familiarizarte con la estadística, es necesario definir algunos conceptos que se utilizarán a lo largo del cuaderno de trabajo:

- **Población.** Es el conjunto de todos los individuos u objetos (de aquí en adelante sólo llamados individuos), cuyas características se han de analizar.
- **Variable.** Es una característica de la población que es común a todos los individuos. Puede tomar valores diferentes para distintos individuos. Dentro de las variables se puede establecer la siguiente clasificación:

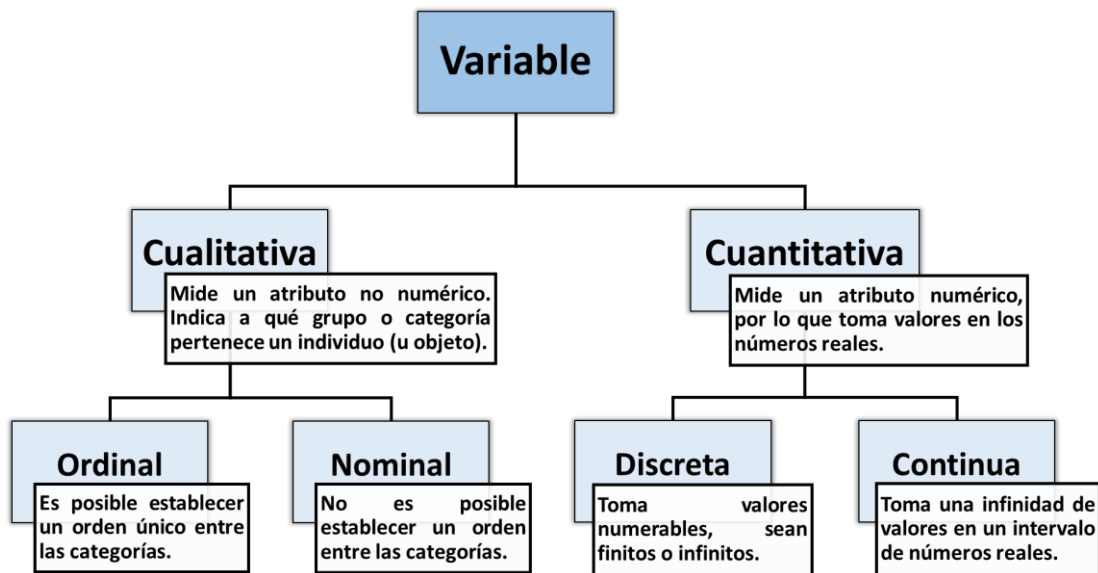


Figura 1.1. Clasificación de las variables.

- **Muestra.** Es un subconjunto de la población. En la práctica, este subconjunto es el que realmente se analiza para obtener información sobre la población. Se pueden distinguir dos tipos de muestras: **Muestra aleatoria (representativa)**, la cual refleja en la medida de lo posible las características de la población de la cual proviene; y **Muestra no aleatoria (sesgada)**, en la cual conjuntos importantes de la población no están representados. Además de la aleatoriedad, la representatividad de la muestra también depende de su tamaño. En este sentido, si la variabilidad de la población es grande, se requiere de una muestra grande; si la variabilidad es poca, con una muestra pequeña basta.

El proceso mediante el cual se obtiene una muestra se conoce como **muestreo**, en el cual se pueden utilizar diversos los instrumentos para la recolección de datos: encuestas, experimentos, la simple observación, etc. Incluso, hay bases de datos proporcionadas

por instituciones nacionales (INEGI) e internacionales (OMS) que pueden ser consultadas en la red.

Esquemáticamente, los conceptos mencionados se pueden representar de la siguiente manera:

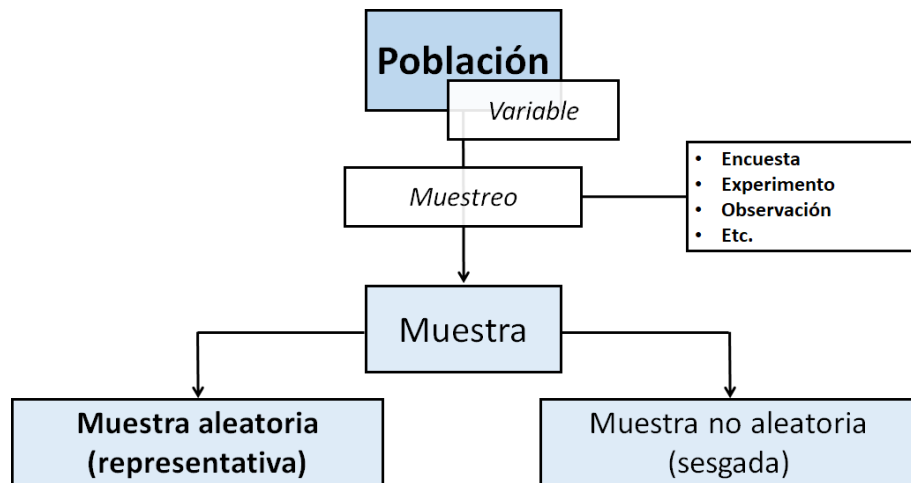


Figura 1.2. Conceptos estadísticos básicos.

Otro concepto fundamental en estadística es el de **distribución**, el cual describe el comportamiento de una variable asociada a un conjunto de datos; consiste de todos los valores diferentes que toma la variable, e incluye las frecuencias con las que se repite cada valor.

Con la finalidad de consolidar los conceptos anteriores, considera los siguientes ejemplos:

Ejemplo 1.1

En cada una de las siguientes situaciones, menciona cuál es la población y cuál la muestra, y si esta última es representativa o no.

- Se seleccionan a las primeras 100 alumnas que llegan al colegio, con el propósito de conocer la estatura promedio de todas las alumnas del plantel.
- Se quiere conocer sobre los hábitos de consumo de tabaco de los alumnos del plantel, con la finalidad de tomar medidas preventivas al respecto. Para ello, se elige a cinco alumnos al azar y se les pregunta acerca de sus hábitos de consumo.

Solución

- a) La población está constituida por todas las alumnas del colegio y la muestra se compone por las primeras 100 alumnas que llegan al plantel. La forma en cómo los alumnos llegan al colegio es aleatoria, por lo que las primeras 100 alumnas pueden ser de cualquier semestre. Por lo tanto, la muestra es representativa.
- b) La población son todos los alumnos del colegio y la muestra los cinco alumnos seleccionados. Como los alumnos son elegidos al azar, estos pueden ser, o no, fumadores. No obstante, el tamaño de la muestra es muy pequeño, lo que podría arrojar como resultado que ninguno sea fumador o que los cinco lo sean. Por lo tanto, la muestra no es representativa.

Ejemplo 1.2

Los datos sobre estudios médicos contienen valores de muchas variables para cada uno de los sujetos de estudio. De las siguientes variables, ¿cuáles son cualitativas y cuáles cuantitativas? ¿Cuáles son ordinales o nominales? ¿Cuáles discretas o continuas?

- a) Género
- b) Edad
- c) Fumador
- d) Presión sanguínea (en milímetros de mercurio)
- e) Concentración de calcio en la sangre (en microgramos por litro)

Solución

- a) La variable género toma los valores 'hombre' o 'mujer', los cuales son categorías entre las que no hay un orden. Por lo tanto, es una variable cualitativa nominal.
- b) Los valores que toma la variable edad son números, normalmente contados en años (0, 1, 2, 3... etc.), por lo que la variable es cuantitativa discreta.
- c) Fumador toma los valores 'sí' o 'no', por lo que es una variable cualitativa nominal, ya que no es posible establecer un orden entre estas categorías.
- d) La presión sanguínea es una medida que toma valores numéricos en los reales, por lo que es cuantitativa continua.
- e) La concentración de calcio en la sangre también es una variable cuantitativa continua, ya que toma valores en los números reales.



Ejercicios 1.1

1. Se quiere hacer un estudio comparativo entre los países que integran la región de Norteamérica (México, Estados Unidos y Canadá). Menciona una variable que describa alguna característica de estos países que sea del tipo:

- a) Ordinal
- b) Nominal
- c) Discreta
- d) Continua

2. Plantea una situación en la que identifiques a la población y a la muestra, de tal manera que esta última:

- a) No sea representativa.
- b) Sea representativa.

1.2. Ciclo de investigación estadística

Una vez definidos los principales conceptos que serán utilizados a lo largo del cuaderno de trabajo, es necesario conocer cómo se lleva a cabo un estudio estadístico. En el siguiente diagrama aparece un modelo de pensamiento en el que se muestran los procesos involucrados al resolver problemas reales utilizando la estadística.

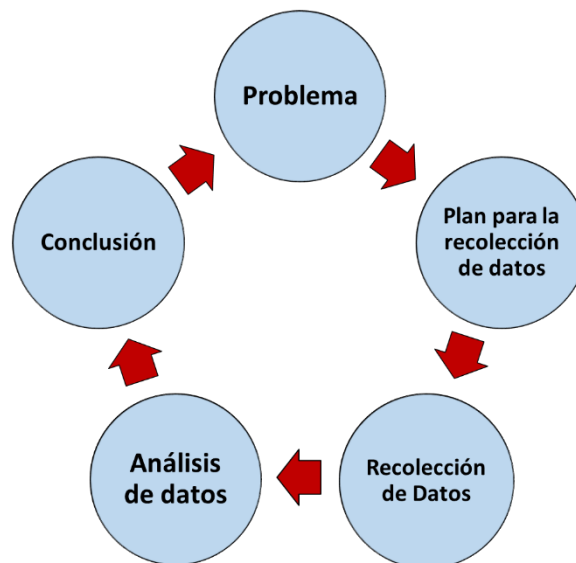


Figura 1.3. Ciclo de investigación estadística.

El ciclo comienza con un problema que debe ser refinado y planteado en términos estadísticos. Enseguida, se desarrolla una estrategia para determinar la variable a

investigar, qué datos proporcionarán información sobre dicha variable, de qué forma obtenerlos o generarlos, y diseñar un plan para su análisis. Una vez recolectados los datos, estos deben ser organizados y depurados para su análisis. El análisis puede hacerse a un nivel exploratorio (que es en lo que nos enfocaremos en este cuaderno de trabajo) para generar una hipótesis, o puede hacerse con la finalidad de hacer inferencias sobre la población de la cual se obtuvieron los datos. Finalmente, los resultados son interpretados en términos del contexto para generar una posible respuesta al problema. Si se considera pertinente, el ciclo se repite, pero teniendo en consideración los resultados obtenidos.

Para facilitar el análisis de los datos, un primer paso es organizarlos tanto de forma tabular como gráfica. Dependiendo del tipo de variable, será el tipo de representación que se utilizará.

1.3. Representación tabular y gráfica de una variable cualitativa

Para representar de forma tabular una variable cualitativa, se colocan en una columna cada una de las categorías que toma la variable. Después, en una segunda columna, se ubica la frecuencia con la que los datos caen en cada categoría. Para ejemplificar esta forma de representación considera el siguiente ejemplo:

Ejemplo 1.3: Redes sociales

Con el crecimiento acelerado de la tecnología, una de las principales actividades entre los jóvenes es interactuar con otras personas a través de las redes sociales. ¿Cuál crees que es la red social más utilizada en México?

Leticia leyó un artículo en el que se mencionaba a Facebook como la red social más utilizada en México en el 2018, seguida por YouTube, Instagram, Twitter, Google+, Pinterest, LinkedIn y Taringa. Para corroborar esta información, Leticia decidió hacer un estudio en su grupo de estadística, para lo cual primero se planteó las siguientes preguntas: ¿Cuál es la variable de estudio? ¿Qué tipo de variable es? ¿Cómo recolectar los datos? ¿Cómo organizarlos? ¿Cómo analizarlos?

Como lo que interesa es conocer la red social más utilizada, para cada individuo esta puede ser cualquiera de las redes sociales existentes, entre las cuales no hay un orden definido, por lo que la variable de estudio es cualitativa nominal. Para recolectar los

datos, Leticia decidió hacer una encuesta rápida, en la cual sólo incluyó la pregunta ‘¿Cuál es la red social que más utilizas?’, y como opciones sólo las redes sociales mencionadas en el artículo. Una vez aplicada la encuesta, Leticia decidió organizar los datos en una tabla como la que se muestra a continuación:

Tabla 1.1. Distribución sobre la preferencia de las redes sociales en un grupo de 40 alumnos.

Categoría	Frecuencia <i>f</i>
Facebook	11
Youtube	10
Instagram	6
Twiter	4
Pinterest	3
Google+	3
Taringa	2
LinkedIn	1

$$\sum f = 40$$

De acuerdo con los datos de la Tabla 1.1., la red social preferida en el grupo de Leticia es Facebook, ya que tiene la mayor frecuencia (11 alumnos), lo cual coincide con lo publicado en el artículo. Después le siguen YouTube (10), Instagram (6), Twitter (4), Pinterest (3), Google+ (3), Taringa (2) y LinkedIn (1).

Una forma más clara y rápida de analizar el conjunto de datos, y obtener el resultado anterior, es mediante el uso de representaciones gráficas. En el caso de una variable cualitativa, como la del ejemplo, el gráfico de barras es el más utilizado.

1.3.1. Gráfico de barras

En este tipo de gráfico, cada categoría se representa por una barra, cuya altura corresponde a la frecuencia de la categoría en cuestión. Si se trata de una variable nominal, el orden en el que se presenten las categorías en el gráfico no es relevante. De esta manera, el gráfico de barras que representa a los datos de la Tabla 1.1 es el que se muestra a continuación:

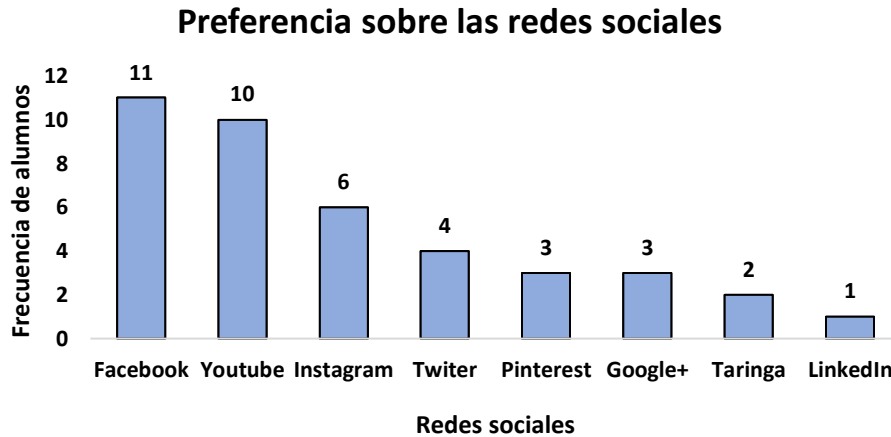


Gráfico 1.1. Gráfico de barras sobre las redes sociales preferidas en un grupo de 40 alumnos.

Otro gráfico utilizado para analizar el comportamiento de variables cualitativas es el de sectores (pastel).

1.3.2. Gráfico de sectores

En este tipo de gráfico, el total de datos se representa mediante un círculo, el cual se divide en un número de sectores igual a la cantidad de categorías. Cada categoría corresponde a un sector, cuya amplitud se determina por la frecuencia de la categoría.

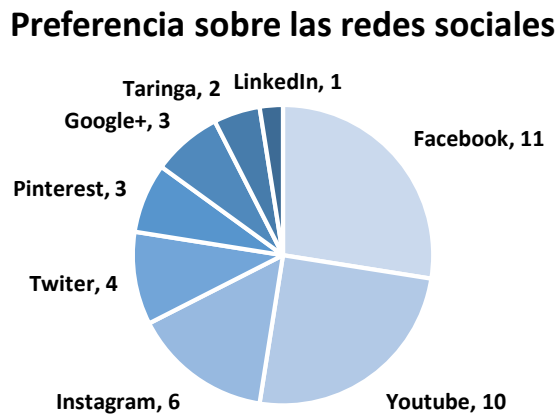


Gráfico 1.2. Gráfico de sectores sobre las redes sociales preferidas en un grupo de 40 alumnos.

Tanto el gráfico de barras como el de sectores pueden construirse cambiando las frecuencias a porcentajes, los cuales se obtienen al dividir la frecuencia de cada categoría entre el total de datos y multiplicando el resultado por 100. Al final, los gráficos construidos con esta nueva escala lucirán como los Gráficos 1.1 y 1.2.

1.4. Representación tabular y gráfica de una variable cuantitativa

Cuando se trata de una variable cuantitativa, la tabla de distribución de frecuencias, el histograma y los polígonos de frecuencias son las representaciones que se utilizan para analizar este tipo de variable. Asimismo, la manera de construir estas representaciones diferirá si la variabilidad en los datos es poca o mucha. Para ejemplificar este último caso, considera la siguiente situación.

Ejemplo 1.4: Duración de la batería del celular

Sofía está planeando comprar un teléfono celular. Entre sus opciones están las marcas Ultraphone y Superphone. No obstante, ha observado que uno de los principales problemas que tienen sus compañeros es con la duración de la batería, por lo que tiene pensado elegir la marca que le ofrezca un mayor tiempo de duración.

- ¿De qué manera Sofía puede determinar qué marca de celular ofrece una mayor duración de la batería?
- Sofía le pide a cinco de sus amigos que tienen un celular de la marca Ultraphone, y a otros cinco que poseen uno de la marca Superphone, que registren el tiempo (en horas) que dura la batería de sus celulares con una sola carga durante 10 días. Transcurrido este tiempo, sus compañeros le proporcionan los datos que se muestran en las Tablas 1.2 y 1.3. A partir de esta información, ¿qué celular ofrece el mayor tiempo de duración de la batería?

Tabla 1.2. Tiempo de duración de la batería Ultraphone (horas).

5.0	33.5	15.0	41.0	18.5	14.0
9.5	13.5	5.5	13.5	30.5	13.0
10.5	15.5	4.5	17.0	13.0	4.5
19.5	14.5	10.5	9.0	14.0	18.0
23.0	18.0	7.5	10.5	15.5	22.0
10.0	13.5	9.0	23.0	30.5	15.5
15.5	24.5	25.5	15.0	17.0	25.0
11.0	18.0	27.0	12.5	8.0	28.0
12.5	27.0				

Tabla 1.3. Tiempo de duración de la batería Superphone (horas).

27.0	27.5	31.5	19.5	23.5	32.5
18.5	20.0	23.0	25.5	23.5	29.0
29.5	31.0	27.5	24.5	33.0	16.0
22.5	24.5	30.0	21.0	23.0	35.0
26.5	29.5	23.0	26.0	24.0	7.5
27.5	15.0	27.0	27.0	24.0	30.0
23.5	23.5	19.5	22.5	20.5	26.0
19.0	32.0	27.0	24.5	27.5	25.5
29.5	19.5				

- c) Una amiga de Sofía, quien tiene los dos tipos de celular, le aconseja comprar el Superphone, ya que la batería dura dos veces más que la del Ultraphone. De acuerdo con los datos de las Tablas 1.2 y 1.3, ¿debe Sofía seguir el consejo de su amiga? Justifica tu respuesta.

Para facilitar el análisis de los datos, un primer paso es organizarlos tanto de forma tabular como gráfica.

1.4.1. Tabla de distribución de frecuencias

En el Ejemplo 1.4, la variable de estudio es el tiempo que dura la batería de cada marca de celular, medido en horas, con una sola carga. Como se puede observar en la Figura 1.1, esta variable es cuantitativa continua, ya que puede tomar un número infinito de valores diferentes en los reales. Al trabajar con este tipo de variable, es conveniente agrupar los datos en clases (o intervalos), las cuales se obtienen al dividir el rango de la variable en el número de clases deseado. Para ejemplificar esto, considera los datos ordenados de la Tabla 1.4 que corresponden al tiempo de duración de la batería del celular Ultraphone.

Tabla 1.4. Datos ordenados del tiempo de duración de la batería del celular Ultraphone.

4.5	4.5	5.0	5.5	7.5	8.0
9.0	9.0	9.5	10.0	10.5	10.5
10.5	11.0	12.5	12.5	13.0	13.0
13.5	13.5	13.5	14.0	14.0	14.5
15.0	15.0	15.5	15.5	15.5	15.5
17.0	17.0	18.0	18.0	18.0	18.5
19.5	22.0	23.0	23.0	24.5	25.0
25.5	27.0	27.0	28.0	30.5	30.5
33.5	41.0				

El rango se define como la diferencia entre el valor máximo ($x_{máx}$) y el valor mínimo ($x_{mín}$). En la Tabla 1.4, el tiempo mínimo de duración de la batería del celular Ultraphone es $x_{mín} = 4.5$ horas, y el tiempo máximo es $x_{máx} = 41$ horas. De esta manera, el rango se determina de la siguiente forma:

- $Rango = x_{máx} - x_{mín} = 41 - 4.5 = 36.5$.

El número de clases se calcula mediante la regla de *Sturges*:

$$\text{Número de clases} = 1 + 3.32\log(n),$$

donde n es el número de datos, que en el ejemplo es de 50. Así, el número de clases es:

- $\text{Número de clases} = 1 + 3.32\log(50) = 6.64 \approx 7$.

Una vez determinado el número de clases, lo siguiente es calcular su amplitud:

- $\text{Amplitud} = \frac{\text{Rango}}{\text{Número de intervalos}} = \frac{36.5}{7} = 5.21 \approx 5.5$.

Si la parte decimal de la amplitud es mayor que cero y menor que .5, la amplitud se redondea a .5; si es mayor que .5, se redondea al entero siguiente.

Debido al redondeo de la amplitud, es necesario hacer un ajuste para determinar en qué valor inicia la primera clase y en cuál termina la séptima. Para ello, se calcula en cuánto se ha excedido el rango, se divide el exceso entre dos, y una parte se resta al valor mínimo y la otra se suma al valor máximo:

- $\text{Rango ampliado} = \text{Número de clases} \times \text{Amplitud} = (7)(5.5) = 38.5$;
- $\frac{\text{Rango ampliado} - \text{Rango}}{2} = \frac{38.5 - 36.5}{2} = \frac{2}{2} = 1$;
- $x_{\min} - 1 = 4.5 - 1 = 3.5$;
- $x_{\max} + 1 = 41 + 1 = 42$.

De esta manera, nuestra primera clase inicia en 3.5 horas y la séptima termina en 42 horas. El límite superior de cada clase se obtendrá al ir sumando la amplitud (5.5 horas) al límite inferior correspondiente:

$$L_{inf} + \text{Amplitud} = L_{sup}.$$

Determinadas las clases, lo que sigue es contar la frecuencia de los valores que se ubican en cada clase (f_i). Por convención, si algún valor coincide con el límite superior de una clase, dicho valor se incluye en esta clase.

La Tabla 1.5 muestra cómo se distribuye el tiempo de duración de la batería del celular Ultraphone; poco más de la mitad de las veces, el tiempo de duración de la batería estuvo entre 9.0 y 20.0 horas (16 entre 9 y 14.5 horas, y 13 entre 14.5 y 20.0 horas). Fueron pocas las veces en las que el tiempo de duración superó estos valores, siendo poco común que la batería durara más de 31 horas.

Tabla 1.5. Distribución de frecuencias del tiempo de duración de la batería del celular Ultraphone.

Clase $L_{inf} - L_{sup}$	Frecuencia f_i
3.5 – 9.0	8
9.0 – 14.5	16
14.5 – 20.0	13
20.0 – 25.5	6
25.5 – 31.0	5
31.0 – 36.5	1
36.5 – 42.0	1

$$\sum f_i = 50$$

A pesar de que la Tabla 1.5 es ilustrativa, una representación gráfica proporciona una idea más clara acerca de cómo es el comportamiento de la variable.

1.4.2. Histograma

Un histograma es un gráfico que muestra cómo se distribuyen los datos entre las clases. A cada clase le corresponde una barra; cada barra tiene una altura igual a la frecuencia de la clase y una base igual a la amplitud. Las barras se tocan una a otra para indicar que la variable es continua.

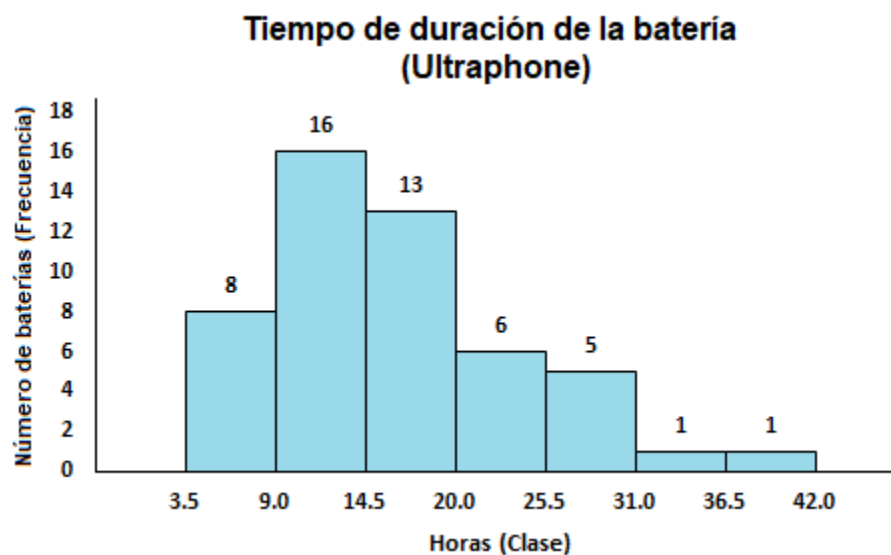


Gráfico 1.3. Histograma del tiempo de duración de la batería del celular Ultraphone.

En el Gráfico 1.3 se muestra el histograma correspondiente al tiempo de duración de la batería del celular Ultraphone. En el caso del celular Superphone, su histograma es el que se expone a continuación:

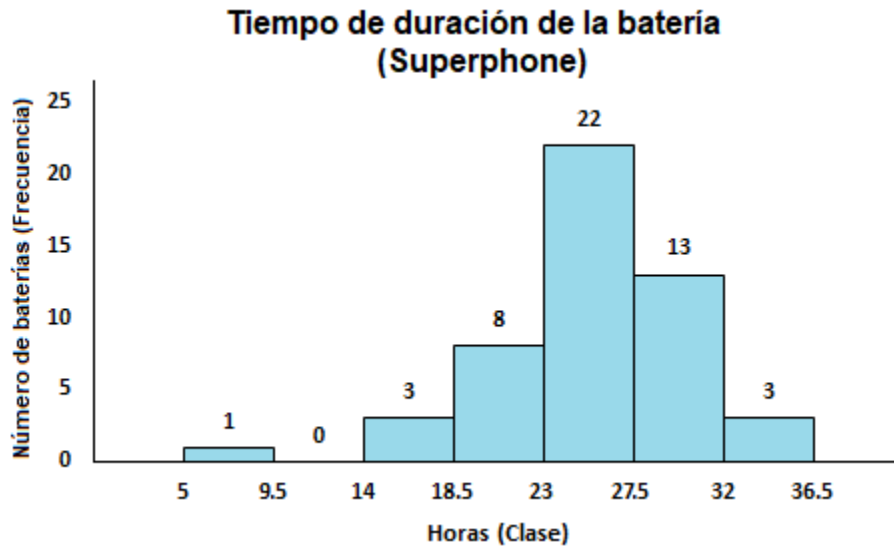


Gráfico 1.4. Histograma del tiempo de duración de la batería del celular Superphone.

1.4.3. Análisis de un histograma

Para el análisis del aspecto general de un histograma, considera lo siguiente:

- **Centro.** Es el valor alrededor del cual se agrupan los datos. En el caso del celular Ultraphone (Gráfico 1.3), los datos se agrupan, aproximadamente, alrededor de 15 horas, mientras que en el Superphone (Gráfico 1.4) es alrededor de 25 horas.
- **Forma.** Se refiere a la simetría del histograma y a si tiene picos. Sobre la simetría, un histograma es simétrico (a), o aproximadamente simétrico (b), si el lado derecho, a partir del centro, es igual o similar al lado izquierdo, respectivamente.



Gráfico 5. Histograma simétrico (a) y aproximadamente simétrico (b).

Un histograma es asimétrico hacia la derecha (d), si su lado derecho se extiende mucho más lejos que el lado izquierdo; y es asimétrico hacia la izquierda (c), si el lado izquierdo se extiende mucho más allá que el lado derecho.

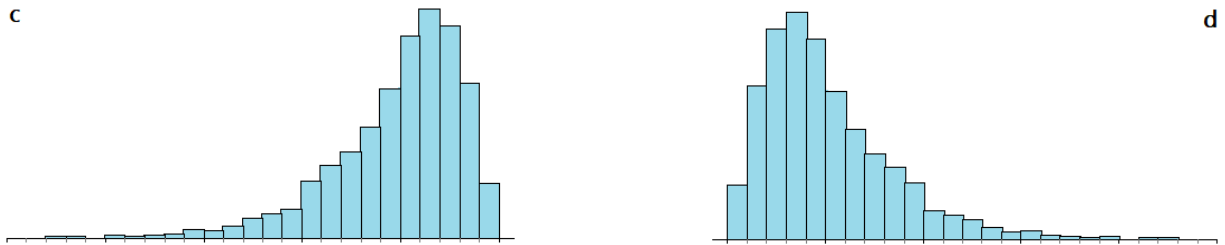


Gráfico 1.6. Histogramas asimétrico hacia la izquierda (c) y asimétrico hacia la derecha (d).

Los picos en un histograma se presentan cuando la frecuencia de una o varias clases es la mayor.

En el caso del celular Ultraphone (Gráfico 1.3), se puede considerar que la distribución es asimétrica hacia la derecha, mientras que la distribución del Superphone (Gráfico 1.4) es asimétrica hacia la izquierda. Respecto a los picos, ambas distribuciones presentan uno en las clases (9 – 14.5] y (23 – 27.5], respectivamente. Cabe señalar que cuando son pocos los datos resulta difícil determinar la forma de una distribución. Ésta se aprecia mejor cuando el número de datos es relativamente grande.

- **Dispersión.** Se refiere a qué tan separados se encuentran los datos respecto al centro. En el Gráfico 1.3 los datos están más dispersos, ya que van de 3.5 a 42 horas, mientras que en el Gráfico 1.4 van de 5 a 36.5 horas (Más adelante se verá cómo hacer un análisis más preciso de la dispersión).
- **Desviaciones y huecos.** Un caso importante de desviación son las observaciones atípicas. Es decir, una observación individual que queda fuera del aspecto general del histograma. Los huecos se presentan cuando alguna clase tiene frecuencia cero. Sólo en el Gráfico 1.4 se presenta un hueco en la clase (9.5 – 14]. El valor que cae en la clase (5 – 9.5] no se puede considerar como atípico, pues no se aleja tanto de donde se agrupa la mayoría de los datos (con base en el rango intercuartil, el cual se verá más adelante, se puede determinar cuándo un valor puede considerarse como atípico).

A partir del análisis anterior, se puede observar que, en promedio, el tiempo de duración de la batería del celular Superphone es mayor que el de la batería del Ultraphone. En el primero, los datos se agrupan alrededor de 25 horas, mientras que en el segundo es alrededor de 15 horas. No obstante, la afirmación de la amiga de Sofía, de que la batería del Superphone dura el doble, no es adecuada, ya que el histograma correspondiente tendría que estar centrado aproximadamente alrededor de 30 horas.

Una forma más clara de observar la diferencia entre los conjuntos de datos es compararlos en un solo gráfico. Sin embargo, no es factible superponer un histograma sobre otro. Para ello, los polígonos de frecuencias son más adecuados.

1.4.4. Polígono de frecuencias

El polígono de frecuencias se construye a partir del histograma, uniendo con líneas rectas los puntos medios de las barras por su parte superior. A estos puntos medios se les llama marcas de clase (y_i), las cuales se obtienen al sumar los límites superior e inferior, y dividir el resultado entre dos:

$$\text{Marca de clase } (y_i) = \frac{L_{inf} + L_{sup}}{2}.$$

Tabla 1.6. Distribución de frecuencias del tiempo de duración de la batería del celular Ultraphone (Marca de clase).

Clase $L_{inf} - L_{sup}$	Marca de Clase y_i	Frecuencia f_i
3.5 – 9.0	6.25	8
9.0 – 14.5	11.75	16
14.5 – 20.0	17.25	13
20.0 – 25.5	22.75	6
25.5 – 31.0	28.25	5
31.0 – 36.5	33.75	1
36.5 – 42.0	39.25	1

$$\sum f_i = 50$$

El polígono de frecuencias correspondiente al tiempo de duración de la batería del celular Ultraphone se muestra a continuación:

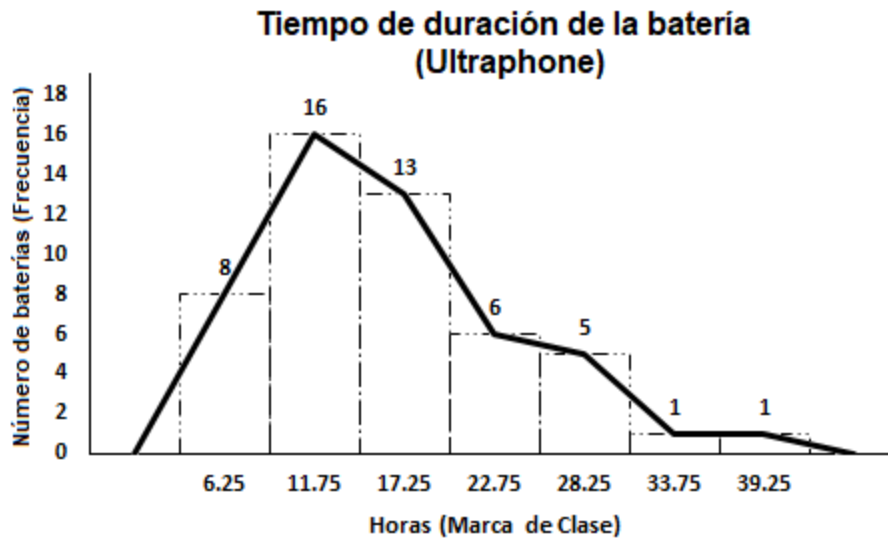


Gráfico 1.7. Polígono de frecuencias del tiempo de duración de la batería del celular Ultraphone.

En el caso del tiempo de duración de la batería del celular Superphone, el polígono de frecuencias es el siguiente:

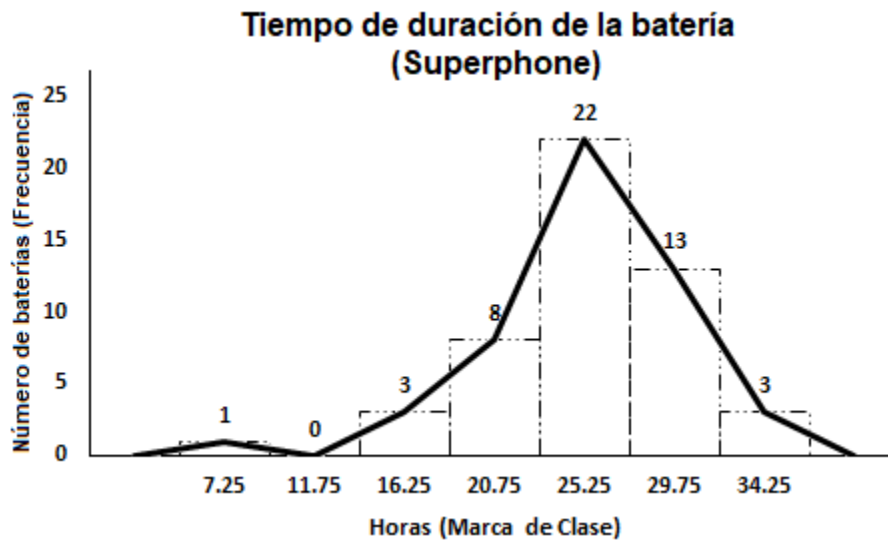


Gráfico 1.8. Polígono de frecuencias del tiempo de duración de la batería del celular Superphone.

Como ya se mencionó, una ventaja de los polígonos de frecuencias es que permiten comparar varias distribuciones en un solo gráfico. En el Gráfico 1.9 se comparan los tiempos de duración de las baterías de los celulares Ultraphone y Superphone. Se puede observar cómo el polígono de frecuencias de este último se recarga hacia valores más altos, lo que indica un mayor tiempo de duración.

Tiempo de duración de las baterías de los celulares Ultraphone y Superphone

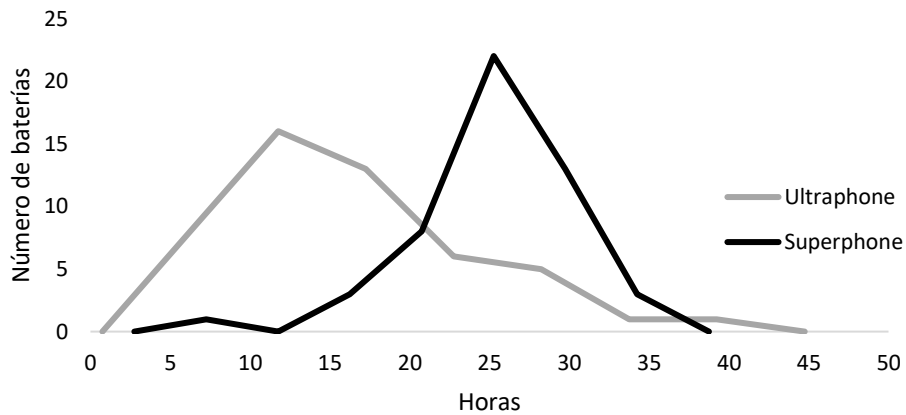


Gráfico 1.9. Polígonos de frecuencias del tiempo de duración de las baterías de los celulares Ultraphone y Superphone.

1.4.5. Polígono de frecuencias acumuladas (Ojiva)

Otra representación gráfica que también resulta útil es el polígono de frecuencias acumuladas. La frecuencia acumulada (F_i) se obtiene al sumar a la frecuencia de la clase, la frecuencia de las clases previas.

Tabla 1.7. Distribución de frecuencias del tiempo de duración de la batería del celular Ultraphone (Frecuencia acumulada).

Clase $L_{inf} - L_{sup}$	Marca de Clase y_i	Frecuencia f_i	Frecuencia acumulada F_i
3.5 – 9.0	6.25	8	8
9.0 – 14.5	11.75	16	16 + 8 = 24
14.5 – 20.0	17.25	13	13 + 24 = 37
20.0 – 25.5	22.75	6	6 + 37 = 43
25.5 – 31.0	28.25	5	5 + 43 = 48
31.0 – 36.5	33.75	1	1 + 48 = 49
36.5 – 42.0	39.25	1	1 + 49 = 50

$$\sum f_i = 50$$

El polígono de frecuencias acumuladas se construye uniendo con líneas rectas las frecuencias acumuladas sobre los límites superiores de los intervalos (L_{sup}). En el Gráfico 1.10 se muestra el polígono de frecuencias acumuladas correspondiente al tiempo de duración de la batería del celular Ultraphone:

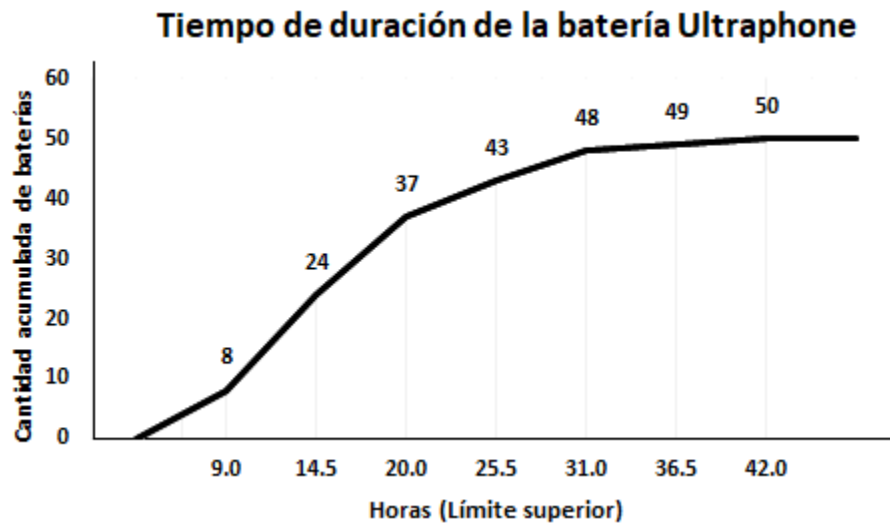


Gráfico 1.10. Polígono de frecuencias acumuladas del tiempo de duración de la batería del Ultraphone. En el caso del tiempo de duración de la batería del celular Superphone, el polígono de frecuencias acumuladas correspondiente se muestra a continuación:

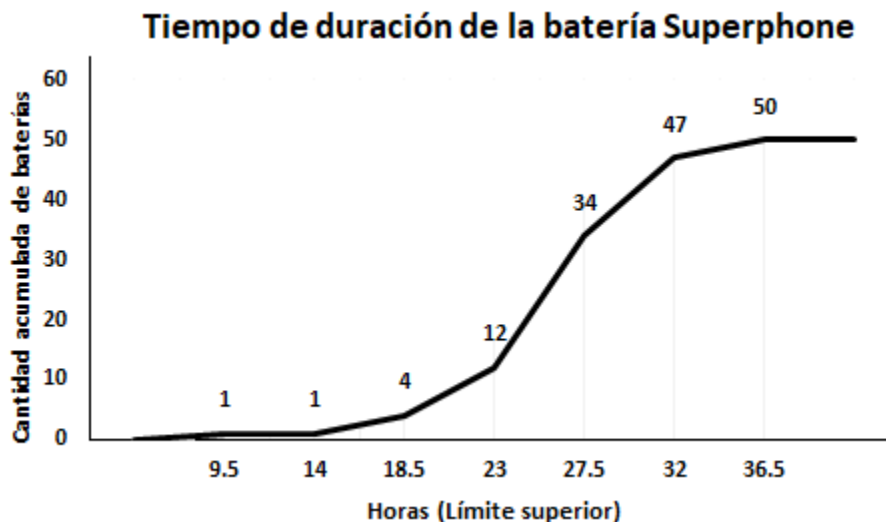


Gráfico 1.11. Polígono de frecuencias acumuladas del tiempo de duración de la batería del Superphone. Al igual que los polígonos de frecuencias (Gráfico 1.9), los de frecuencias acumuladas también se pueden exponer en un solo gráfico:

Tiempo de duración de las baterías de los celulares Ultraphone y Superphone

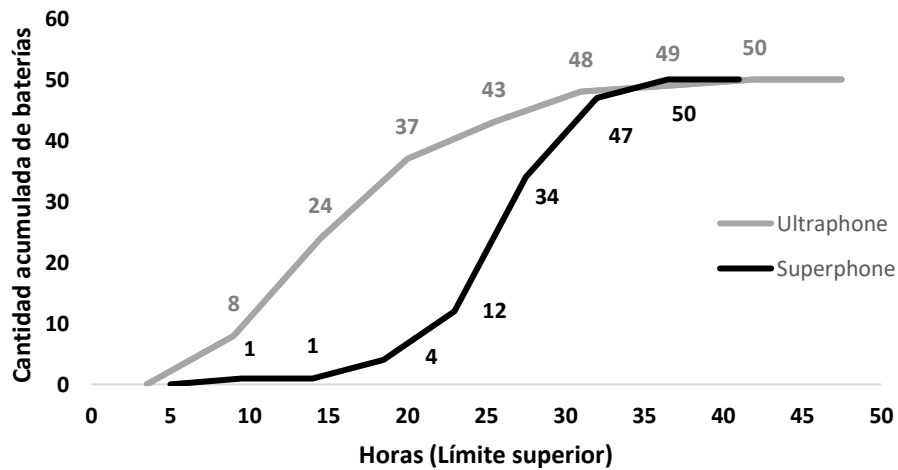


Gráfico 1.12. Polígonos de frecuencias acumuladas de los tiempos de duración de las baterías de los celulares Ultraphone y Superphone.

En el Gráfico 1.12 se puede observar cómo, aproximadamente, 43 baterías del celular Ultraphone (casi el 100%) duran a lo más 25 horas; mientras que poco menos de 34 baterías del celular Superphone (alrededor del 67%) se ubican en ese mismo rango. De esta manera, para estas muestras, la batería del celular Superphone ofrece un mayor tiempo de duración de la batería.

1.4.6. Polígono de frecuencias relativas y relativas acumuladas

Por otro lado, si el número de datos en alguno de los dos conjuntos hubiera sido marcadamente superior, el hacer la comparación sólo considerando las frecuencias no hubiera sido apropiado. En tal caso, lo más adecuado es transformar las frecuencias en cantidades relativas, ya que esto permite trabajar con los datos en una misma escala. La frecuencia relativa de una clase es la proporción de todos los datos que se ubican en la clase. Se obtiene al dividir la frecuencia de la clase entre el total de datos. Una vez determinada la frecuencia relativa, la frecuencia relativa acumulada se obtiene al sumar a la frecuencia relativa de la clase las frecuencias relativas previas (o dividiendo la frecuencia acumulada de la clase entre el total de datos), tal como se muestra en la siguiente tabla:

Tabla 1.8. Distribución de frecuencias del tiempo de duración de la batería del celular Ultraphone (Frecuencia relativa y Frecuencia relativa acumulada).

Clase $L_{inf} - L_{sup}$	Marca de Clase y_i	Frecuencia f_i	Frecuencia acumulada F_i	Frecuencia relativa fr_i	Frecuencia relativa acumulada Fr_i
3.5 – 9.0	6.25	8	8	$\frac{8}{50} = 0.16$	0.16
9.0 – 14.5	11.75	16	24	$\frac{16}{50} = 0.32$	$0.32 + 0.16 = 0.48$
14.5 – 20.0	17.25	13	37	$\frac{13}{50} = 0.26$	$0.26 + 0.48 = 0.74$
20.0 – 25.5	22.75	6	43	$\frac{6}{50} = 0.12$	$0.12 + 0.74 = 0.86$
25.5 – 31.0	28.25	5	48	$\frac{5}{50} = 0.1$	$0.1 + 0.86 = 0.96$
31.0 – 36.5	33.75	1	49	$\frac{1}{50} = 0.02$	$0.02 + 0.96 = 0.98$
36.5 – 42.0	39.25	1	50	$\frac{1}{50} = 0.02$	$0.02 + 0.98 = 1$

$$\sum f_i = 50$$

$$\sum fr_i = 1$$

En el Gráfico 1.13 se muestran los polígonos de frecuencias relativas del tiempo de duración de la batería de ambos tipos de celular. Como se puede observar, este gráfico es similar al Gráfico 1.9, ya que lo único que cambia es la escala.

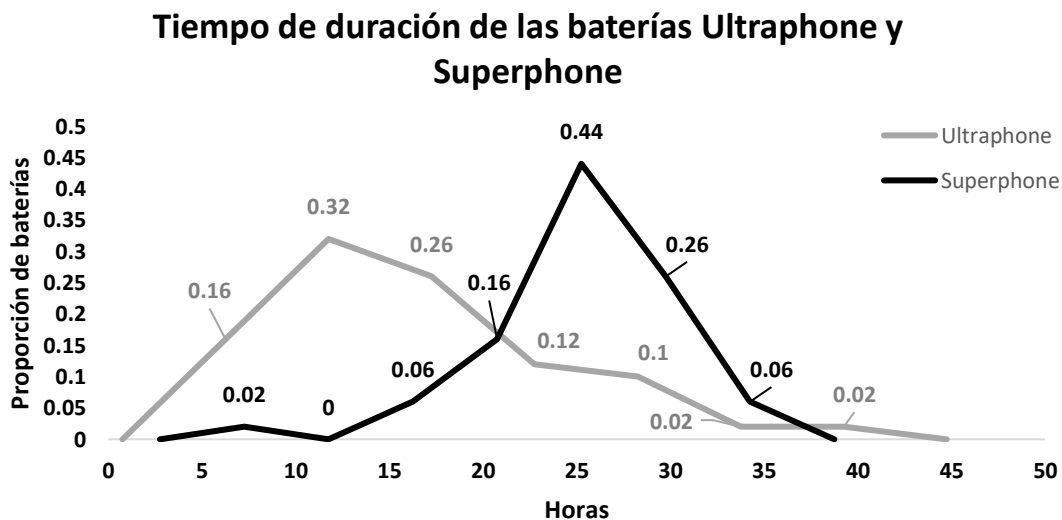


Gráfico 1.13. Polígonos de frecuencia relativa del tiempo de duración de la batería de los celulares Ultraphone y Superphone.

En el caso de los polígonos de frecuencias relativas acumuladas, estos son similares a los que se presentan en el Gráfico 1.12, tal como se muestra a continuación:

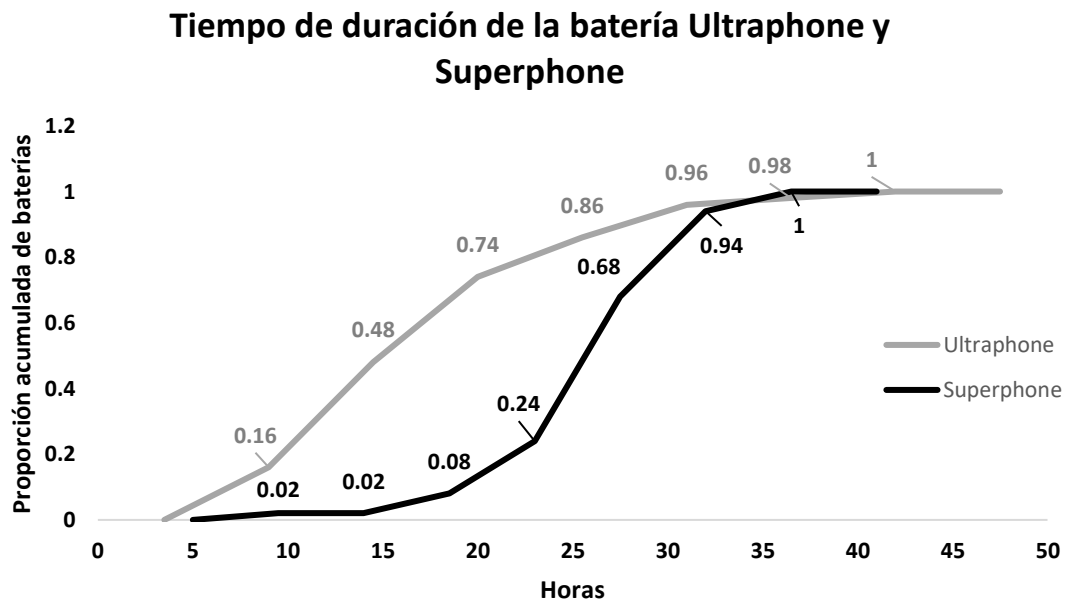


Gráfico 1.14. Polígonos de frecuencias relativas acumuladas Ultraphone y Superphone.

Hasta aquí, se ha visto cómo únicamente mediante las representaciones gráficas se puede dar respuesta al ejemplo 1.4. Sin embargo, habrá ocasiones en las que la información que proporcionen los gráficos no sea suficiente. Por ejemplo, supón que Sofía cree que necesita más datos antes de llegar a una conclusión, por lo que decide tomar otra muestra del tiempo de duración de la batería de cada celular. En el siguiente gráfico se muestra la distribución de los datos que obtuvo.

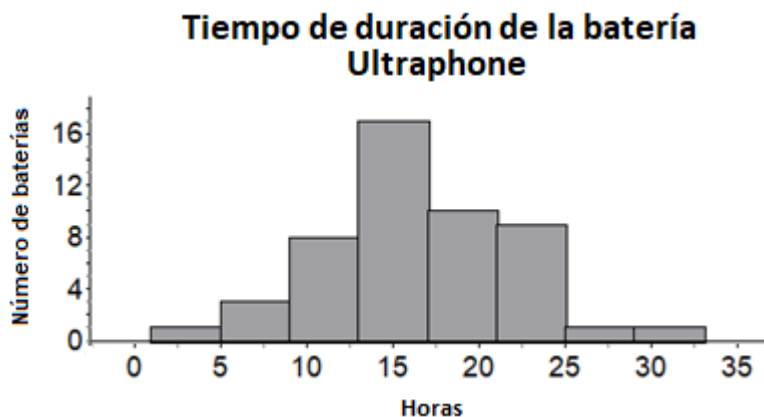


Gráfico 1.15. Tiempo de duración de la batería del celular Ultraphone.

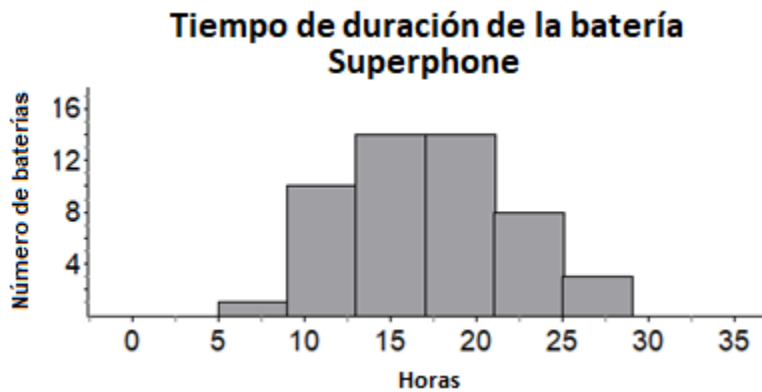


Gráfico 1.16. Tiempo de duración de la batería del celular Superphone.

Bajo este nuevo escenario, resulta complicado tomar una decisión apoyados sólo en la información que proporcionan los histogramas. De esta manera, será necesario recurrir a descripciones más precisas, como lo son los resúmenes numéricos. No obstante, antes de pasar a este tema, es necesario mostrar las representaciones gráficas que se utilizan cuando la variabilidad en los datos es poca.

1.5. Representación tabular y gráfica de una variable cuantitativa puntual

A diferencia del ejemplo 1.4, hay situaciones en las que es mejor trabajar con un conjunto de datos de manera puntual. En particular, cuando la variabilidad es poca. Por ejemplo, cuando se trata de las calificaciones finales de un grupo de estudiantes en cierta materia, estas no pueden ser menores que cinco, ni mayores que 10, por lo que el rango es sólo de cinco unidades. De esta manera, convendrá trabajar con cinco clases, cada una con amplitud igual a la unidad; la primera clase para la calificación 5, la segunda para el 6, la tercera para el 7, y así sucesivamente. Para ejemplificar esto, considera el siguiente ejemplo.

Ejemplo 1.5: Calificaciones finales

En la Tabla 1.9 se presentan las calificaciones finales de 35 alumnos de un grupo de estadística. A partir de estos datos, construye la tabla de distribución de frecuencias y el histograma correspondiente.

Tabla 1.9. Calificación final de 35 alumnos.

6	5	9	6	8
7	7	8	8	5
5	7	9	8	7
8	5	7	9	8
5	6	5	6	6
8	6	10	6	9
9	9	6	8	10

Solución

De acuerdo con lo mencionado, para este conjunto de datos se pueden utilizar cinco clases, cada una con amplitud igual a la unidad. Así, la tabla de distribución de frecuencias es la siguiente:

Tabla 1.10. Tabla de distribución de frecuencia de 35 calificaciones finales.

Calificación	Frecuencia f_i	Frecuencia acumulada F_i	Frecuencia relativa fr_i	Frecuencia relativa acumulada Fr_i
5	6	6	$\frac{6}{35} = 0.17$	0.17
6	8	14	$\frac{8}{35} = 0.23$	$0.17 + 0.23 = 0.40$
7	5	19	$\frac{5}{35} = 0.14$	$0.40 + 0.14 = 0.54$
8	8	27	$\frac{8}{35} = 0.23$	$0.54 + 0.23 = 0.77$
9	6	33	$\frac{6}{35} = 0.17$	$0.77 + 0.17 = 0.94$
10	2	35	$\frac{2}{35} = 0.06$	$0.94 + 0.06 = 1.00$

$$\sum f_i = 35$$

$$\sum fr_i = 1$$

Aplicar el procedimiento utilizado en el ejemplo 1.4 en esta nueva situación es posible, pero requerirá de más tiempo y se podrían obtener resultados fuera del contexto del problema. Por ejemplo, tener una clase que considere calificaciones menores que 5 o mayores que 10.

Una vez obtenida la tabla de distribución de frecuencias, el histograma correspondiente a las calificaciones finales es el que se muestra a continuación:

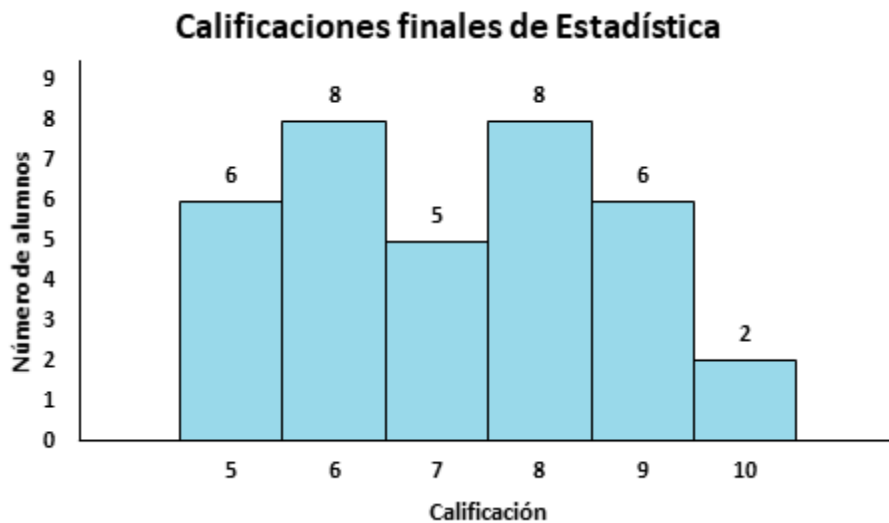


Gráfico 1.17. Calificaciones finales de 35 estudiantes de estadística.

Construido el histograma, los polígonos de frecuencias, de frecuencias acumuladas, de frecuencias relativas y de frecuencias relativas acumuladas, se obtienen de la misma forma en cómo se hizo en el ejemplo 1.4.



Ejercicios 1.2

1. Los siguientes datos representan la presión arterial sistólica (medida en milímetros de mercurio) de 50 hombres adultos saludables:

Tabla 1.11. Presión arterial sistólica de 50 hombres adultos.

110	105	120	135	150	140	129	128	122	120
115	110	130	125	123	120	123	126	140	146
150	145	128	130	120	120	120	125	127	110
115	115	129	120	123	123	140	145	145	128
129	130	118	120	125	130	132	150	150	148

- a) Construye la tabla de distribución de frecuencias y descríbela.
 - b) Elabora el histograma y el polígono de frecuencias acumuladas e interprétalos.
2. Con los datos de la Tabla 1.11, construye los gráficos de barras y de sectores considerando como categorías los niveles de hipertensión que se muestran en la siguiente tabla:

Tabla 1.12. Niveles de hipertensión.

Nivel	Presión arterial
Presión baja	Menor a 120
Prehipertensión	De 120 a 139
Hipertensión	140 o más

- ¿Qué porcentaje de hombres tiene presión baja?
- ¿Qué porcentaje de hombres tiene hipertensión?

1.6. Medidas de tendencia central para datos no agrupados

Las medidas de tendencia central resumen un conjunto de datos en un único valor que describe o representa el centro de dicho conjunto.

1.6.1. Moda (\hat{x})

La moda de un conjunto de datos se define como el valor que aparece con mayor frecuencia.

- Cuando dos valores ocurren con la misma frecuencia, y esta es la más alta, ambos valores son modas, por lo que el conjunto de datos es bimodal.
- Cuando son más de dos valores los que tienen la frecuencia más alta, todos son modas, por lo que el conjunto de datos es multimodal.
- Cuando ningún valor se repite se dice que no hay moda.

1.6.2. Mediana (\tilde{x})

Suponiendo que los datos de una muestra están ordenados de menor a mayor, si el número de datos es impar, la mediana de la muestra se define como el valor del dato central. Si el número de datos es par, la mediana de la muestra es el promedio de los valores de los dos datos centrales.

1.6.3. Media (\bar{x})

Si los datos de la muestra son x_1, x_2, \dots, x_n , entonces la media \bar{x} se define como la suma de los valores de los datos dividida entre el tamaño n de la muestra. Esto es,

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i$$

Donde $i = 1, 2, \dots, n$.

Ejemplo 1.6: Ventas de petróleo

En la siguiente tabla se muestran los precios mensuales (por barril) de la Mezcla Mexicana de petróleo de exportación durante el año 2014.

Tabla 1.13. Precios mensuales de la Mezcla Mexicana de Exportación

Mes	Precio (Dólares)
Diciembre	52
Noviembre	71
Octubre	75
Septiembre	86
Agosto	91
Julio	95
Junio	99
Mayo	97
Abril	96
Marzo	93
Febrero	93
Enero	91

a) Calcula la moda, la mediana y la media del precio de la mezcla mexicana de ese año.

Solución

Para dar solución, primero organiza los datos de menor a mayor.

52 71 75 86 91 91 93 93 95 96 97 99

Recuerda que la moda es el valor con la mayor frecuencia. En este caso, hay dos valores con frecuencia dos. Por lo tanto, el conjunto de datos es bimodal:

$$\hat{x} = 91 \text{ y } 93 \text{ dólares}$$

Como el conjunto de datos es par, la mediana se determina a partir del promedio de los valores de los dos datos centrales:

52 71 75 86 91 91 93 93 95 96 97 99

$$\tilde{x} = \frac{91 + 93}{2} = 92 \text{ dólares}$$

Finalmente, la media es la suma de los valores dividida entre el número total de datos:

$$\bar{x} = \frac{1}{12} \sum_{i=1}^{12} x_i = \frac{52 + 71 + 75 + 86 + 91 + 91 + 93 + 93 + 95 + 96 + 97 + 99}{12}$$

$$\bar{x} = 86.58 \text{ dólares}$$

Por lo tanto, el precio promedio de la mezcla mexicana durante el año 2014 fue de 86.58 dólares por barril.

b) En el mes de Diciembre el precio de la mezcla fue el más bajo. Calcula el precio promedio y la mediana suponiendo que el precio en ese mes fue de 10 dólares. ¿Qué observas?

Solución

Considera nuevamente los datos en orden de magnitud:

10 71 75 86 91 91 93 93 95 96 97 99

Como el número de datos sigue siendo el mismo, la mediana no se modifica:

$$\tilde{x} = \frac{91 + 93}{2} = 92 \text{ dólares}$$

En el caso de la media se tiene lo siguiente:

$$\bar{x} = \frac{1}{12} \sum_{i=1}^{12} x_i = \frac{10 + 71 + 75 + 86 + 91 + 91 + 93 + 93 + 95 + 96 + 97 + 99}{12}$$

$$\bar{x} = 83.08 \text{ dólares}$$

Como se puede observar, la disminución del valor de un dato no afecta el valor de la mediana. No obstante, el valor de la media pasa de 86.58 a 83.08 dólares. Este cambio se debe a que la media depende del valor de cada dato, mientras que la mediana no, ya que esta última es una medida de posición. Este fenómeno es más claro en el siguiente inciso.

c) Ahora supón que en el mes de Diciembre el precio de la mezcla fue de 500 dólares. Calcula el precio promedio y la mediana. ¿Qué observas?

Solución

Con la nueva modificación, la organización de los datos queda de la siguiente manera:

71 75 86 91 91 93 93 95 96 97 99 500

Para este nuevo arreglo, la mediana es la siguiente:

$$\tilde{x} = \frac{93 + 93}{2} = 93 \text{ dólares}$$

Por su parte, el valor de la media es el que se obtiene a continuación:

$$\bar{x} = \frac{1}{12} \sum_{i=1}^{12} x_i = \frac{500 + 71 + 75 + 86 + 91 + 91 + 93 + 93 + 95 + 96 + 97 + 99}{12}$$

$$\bar{x} = 123.92 \text{ dólares}$$

De nueva cuenta, se puede apreciar cómo el valor de la mediana no se ve tan afectado por la presencia del valor atípico 500, ya que pasa 92 a 93 dólares. En cambio, la media sí se ve fuertemente afectada, pues su valor cambia de 86.58 a 123.92 dólares.

El fenómeno observado en los incisos b y c se conoce como *robustez*, y se puede resumir de la siguiente manera: Se dice que la media \bar{x} no es una medida **robusta** de centro, debido a que es sensible a pocas observaciones atípicas. Por el contrario, debido a que la mediana \tilde{x} no es sensible a pocas de tales observaciones, se dice que es robusta. Esta es la razón por la que es preferible utilizar la mediana en lugar de la media para describir el centro de una distribución que es asimétrica o que tiene valores atípicos. La media es apropiada cuando la distribución de los datos es simétrica o aproximadamente simétrica.

1.6.4. Comparación entre la media y la mediana

- La media (\bar{x}) y la mediana (\tilde{x}) de una distribución aproximadamente simétrica se encuentran muy cerca.

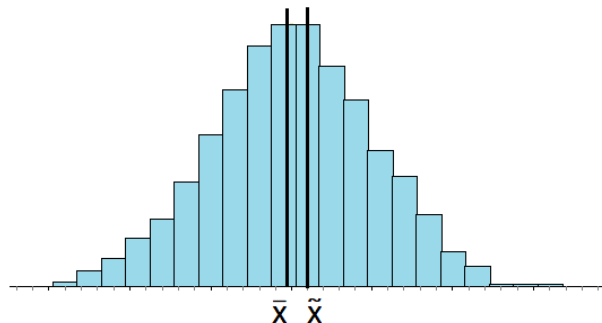


Gráfico 1.18. Ubicación de la media y la mediana en una distribución aproximadamente simétrica.

- En una distribución que es exactamente simétrica, la media (\bar{x}) y la mediana (\tilde{x}) coinciden.

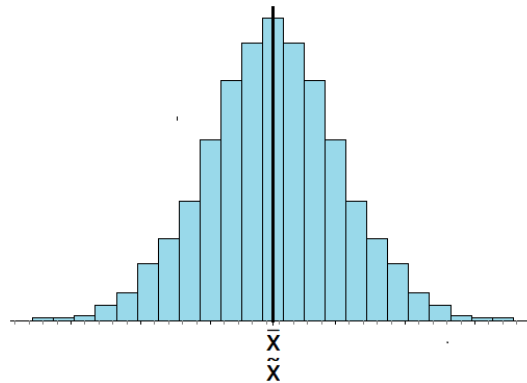


Gráfico 1.19. Ubicación de la media y la mediana en una distribución exactamente simétrica.

- En una distribución asimétrica la media (\bar{x}) queda desplazada hacia la cola más larga.

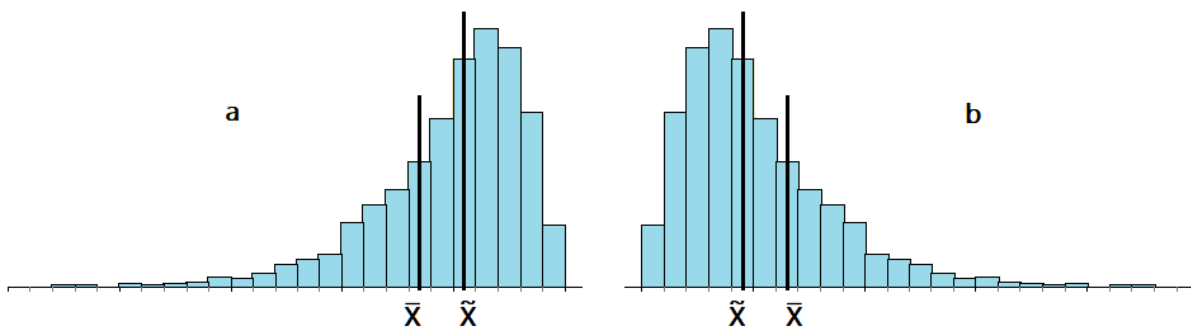


Gráfico 1.20. Ubicación de la media y la mediana en una distribución asimétrica hacia la izquierda (a) y asimétrica hacia la derecha (b).

1.7. Medidas de tendencia central para datos agrupados

Ejemplo 1.6: Estaturas de estudiantes

En la Tabla 1.14 se muestra la distribución de las estaturas (en centímetros) de 45 estudiantes de bachillerato. ¿Cómo determinarías la media, la mediana y la moda de este conjunto de datos?

Tabla 1.14. Distribución de las estaturas de 45 estudiantes de bachillerato.

Clase $L_{inf} - L_{sup}$	Marca de Clase y_i	Frecuencia f_i	Frecuencia relativa fr_i	Frecuencia acumulada F_i	Frecuencia relativa acumulada Fr_i
149.75 — 154.25	152	7	0.16	7	0.16
154.25 — 158.75	156.5	2	0.04	9	0.20
158.75 — 163.25	161	12	0.27	21	0.47
163.25 — 167.75	165.5	7	0.16	28	0.62
167.75 — 172.25	170	5	0.11	33	0.73
172.25 — 176.75	174.5	6	0.13	39	0.87
176.75 — 181.25	179	6	0.13	45	1.00

$$\sum f_i = 45 \quad \sum fr_i = 1$$

1.7.1. Media (\bar{x})

La Tabla 1.14 proporciona la cantidad de datos (frecuencia) que se ubican en cada clase, pero no el valor de cada uno de ellos. Bajo este escenario, para poder estimar el valor de la media se hace el supuesto de que cada dato toma el valor de la marca de clase correspondiente. Así, la fórmula para calcular la media toma la siguiente forma:

$$\bar{x} = \frac{1}{n} (f_1 y_1 + f_2 y_2 + \dots + f_n y_n) = \frac{1}{n} \sum_{i=1}^N f_i y_i$$

Donde,

- y_1, y_2, \dots, y_n son las marcas de clase de cada clase.
- f_1, f_2, \dots, f_n son las frecuencias absolutas de cada clase.
- n es el número total de datos.
- N es el número de clases.

Teniendo en consideración la fórmula anterior, la media de las 45 estaturas se determina de la siguiente manera:

$$\begin{aligned} \bar{x} &= \frac{1}{45} \sum_{k=1}^7 f_i y_i \\ &= \frac{(7)(152) + (2)(156.5) + (12)(161) + (7)(165.5) + (5)(170) + (6)(174.5) + (6)(179)}{45} \end{aligned}$$

$$\bar{x} = 165.3 \text{ centímetros.}$$

De esta manera, si se elige a un estudiante al azar de este grupo, se espera que tenga una estatura aproximada de 165.3 centímetros.

1.7.2. Mediana (\tilde{x})

Como la mediana es el valor que se ubica a la mitad del arreglo, ordenado de menor a mayor, lo primero que se debe hacer es identificar la clase que contiene a dicho valor. Para ello, se utiliza la siguiente expresión:

$$P_m = \frac{n + 1}{2}$$

Donde n es el total de datos. De esta manera,

$$P_m = \frac{45 + 1}{2} = 23$$

Así, la mediana corresponde al dato 23, el cual se ubica en la cuarta clase (163.25 – 167.75]. Hasta antes de esta clase, sólo se han contabilizado 21 datos (de acuerdo con la frecuencia acumulada), por lo que los datos del 22 hasta el 28, incluido el 23, se ubican en la cuarta clase.

Una vez identificada la clase que contiene a la mediana, su valor se calcula mediante la siguiente fórmula:

$$\tilde{x} = L_{inf} + \left(\frac{\frac{n}{2} - F_{ant}}{f} \right) * A$$

Donde,

- L_{inf} es el límite inferior del intervalo que contiene la mediana.
- f es la frecuencia absoluta del intervalo que contiene a la mediana.
- F_{ant} es la frecuencia absoluta acumulada hasta el intervalo anterior al que contiene a la mediana.
- n es el total de datos.
- A es la amplitud del intervalo que contiene a la mediana.

De acuerdo con la fórmula anterior, la mediana de las 45 estaturas se determina de la siguiente manera:

$$\tilde{x} = L_{inf} + \left(\frac{\frac{n}{2} - F_{ant}}{f} \right) * A = 163.25 + \left(\frac{\frac{45}{2} - 21}{7} \right) (4.5) = 164.21 \text{ centímetros.}$$

Por lo tanto, la mitad de los estudiantes mide menos de 164.21 centímetros y la otra mitad tiene una estatura mayor a este valor.

1.7.3. Moda (\hat{x})

Al igual que con la mediana, lo primero es identificar la clase que la contiene a la moda. Como la moda es el dato que más se repite, ésta se ubica en la clase con mayor frecuencia. Así, nuestra clase de interés es (158.75 – 163.25].

Una vez identificada la clase que contiene a la moda, su valor se calcula mediante la siguiente fórmula:

$$\hat{x} = L_{inf} + \left[\frac{f - f_{ant}}{(f - f_{ant}) + (f - f_{sup})} \right] * A$$

Donde

- L_{inf} es el límite inferior de la clase que contiene a la moda.
- f es la frecuencia de la clase que contiene a la moda.
- f_{ant} es la frecuencia de la clase anterior a la clase que contiene a la moda.
- f_{sup} es la frecuencia de la clase superior a la clase que contiene a la moda.
- n es el total de datos.
- A es la amplitud de la clase que contiene a la moda.

De acuerdo con la fórmula anterior, la moda de las 45 estaturas es la siguiente:

$$\hat{x} = L_{inf} + \left[\frac{f - f_{ant}}{(f - f_{ant}) + (f - f_{sup})} \right] * A = 158.75 + \left[\frac{12 - 2}{(12 - 2) + (12 - 7)} \right] (4.5)$$

$$\hat{x} = 161.75 \text{ centímetros.}$$

Por lo tanto, se espera que la estatura con mayor frecuencia en este grupo sea próxima a 161.75 centímetros. En el caso de que haya dos clases, o más, con la mayor frecuencia, la fórmula anterior se aplica para cada una de ellas. De presentarse esta situación, se estaría hablando de un conjunto de datos bimodal o multimodal, según sea el caso.

Se debe tener en cuenta que los procedimientos anteriores sólo dan valores aproximados de la media, la mediana y la moda, ya que no se conoce el valor específico de cada dato. Además, tanto para la mediana como para la moda, una forma de corroborar que el valor que se obtuvo es adecuado, es verificando que dichos valores no queden fuera de los intervalos que contienen a estas medidas.



Ejercicios 1.3

1. Para medir el índice de la calidad del aire en la Ciudad de México se utiliza la siguiente escala: Buena (0 – 50), Regular (51 – 100), Mala (101 – 150), Muy mala (151 – 200), Extremadamente mala (201 – 300), y Peligrosa (301 – 500).

En la Tabla 1.15 se muestran los niveles de contaminación promedio mensual por ozono (O_3), medido en partes por millón (ppm), de los 10 primeros meses del año 2017.

Tabla 1.15. Promedios mensuales de contaminación por ozono (O_3).

Mes	Nivel de O_3
Enero	87
Febrero	93
Marzo	85
Abril	95
Mayo	128
Junio	90
Julio	86
Agosto	83
Septiembre	64

- Calcula la media, la mediana y la moda de este conjunto de datos e interpreta los resultados.
 - ¿De qué manera el valor del nivel de contaminación del mes de mayo influye en los valores de la media y la mediana?
2. En la siguiente tabla se muestra la distribución del peso (kg) de 38 estudiantes de bachillerato.

Tabla 1.16. Distribución del peso de 38 estudiantes de bachillerato.

Clase $L_{inf} - L_{sup}$	Marca de Clase y_i	Frecuencia f_i	Frecuencia relativa fr_i	Frecuencia acumulada F_i	Frecuencia relativa acumulada Fr_i
42 – 49	45.5	4	0.11	4	0.11
49 – 56	52.5	6	0.16	10	0.26
56 – 63	59.5	13	0.34	23	0.61
63 – 70	66.5	9	0.24	32	0.84
70 – 77	73.5	4	0.11	36	0.95
77 – 84	80.5	0	0.00	36	0.95
84 – 91	87.5	2	0.05	38	1.00

Calcula le media, la mediana y la moda, e interpreta los resultados.

1.8. Medidas de posición y de dispersión

La media, la mediana y la moda son medidas de centro de una distribución. Sin embargo, caracterizar a una distribución sólo con una de estas medidas puede ser engañoso.

Ejemplo 7: Compra de medicamentos

En la rutina de un laboratorio, se toma una muestra de 13 píldoras de cada uno de los lotes que recibe y se calcula la media del componente activo (medido en miligramos), la cual debe ser lo más parecida posible a un valor deseado (supongamos 0.62 mg). ¿De cuál de las siguientes dos farmacéuticas conveniente comprar un lote de medicamentos?

Tabla 1.17. Muestras del componente activo de 13 píldoras de dos farmacéuticas.

Farmacéutica 1												
0.65	0.61	0.59	0.6	0.55	0.64	0.6	0.58	0.62	0.57	0.56	0.6	0.63
Farmacéutica 2												
0.7	0.55	0.35	0.6	0.4	0.75	0.6	0.85	0.5	0.6	0.45	0.8	0.65

Solución

Como lo que interesa es que la media del componente activo de cada una de las muestras sea lo más parecida al 0.62 mg, se determina su valor en cada caso:

Farmacéutica 1

$$\bar{x} = \frac{0.65 + 0.61 + 0.59 + 0.6 + 0.55 + 0.64 + 0.6 + 0.58 + 0.62 + 0.57 + 0.56 + 0.6 + 0.63}{13}$$

$$\bar{x} = 0.6 \text{ mg}$$

Farmacéutica 2

$$\bar{x} = \frac{0.7 + 0.55 + 0.35 + 0.6 + 0.4 + 0.75 + 0.6 + 0.85 + 0.5 + 0.6 + 0.45 + 0.8 + 0.65}{13}$$

$$\bar{x} = 0.6 \text{ mg}$$

De ambas muestras se obtiene la misma media del componente activo. Por lo tanto, de cualquiera de las dos farmacéuticas sería adecuado comprar el lote de medicamentos. Sin embargo, se debe tener en cuenta que un lote con una concentración media adecuada en su componente activo puede ser muy peligroso, si hay píldoras con contenidos del componente muy elevados y otras con contenidos muy bajos. En el primer caso, el medicamento podría ser nocivo, y en el segundo, no tendría efecto alguno en quien lo tome. De esta manera, además de estar interesados en la media de la concentración del componente activo de cada muestra, estamos interesados en su dispersión o variabilidad.

1.8.1. Rango

El rango es la medida de dispersión más sencilla. Se define como la diferencia entre el valor máximo ($x_{m\acute{a}x}$) y el valor mínimo ($x_{m\acute{i}n}$).

$$Rango = x_{m\acute{a}x} - x_{m\acute{i}n}$$

En el caso de las farmacéuticas, los rangos correspondientes se calculan a continuación:

Farmacéutica 1

$$Rango = x_{m\acute{a}x} - x_{m\acute{i}n} = 0.65 - 0.5 = 0.1 \text{ mg}$$

Farmacéutica 2

$$Rango = x_{m\acute{a}x} - x_{m\acute{i}n} = 0.85 - 0.35 = 0.5 \text{ mg}$$

Como se puede observar, conviene comprar el lote de medicamentos de la farmacéutica 1, ya que la dispersión de los datos es más pequeña.

Sin embargo, se debe tener en cuenta que la presencia de una o más observaciones atípicas podrían enmascarar la conclusión anterior, pues el rango sólo considera los dos

valores extremos de un conjunto de datos, pero no informa sobre lo que ocurre entre estos valores. Por ejemplo, considera que las muestras obtenidas de cada uno de los lotes son las que se presentan a continuación:

Tabla 1.18. Muestras del componente activo de 13 píldoras de dos farmacéuticas.

Farmacéutica 1												
0.88	0.61	0.59	0.6	0.32	0.64	0.6	0.58	0.62	0.57	0.56	0.6	0.63

Farmacéutica 2												
0.7	0.55	0.35	0.6	0.4	0.75	0.6	0.85	0.5	0.6	0.45	0.8	0.65

Bajo este nuevo escenario, los rangos de cada una de las farmacéuticas son los siguientes:

Farmacéutica 1

$$Rango = x_{m\acute{a}x} - x_{m\acute{i}n} = 0.88 - 0.32 = 0.56 \text{ mg}$$

Farmacéutica 2

$$Rango = x_{m\acute{a}x} - x_{m\acute{i}n} = 0.85 - 0.35 = 0.5 \text{ mg}$$

De esta manera, pareciera que la mejor opción es comprar el lote de medicamentos de la farmacéutica 2. No obstante, la mayoría de los datos de la farmacéutica 1 están agrupados en torno a la media, tal como se observa en los siguientes gráficos:

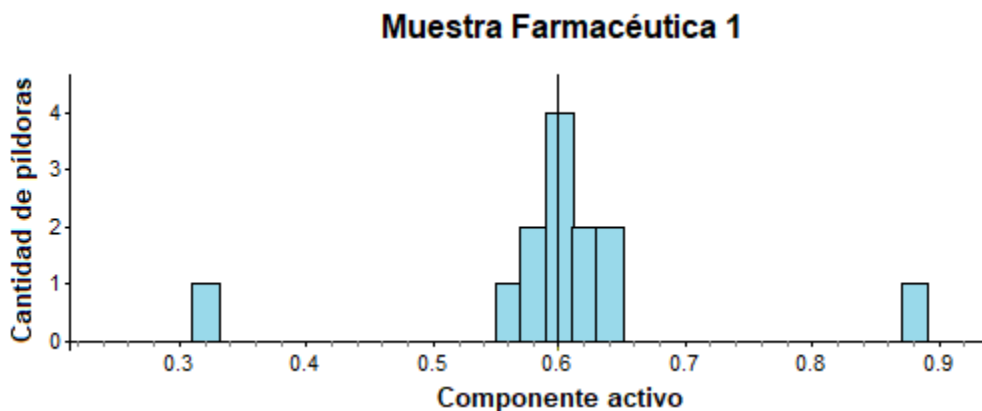


Gráfico 1.21. Distribución del componente activo de la farmacéutica 1.

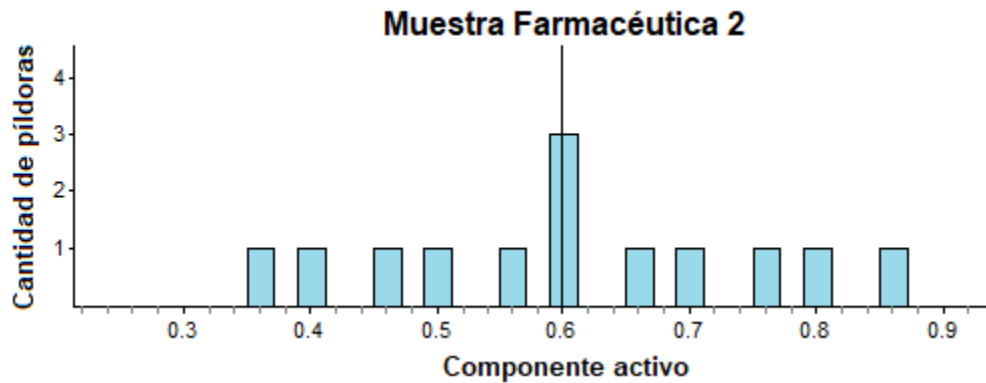


Gráfico 1.22. Distribución del componente activo de la farmacéutica 2.

Para evitar los problemas por la presencia de valores atípicos, se puede mejorar la descripción de la dispersión observando sólo el 50% de los datos centrales. Para ello, conviene dividir nuestra distribución en cuatro partes iguales.

1.8.2. Cuartiles

Los **cuartiles** determinan entre qué valores se encuentra la mitad central de los datos.

1. El primer cuartil (Q_1) separa el primer 25% de los datos.
2. El segundo cuartil separa el 50% de los datos y corresponde a la mediana.
3. El tercer cuartil (Q_3) separa el 75% de los datos.

Para determinar los cuartiles, se puede hacer mediante la regla de las medianas:

1. Ordena los datos en orden creciente y localiza la mediana global \tilde{x} .
2. El primer cuartil Q_1 es la mediana de los datos situados a la izquierda de \tilde{x} .
3. El tercer cuartil Q_3 es la mediana de los datos situados a la derecha de \tilde{x} .

Los cuartiles correspondientes a las farmacéuticas 1 y 2 (Tabla 1.18) son:

Farmacéutica 1

El primer paso es ordenar los datos de menor a mayor y encontrar el valor de la mediana (\tilde{x}). Como el número de datos es impar (13 datos), la mediana es el valor del dato que se ubica en el centro:

0.32 0.56 0.57 0.58 0.59 0.6 0.6 0.6 0.61 0.62 0.63 0.64 0.88

De esta manera, $\tilde{x} = 0.6 \text{ mg}$.

El siguiente paso es determinar el primer cuartil (Q_1), el cual es la mediana de los datos situados a la izquierda de \tilde{x} . Como la cantidad de datos es impar (7 datos), el primer cuartil es el valor del dato central:

0.32 0.56 0.57 0.58 0.59 0.6 0.6

De esta manera, $Q_1 = 0.58 \text{ mg}$.

Finalmente, se determina el tercer cuartil (Q_3), el cual es la mediana de los datos situados a la derecha de \tilde{x} :

0.6 0.6 0.61 0.62 0.63 0.64 0.88

Así, $Q_3 = 0.62 \text{ mg}$.

Por lo tanto, el 50% de los datos centrales de la farmacéutica 1 se ubican entre 0.58 y 0.62 mg.

Farmacéutica 2

Siguiendo el procedimiento anterior, los cuartiles para la farmacéutica 2 son; $\tilde{x} = 0.6$, $Q_1 = 0.5$ y $Q_3 = 0.7$:

0.35 0.4 0.45 0.5 0.55 0.6 0.6 0.6 0.65 0.7 0.75 0.8 0.85

De esta manera, el 50% de los datos centrales de la farmacéutica 2 se ubican entre 0.5 y 0.7 mg. Por lo tanto, el 50% de los datos centrales está más disperso en la farmacéutica 2, por lo que convendría comprar el lote de medicamentos de la farmacéutica 1.

Una forma más clara de observar la posición de los cuartiles es mediante una representación gráfica.

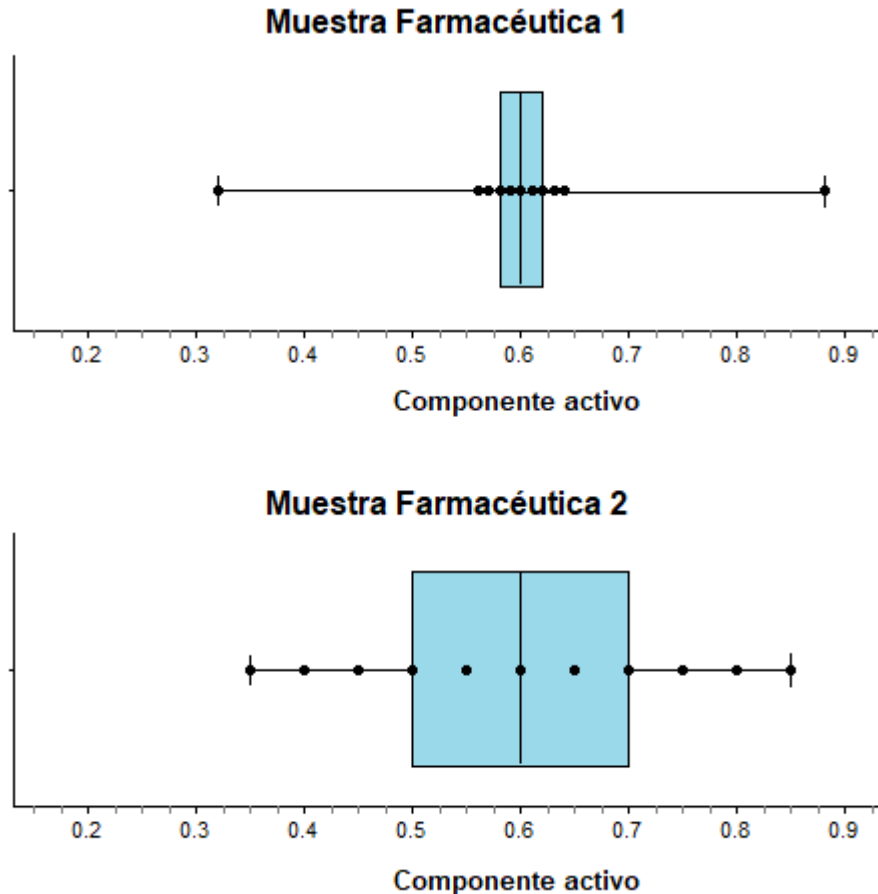
1.8.3. Diagrama de caja

Un diagrama de caja muestra gráficamente los cinco números resumen (x_{\min} , Q_1 , \tilde{x} , Q_3 , x_{\max}).

- Los lados izquierdo y derecho de la caja corresponden a los cuartiles Q_1 y Q_3 , respectivamente.
- El segmento del interior de la caja corresponde a la mediana \tilde{x} .

- Los extremos de los segmentos perpendiculares al lado izquierdo y derecho de la caja corresponden a los valores máximo ($x_{m\acute{a}x}$) y m nimo ($x_{m n}$), respectivamente.

Los diagramas de caja de las farmac uticas 1 y 2 se muestran a continuaci n:



Gr fico 1.23. Diagramas de caja de las farmac uticas 1 y 2.

En el Gr fico 1.23 se puede observar c mo el 50% central de los datos de la farmac utica 2 se encuentran m s dispersos que el 50% central de los datos de la farmac utica 1. Una descripci n m s precisa se puede obtener mediante el rango intercuartil.

1.8.4. Rango intercuartil

El rango intercuartil se define como la diferencia entre Q_3 y Q_1 .

$$RIC = Q_3 - Q_1$$

El RIC de las farmac uticas 1 y 2 son los siguientes:

Farmac utica 1

$$RIC = Q_3 - Q_1 = 0.62 - 0.58 = 0.04 \text{ mg}$$

Farmacéutica 2

$$RIC = Q_3 - Q_1 = 0.7 - 0.5 = 0.2 \text{ mg}$$

A partir del RIC, queda claro que el 50% de los datos centrales en la farmacéutica 2 se encuentran más dispersos que el 50% de los datos centrales de la farmacéutica 1. No obstante, a pesar de que esta medida nos ayuda a tomar una decisión sobre la dispersión de los datos frente a la presencia de observaciones atípicas, no nos proporciona una imagen completa de la dispersión total. Esta distinción corresponde a la varianza y a la desviación estándar.

1.8.5. Varianza (s^2)

Al hablar sobre la variación de un conjunto de datos es necesario especificar un valor de referencia respecto al cual varían. Este valor corresponde a la media, ya que es el que mejor representa el comportamiento de dicho conjunto. A la diferencia que existe entre cada dato y la media se le llama desviación. Como lo que nos interesa es un valor que informe sobre la desviación general del conjunto de datos, este se determina calculando la desviación promedio. No obstante, cada desviación debe elevarse al cuadrado, si no las desviaciones de los datos inferiores a la media se eliminarán con las desviaciones de los datos que son superiores a esta medida. A la desviación promedio se le conoce como varianza.

En otros términos, la varianza s^2 de un conjunto de datos es la suma de los cuadrados de las desviaciones de dichos datos respecto a su media \bar{x} , dividido entre $n - 1$. Algebraicamente, la varianza de n observaciones x_1, x_2, \dots, x_n se obtiene mediante la siguiente expresión:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1} = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n - 1}$$

Con base en la fórmula anterior, las varianzas de las farmacéuticas 1 y 2 son las siguientes:

Farmacéutica 1

Para facilitar los cálculos, considera el siguiente arreglo:

Tabla 1.19. Cálculo de la varianza.

Datos x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0.88	$0.88 - 0.6 = \mathbf{0.28}$	$(0.28)^2 = \mathbf{0.0784}$
0.61	$0.61 - 0.6 = \mathbf{0.01}$	$(0.01)^2 = \mathbf{0.0001}$
0.59	$0.59 - 0.6 = \mathbf{-0.01}$	$(-0.01)^2 = \mathbf{0.0001}$
0.6	$0.6 - 0.6 = \mathbf{0}$	$(0)^2 = \mathbf{0}$
0.32	$0.32 - 0.6 = \mathbf{-0.28}$	$(-0.28)^2 = \mathbf{0.0784}$
0.64	$0.64 - 0.6 = \mathbf{0.04}$	$(0.04)^2 = \mathbf{0.0016}$
0.6	$0.6 - 0.6 = \mathbf{0}$	$(0)^2 = \mathbf{0}$
0.58	$0.58 - 0.6 = \mathbf{-0.02}$	$(-0.02)^2 = \mathbf{0.0004}$
0.62	$0.62 - 0.6 = \mathbf{0.02}$	$(0.02)^2 = \mathbf{0.0004}$
0.57	$0.57 - 0.6 = \mathbf{-0.03}$	$(-0.03)^2 = \mathbf{0.0009}$
0.56	$0.56 - 0.6 = \mathbf{-0.04}$	$(-0.04)^2 = \mathbf{0.0016}$
0.6	$0.6 - 0.6 = \mathbf{0}$	$(0)^2 = \mathbf{0}$
0.63	$0.63 - 0.6 = \mathbf{0.03}$	$(0.03)^2 = \mathbf{0.0009}$

$$\sum (x_i - \bar{x})^2 = \mathbf{0.1628}$$

Como el total de datos es 13, la varianza para la farmacéutica 1 es:

$$s^2 = \sum_{i=1}^{13} \frac{(x_i - \bar{x})^2}{13 - 1} = \frac{0.1628}{13 - 1} = \frac{0.1628}{12} = 0.0136 \text{ mg}^2$$

Así, se tiene que, en promedio, los datos se desvían de la media en 0.0136 mg^2 .

Farmacéutica 2

Mediante el procedimiento anterior, la varianza para farmacéutica 2 es:

$$s^2 = \sum_{i=1}^{13} \frac{(x_i - \bar{x})^2}{13 - 1} = \frac{0.275}{13 - 1} = \frac{0.275}{12} = 0.0229 \text{ mg}^2$$

Como la varianza de la farmacéutica 2 es mayor que la varianza de la farmacéutica 1, la dispersión de los datos en esta última es mayor.

Por otra parte, observa que las unidades de medida de la varianza están al cuadrado, lo que resulta poco intuitivo. Para evitar la posible confusión que esto podría provocar, es

conveniente sacar la raíz cuadrada de la varianza. El valor resultante se conoce como desviación estándar.

1.8.6. Desviación estándar (s)

La desviación estándar es la raíz cuadrada positiva de la varianza s^2 :

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}} = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n - 1}}$$

Considerando la expresión anterior, las desviaciones estándar de las farmacéuticas 1 y 2 son las siguientes:

Farmacéutica 1

$$s = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{0.1628}{13 - 1}} = \sqrt{\frac{0.1628}{12}} = \sqrt{0.0136} = 0.1165 \text{ mg}$$

Farmacéutica 2

$$s = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{0.275}{13 - 1}} = \sqrt{\frac{0.275}{12}} = \sqrt{0.0229} = 0.1514 \text{ mg}$$

Por lo tanto, para las farmacéuticas 1 y 2 los datos se desvían de la media, en promedio, en 0.1165 mg y 0.1514 mg, respectivamente. Como la segunda desviación es mayor, la dispersión de los datos de la farmacéutica 2 es mayor.

1.9. Varianza y desviación estándar para datos agrupados

Considera nuevamente la distribución de las estaturas (en centímetros) de 45 estudiantes de bachillerato. ¿Cómo determinarías la varianza y la desviación estándar para este conjunto de datos?

Tabla 1.20. Distribución de las estaturas de 45 estudiantes de bachillerato.

Clase $L_{inf} - L_{sup}$	Marca de Clase y_i	Frecuencia f_i	Frecuencia relativa fr_i	Frecuencia acumulada F_i	Frecuencia relativa acumulada Fr_i
149.75 – 154.25	152	7	0.16	7	0.16
154.25 – 158.75	156.5	2	0.04	9	0.20
158.75 – 163.25	161	12	0.27	21	0.47
163.25 – 167.75	165.5	7	0.16	28	0.62
167.75 – 172.25	170	5	0.11	33	0.73
172.25 – 176.75	174.5	6	0.13	39	0.87
176.75 – 181.25	179	6	0.13	45	1.00

$$\sum f_i = 45 \quad \sum fr_i = 1$$

1.9.1. Varianza (s^2)

Como la Tabla 1.20 sólo proporciona la cantidad de datos que se ubican en cada clase, pero no el valor de cada uno de ellos, para estimar el valor de la varianza se hace el supuesto de que cada dato toma el valor de la marca de clase correspondiente. De esta manera, la fórmula para calcular la varianza toma la siguiente forma:

$$s^2 = \frac{f_1 \cdot (y_1 - \bar{x})^2 + f_2 \cdot (y_2 - \bar{x})^2 + \dots + f_n \cdot (y_n - \bar{x})^2}{n - 1} = \sum_{i=1}^N \frac{f_i \cdot (y_i - \bar{x})^2}{n - 1}$$

En donde:

- y_1, y_2, \dots, y_n son las marcas de clase de cada clase.
- f_1, f_2, \dots, f_n son las frecuencias de cada clase.
- n es el número total de datos.
- N es el número de clases.
- \bar{x} es la media de este conjunto de datos.

Para determinar la varianza de las 45 estaturas es necesario calcular primero la media. Como se trata de un conjunto de datos agrupados, recuerda que la media se calcula mediante la expresión,

$$\bar{x} = \frac{1}{n}(f_1y_1 + f_2y_2 + \dots + f_ny_n) = \frac{1}{n} \sum_{i=1}^N f_i y_i$$

La media para este conjunto de datos ya se calculó en el ejemplo 1.6, dando como resultado $\bar{x} = 165.3$ cm. Una vez obtenida la media, para calcular el valor de la varianza considera el siguiente arreglo:

Tabla 1.21. Cálculo de la varianza.

Marca de Clase y_i	Frecuencia f_i	$y_i - \bar{x}$	$(y_i - \bar{x})^2$	$f_i \cdot (y_i - \bar{x})^2$
152	7	$152 - 165.3 = -13.3$	$(-13.3)^2 = 176.89$	$7(176.89) = 1238.23$
156.5	2	$156.5 - 165.3 = -8.8$	$(-8.8)^2 = 77.44$	$2(77.44) = 154.88$
161	12	$161 - 165.3 = -4.3$	$(-4.3)^2 = 18.49$	$12(18.49) = 221.88$
165.5	7	$165.5 - 165.3 = 0.2$	$(0.2)^2 = 0.04$	$7(0.04) = 0.28$
170	5	$170 - 165.3 = 4.7$	$(4.7)^2 = 22.09$	$5(22.09) = 110.45$
174.5	6	$174.5 - 165.3 = 9.2$	$(9.2)^2 = 84.64$	$6(84.64) = 507.84$
179	6	$179 - 165.3 = 13.7$	$(13.7)^2 = 187.69$	$6(187.69) = 1226.14$

$$\sum f_i \cdot (y_i - \bar{x})^2 = 3359.7$$

Como el total de datos es 45, la varianza para las estaturas del grupo es:

$$s^2 = \sum_{i=1}^N \frac{f_i \cdot (y_i - \bar{x})^2}{n - 1} = \frac{3359.7}{45 - 1} = \frac{3359.7}{44} = 76.36 \text{ cm}^2$$

Para evitar la posible confusión que puedan provocar las unidades al cuadrado, es conveniente sacar la raíz cuadrada de la varianza. Como se vio anteriormente, el valor resultante se conoce como desviación estándar.

1.9.2. Desviación estándar (s)

La desviación estándar es la raíz cuadrada positiva de la varianza s^2 ,

$$s = \sqrt{\frac{f_1 \cdot (y_1 - \bar{x})^2 + f_2 \cdot (y_2 - \bar{x})^2 + \dots + f_n \cdot (y_n - \bar{x})^2}{n - 1}} = \sqrt{\sum_{i=1}^N \frac{f_i \cdot (y_i - \bar{x})^2}{n - 1}}$$

De esta manera, el valor de la desviación estándar de las 45 estaturas es:

$$s = \sqrt{\sum_{i=1}^N \frac{f_i \cdot (y_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{3359.7}{45 - 1}} = \sqrt{\frac{3359.7}{44}} = \sqrt{76.36} = 8.74 \text{ cm}$$

Por lo tanto, si se elige a un estudiante al azar de los 45, se espera que su estatura se encuentre entre 156.56 y 174.04 cm.

1.9.3. Propiedades de la desviación estándar.

- Cuando no hay dispersión, la desviación estándar es igual a cero ($s = 0$). Esto ocurre cuando todos los datos toman el mismo valor. En caso contrario, siempre es positiva ($s > 0$).
- A medida que los datos se separan más de la media, el valor de la desviación estándar se hace mayor.
- Las unidades de medida de la desviación estándar son las mismas que las unidades de medida de los datos originales.
- Fuertes asimetrías, o pocas observaciones atípicas, pueden hacer que aumente mucho el valor de la desviación estándar. Es decir, s no es robusta.



Ejercicios 1.4

1. De acuerdo con el Servicio Sismológico Nacional (SSN), ocurrieron 36 sismos en Chiapas con una intensidad mayor o igual a cuatro grados en la escala de Richter entre el cinco y el doce de septiembre de 2017. La intensidad de estos sismos se presenta a continuación:

Tabla 1.22. Intensidad de los sismos.

4.6	4.7	8.2	5.1	5.9	4.5	4.7	5.3	4.8	4.5	4.5	4.6	4.5	4.5	4.5
5.4	4.6	5.6	4.6	4.5	4.7	4.5	4.5	4.5	5	4.5	4.7	4.6	4.5	5.6
4.8	5.2	4.7	5.3	4.6	4.7									

- Dibuja un diagrama de caja con estos datos.
- Calcula la desviación estándar de la intensidad de los sismos e interpreta el resultado.

- c) Si se elimina el valor 8.2, ¿qué ocurre con el diagrama de caja y con la desviación estándar?
2. En la siguiente tabla se muestra la distribución del tiempo de traslado al colegio (en minutos) de 38 estudiantes.

Tabla 1.23. Tiempo de traslado al colegio de 38 estudiantes.

Clase $L_{inf} - L_{sup}$	Marca de Clase y_i	Frecuencia f_i	Frecuencia relativa fr_i	Frecuencia acumulada F_i	Frecuencia relativa acumulada Fr_i
3.5 — 17.5	10.5	3	0.08	3	0.08
17.5 — 31.5	24.5	6	0.16	9	0.24
31.5 — 45.5	38.5	7	0.18	16	0.42
45.5 — 59.5	52.5	0	0.00	16	0.42
59.5 — 73.5	66.5	14	0.37	30	0.79
73.5 — 87.5	80.5	1	0.03	31	0.82
87.5 — 101.5	94.5	7	0.18	38	1.00

Calcula la desviación estándar del tiempo de traslado de los estudiantes.

Evaluación de la Unidad I

1. Relaciona ambas columnas.

- a. La mediana muestral es 9. \tilde{x}
- b. Si se eliminan todos los ceros en un conjunto de datos, esto no afecta el valor de esta medida de centro. 18, 18, 18, 18, 18, 18
- c. La moda muestral es 9. 5, 7, 8, 10, 13, 14
- d. Un conjunto de datos puede no tener esta medida de centro. \bar{x}
- e. La media muestral es 9. 1, 2, 5, 9, 9, 15
- f. Tiene la desviación estándar más pequeña. 18, 0, 0, 0, 0, 0
- g. Tiene la desviación estándar más grande. 1, 2, 9, 12, 12, 18

2. México se encuentra en dos de las siete principales zonas de fenómenos tropicales. Los siguientes datos corresponden a la intensidad (km/h) de los fenómenos que golpearon las costas del país entre los años 2005 y 2011.

Tabla 1.24. Intensidad de los fenómenos tropicales (2005 – 2011).

95	55	160	95	50	55	100	150
55	55	65	185	95	55	55	165
75	150	90	45	45	165	110	55
100	165	65	45	75	85	55	65
130	130	260	85	45	55	205	175
230	130	85	75	215	65	55	65

- a) Organiza los datos en una tabla de distribución de frecuencias.
 b) Construye el histograma y el polígono de frecuencias acumuladas e interprétalos.
3. De acuerdo con la escala Saffir–Simpson, los fenómenos tropicales pueden ser clasificados según su intensidad de la siguiente manera:

Tabla 1.25. Clasificación de acuerdo con la escala Saffir–Simpson.

Categoría	Intensidad Km/h
Depresión tropical	Menos de 62
Tormenta tropical	63 – 118
Huracán categoría 1	119 – 153
Huracán categoría 2	154 – 177
Huracán categoría 3	178 – 208
Huracán categoría 4	209 – 251
Huracán categoría 5	Mayor de 252

- a) Con base en los datos del problema 2, construye los gráficos de barras y de sectores considerando la clasificación de la Tabla 1.25.
 b) Calcula la media, la mediana y la moda de la intensidad de los fenómenos tropicales e interpreta los resultados.
4. Un conjunto de 200 datos se dividió en ocho clases, todas de tamaño 3. Después se determinaron las frecuencias de cada clase y se construyó una tabla de distribución de frecuencias. Sin embargo, ciertas entradas de la tabla se perdieron. Supón que la parte de la tabla que se conservó es la siguiente:

Tabla 1.27. Distribución de frecuencias de 200 datos.

Clase $L_{inf} - L_{sup}$	Frecuencia f_i	Frecuencia relativa fr_i
—		0.05
—	14	
—	18	
15 — 18	38	
—		0.10
—	42	
—	11	
—		

- Completa la tabla y construye el histograma correspondiente.
 - Calcula la media, la mediana y la moda.
 - Determina la desviación estándar.
5. En la siguiente tabla se muestran las temperaturas promedio mensuales (°C) en la Ciudad de México de los años 2017 y 2018.

Tabla 1.26. Temperaturas promedio mensuales 2017 y 2018.

Mes	Temperatura 2017	Temperatura 2018
Enero	15.5	13.4
Febrero	16.9	17
Marzo	17.3	19.2
Abril	19.3	19.3
Mayo	21.3	20.3
Junio	20.3	19.6
Julio	18.7	19
Agosto	19.6	18.2
Septiembre	18.5	18.7
Octubre	17.6	18
Noviembre	16.2	16
Diciembre	14.6	14.9

- a) Representa en un sólo gráfico los diagramas de cajas correspondientes a las temperaturas de cada año.
- b) Calcula la desviación estándar de las temperaturas promedio de cada año.
- c) ¿Cómo se relacionan las respuestas dadas en los incisos a y b?
6. Supón que debes aconsejar a una persona que sufre una enfermedad mortal que puede ser tratada con una droga, la cual podría extender la vida del paciente. Es posible elegir entre tres tratamientos. Las personas muestran efectos secundarios a los medicamentos; mientras que en algunos casos la droga tiene los resultados deseados, en otros los resultados pueden ser más favorables o más adversos. Las siguientes listas muestran el número de años que los pacientes han vivido después de ser tratados con una de las tres opciones; cada número de la lista corresponde al tiempo en años que un paciente ha sobrevivido con el respectivo tratamiento.

Tabla 1.28. Tratamientos.

Tratamiento 1	Tratamiento 2	Tratamiento 3
5.2	6.8	6.8
5.6	6.9	6.8
6.5	6.9	6.9
6.5	7	7
7	7	7
7	7	7.1
7	7.1	7.1
7.8	7.1	7.1
8.7	7.2	7.2
9.1	7.4	7.4

¿Cuál de los tres tratamientos le sugerirías tomar a la persona? Justifica tu respuesta.

7. En una feria se invita a los asistentes a participar en uno de dos juegos, pero no en ambos. Con el propósito de saber en cuál juego participar, Juan observa, toma notas y organiza los resultados de 10 personas que participan en cada uno de los juegos. Las pérdidas (-) y las ganancias (+) obtenidas por 10 personas que participan en el juego 1, y por 10 que participan en el juego 2 son las siguientes:

Juego 1:

15	-21	-4	50	-2	11	13	-25	16	-4
----	-----	----	----	----	----	----	-----	----	----

Juego 2:

120	-120	60	-24	-21	133	-81	96	-132	18
-----	------	----	-----	-----	-----	-----	----	------	----

Si tuvieras la posibilidad de participar en uno de los dos juegos, ¿cuál elegirías?

Justifica tu respuesta.

UNIDAD 2. OBTENCIÓN E INTERPRETACIÓN DE INFORMACIÓN ESTADÍSTICA CON DATOS BIVARIADOS.

Presentación.

En el caso de datos Bivariados, ahora se trabajará con dos características presentadas en dos variables que se observan en una población o muestra, como ejemplo el tiempo que estudias para tus exámenes y la calificación que obtienes, o la superficie y el precio de las viviendas en una ciudad. En ambas situaciones es posible identificar como están relacionadas las dos variables.

La finalidad del estudio de dos variables al mismo tiempo, es observar si existe relación entre ellas o son independientes. Respecto a los métodos que se utilizarán, se debe considerar el tipo de variables en estudio, ya sean cuantitativas o cualitativas.

Propósito.

Analizarás la relación entre dos variables estadísticas y realizarás predicciones, a partir del reconocimiento y la modelación de dicha relación, evaluando el grado de intensidad en ella, con la finalidad de elevar tu capacidad de interpretar y evaluar críticamente la información estadística en dos variables aparejadas.

2.1 Introducción.

Al hacer de manera conjunta la revisión de dos diferentes variables en una población o muestra, estadísticamente tendremos datos bivariados, como se presenta en los siguientes casos:

Ejemplo 2.1

- a) En una Compañía de seguros se elige una muestra del personal registrando el género de cada persona y la localidad en que vive. Como ejemplo de sus respuestas tenemos datos bivariados (hombre, Azcapotzalco), (mujer, Cuajimalpa), (mujer, Xochimilco) y (hombre, Tlalpan).
- b) En una bodega se revisa el peso en kg y la cantidad de artículos que contienen las diferentes cajas que se reciben de un centro comercial. Los registros de esos datos serían. (7, 34), (25, 12), (18, 6) y (40, 2).

Como puedes observar en el primer caso se trata de variables cualitativas y en el segundo son variables de tipo cuantitativo.

2.2 Asociación entre dos Variables Cualitativas.

2.2.1 Tablas de Contingencia

En el caso de dos variables de tipo cualitativo aparejadas, estas se representan en una tabla bidimensional, la cual recibe el nombre de tabla de contingencia cuando las características en estudio no son cuantitativas. Dentro de la tabla, las filas (i) representan las categorías de la primer Variable y las columnas (j) las categorías de la segunda Variable, como se muestra en la Tabla 2.1.

Tabla 2.1. Ubicación de la Variable 1 y Variable 2

Variable 1 \ Variable 2	Columna 1	Columna 2	TOTAL
Fila 1		(1,2)	
Fila 2			
TOTAL			n

A cada celda se le reconoce identificando primero la fila y luego la columna en la que se ubica (i, j), por ejemplo en la Tabla 2.1, la celda sombreada es la (1,2) ya que se localiza en la fila 1, columna 2.

Ejemplo 2.2

En la Tabla 2.2, se muestran los registros de solicitudes a cierto bachillerato, clasificadas por promedio escolar y el tipo de secundaria de procedencia de los estudiantes.

Las variables presentadas son:

- Variable 1, Promedio escolar en las categorías “Bajo”, “Regular” y “Alto”
- Variable 2, Tipo de secundaria con categorías “Diurna”, “Técnica” y “Privada”

Tabla 2.2. Representación de dos variables cualitativas.

Promedio escolar	Tipo de secundaria		
	Diurna	Técnica	Privada
Bajo	263	184	108
Regular	419	522	396
Alto	637	760	845

Podemos distinguir que hubo 263 estudiantes con Bajo promedio escolar procedentes de secundaria diurna, mientras que 845 estudiantes procedían de secundaria privada y tenían un promedio escolar alto.

Al sumar las frecuencias absolutas de cada fila y de cada columna, se obtienen las **frecuencias absolutas marginales**, como se observa en la Tabla 2.3.

Tabla 2.3. Tabla de frecuencias marginales.

Promedio escolar	Tipo de secundaria			TOTAL
	Diurna	Técnica	Privada	
Bajo	263	184	108	
Regular	419	522	396	
Alto	637	760	845	1505
TOTAL	1319			

Al realizar los cálculos, vemos que:

- 1319 estudiantes habían cursado la secundaria diurna.
- 1505 estudiantes del total de los solicitantes tenían promedio escolar alto.

Ejercicio 2.1



Completa las sumas de la Tabla 2.3 y contesta las siguientes preguntas:

- _____ estudiantes tenían promedio escolar regular.
- _____ estudiantes procedían de secundaria técnica.
- _____ estudiantes tenían promedio escolar bajo.
- El total de solicitudes recibidas fue:_____

La representación de datos bivariados de variables cualitativas en una tabla de contingencia, nos permite observar de forma más puntual características de los datos.

Retomando el ejemplo anterior:

- La mayor cantidad de estudiantes con un promedio alto procedían de una escuela secundaria privada.
- Estudiantes de secundaria técnica con promedio bajo, fueron la menor cantidad de solicitantes.

Ahora bien, si se quiere visualizar la información contenida en la tabla de contingencia bajo un análisis diferenciado, es recomendable la elaboración de otros tres tipos de tablas adicionales, tabla de frecuencias relativas, tabla de porcentajes por fila y tabla de porcentajes por columna.

2.2.2 Tablas de Frecuencias relativas

Para el cálculo de las frecuencias relativas de cada celda, lo que vamos a hacer es dividir el valor de cada celda entre el número total de datos (n) y multiplicar el resultado por cien para expresarlo en porcentaje; como se observa en la Tabla 2.4.

Ejemplo 2.4

Para la celda (3,2) alumnos de secundaria técnica con promedio alto.

$$f.r.(3,2) = \frac{760}{4134}(100) = 18.38 \%$$

De esta manera es que se calcularon los valores de las otras celdas.

Tabla 2.4. Tabla de frecuencias relativas.

Promedio escolar	Tipo de secundaria			TOTAL
	Diurna	Técnica	Privada	
Bajo	6.36 %	4.45 %	2.61 %	13.42 %
Regular	10.14 %	12.63 %	9.58 %	32.35 %
Alto	15.41 %	18.38 %	20.44 %	54.23 %
TOTAL	31.91 %	35.46 %	32.63 %	100 %

Respecto a la descripción de los valores obtenidos, recuerda que se deben mencionar ambas categorías que intersectan en cada celda, por ejemplo:

- Celda (1,3), el **2.61 %** de los estudiantes solicitantes tenían promedio escolar bajo y cursaron la secundaria en una institución privada.
- Celda (2,1), el **10.14 %** de los estudiantes que presentaron solicitud, tuvieron un promedio escolar regular con estudios de secundaria en escuela diurna.

En el caso de las celdas donde se ubican los totales de fila o columna, estos porcentajes se nombran “**frecuencias marginales**”, como ejemplo:

- **54.23 %** del total de los estudiantes tenían un promedio escolar alto.
- **31.91 %** del total de los solicitantes procedían de secundaria diurna.



Ejercicio 2.2

Con la información de la Tabla 2.4, completa los espacios en blanco de las siguientes afirmaciones:

- _____ de los estudiantes tenían promedio escolar bajo con estudios de secundaria en escuela privada.
- _____ de los solicitantes obtuvieron promedio escolar regular con estudios en escuela secundaria técnica.
- 32.63 % de los estudiantes en total procedían de escuela _____

2.2.3 Tablas de porcentajes por fila

Para este tipo de tabla, lo que haremos es calcular el valor de cada celda respecto al valor total (100 %) por fila, así en cada celda tendremos el porcentaje que le corresponde considerando cada fila como un total parcial es decir cada una de las categorías de la variable 1 (bajo, regular y alto), por ejemplo en la celda (1, 3) el valor se obtiene dividiendo el valor original de la celda (108) en la Tabla 2.2, entre el total de la fila 3 (555) y al final multiplicando por cien, como se muestra en la Tabla 2.5.

$$\text{porcentaje por fila (1, 3)} = \frac{108}{555} (100) = 19.46 \%$$

Tabla 2.5. Tabla de porcentajes por fila.

Promedio escolar	Tipo de secundaria			TOTAL
	Diurna	Técnica	Privada	
Bajo			19.46 %	100 %
Regular			29.62 %	100 %
Alto	28.41 %			100 %
TOTAL				100 %

Observaciones de la Tabla 2.5:

- De los estudiantes con un promedio escolar regular el 31.34 % estudiaron en secundaria diurna, 39.04 % en secundaria técnica y el 29.62 % eran procedentes de secundaria privada.



Ejercicio 2.4

Completa la información de la Tabla 2.6, y responde en los espacios en blanco las siguientes afirmaciones:

- De los estudiantes procedentes de secundaria diurna _____ tuvieron promedio bajo, _____ un promedio regular y _____ promedio escolar alto.
- De los estudiantes procedentes de secundaria privada _____ tuvieron promedio bajo, _____ un promedio regular y _____ promedio escolar alto.

Evaluación tema 1. Unidad 2. Asociación entre dos Variables Cualitativas.

1. ¹Se revisó la información respecto al área de conocimiento para estudios de Licenciatura en la UNAM, que eligieron las alumnas y alumnos egresados de CCH plantel Vallejo generación 2015-2018, los datos se muestran en la Tabla 2.7.

- a) Completa la siguiente tabla de contingencia y construye las tablas de frecuencias relativas, de porcentajes por renglón y de porcentajes por columnas.

¹ Muñoz, L. (2019). Estudio de las Trayectorias escolares del Colegio de Ciencias y Humanidades, Generaciones 2013, 2014, 2015, 2016, 2017 y 2018. México. UNAM. Dirección General Colegio de Ciencias y Humanidades. (p. 158).

Tabla 2.7. Asignación de las tres carreras más solicitadas por plantel, generación 2017 CCH.

Carrera	Plantel					TOTAL
	Azcapotzalco	Naucalpan	Vallejo	Oriente	Sur	
Derecho	287	280	321	339	315	
Médico Cirujano	231	219	212	290	210	
Psicología	221	193	191	259	199	
TOTAL						

b) Con la información de las tablas, responde las preguntas y completa los enunciados siguientes.

- _____ personas eligieron Derecho.
- El total de alumnos del plantel Sur de la generación 2017, fue _____
- ¿Cuántos alumnos del plantel Vallejo eligieron Psicología? _____
- ¿Qué porcentaje de alumnos eligieron Médico Cirujano y fueron del plantel Azcapotzalco? _____
- ¿Qué porcentaje de alumnos fueron del plantel Oriente? _____
- ¿Qué porcentaje de alumnos no eligieron Derecho? _____
- De los alumnos registrados del plantel Naucalpan, el _____ % eligió estudios de Psicología
- De los alumnos que eligieron Derecho, ¿qué porcentaje eran del plantel Sur?

- ¿Cuál de las carreras tiene el mayor porcentaje de alumnos, del total de los registrados y de que plantel fueron? _____

2. Se realizó una encuesta a 400 personas para analizar cómo influyen en la salud los hábitos relacionados con el tabaquismo. Se obtuvieron los siguientes resultados:

- 102 personas fuman mucho y tienen problemas respiratorios.
- 35 personas tienen un nivel moderado de tabaquismo y no tienen problemas respiratorios.
- El 55.25% del total tiene problemas respiratorios.
- Del total, el 30% fuma mucho, mientras que el 25% fuma de forma moderada, otro 25% fuma poco y sólo el 20% no fuma.

- 77 personas no tienen problemas respiratorios y no fuman.
 - a) Elabora la tabla de contingencia y construye las tablas de frecuencias relativas, de porcentajes por renglón y de porcentajes por columnas.
(Escribe el procedimiento completo para el cálculo de los valores de cada una de las celdas)
 - b) Describe debajo de cada tabla, por lo menos la información de dos celdas, filas y columnas.
 - c) Escribe tus conclusiones finales, después de analizar las dos variables en sus diferentes categorías.
 - d) Comenta con tus compañeros de grupo, tus resultados y conclusiones.

2.3 Variables Cuantitativas

Existen situaciones de la vida cotidiana que requieren analizarse mediante el comportamiento de dos variables de tipo cuantitativo.

Las variables en estudio se deben identificar como Variable Independiente “X” y Variable dependiente “Y”, con la intención de observar adecuadamente cuál de ellas determina el comportamiento de la otra.

En Estadística el concepto que define la presencia de relación entre variables cuantitativas es “correlación”. Para poder determinar la relación que mantienen las dos variables, se emplea una representación gráfica y un método numérico.

Un ejemplo de análisis de la relación entre dos variables ocurre cuando a los niños en su etapa de crecimiento se les da seguimiento con los datos de la talla y peso que se registran durante los primeros 12 años de vida.

Como se mencionó anteriormente, el primer paso será establecer cuál de las dos variables es la independiente “X” y cuál la dependiente “Y”.

2.3.1 Diagrama de Dispersión.

En nuestro caso estableceremos que el comportamiento de la talla de los niños debe reflejarse en el peso correspondiente, así que denominaremos:

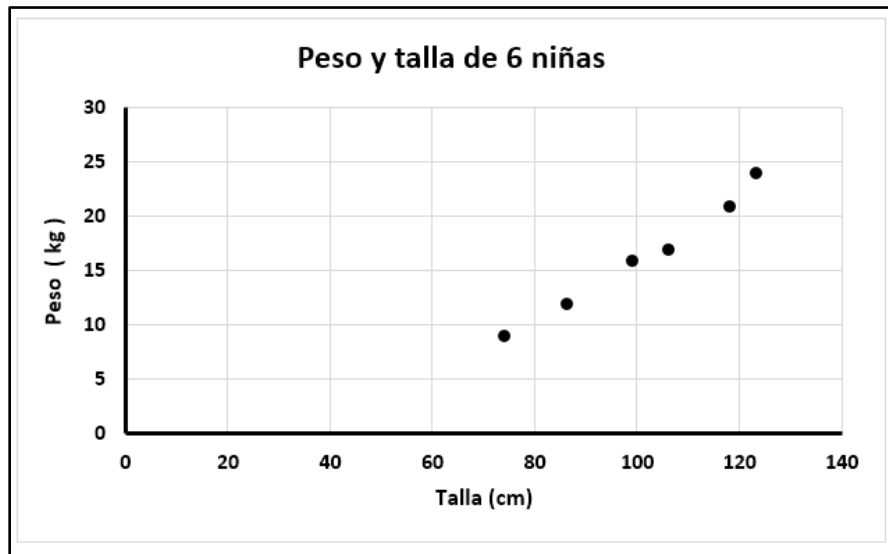
Talla (cm) “X” variable independiente Peso (kg) “Y” variable dependiente

En la Tabla 2.8 se presentan los datos de peso (kg) y talla (cm), de 6 niñas de entre 3 y 11 años.

Tabla 2.8 Datos de Peso y Talla.

X Talla (cm)	Y Peso (kg)
86	12
123	24
74	9
106	17
118	21
99	16

Iniciaremos representando mediante parejas ordenadas (x, y), el comportamiento de las variables en una gráfica de dispersión, el cual se muestra en la Gráfica 1.



Gráfica 1. Gráfica de dispersión.

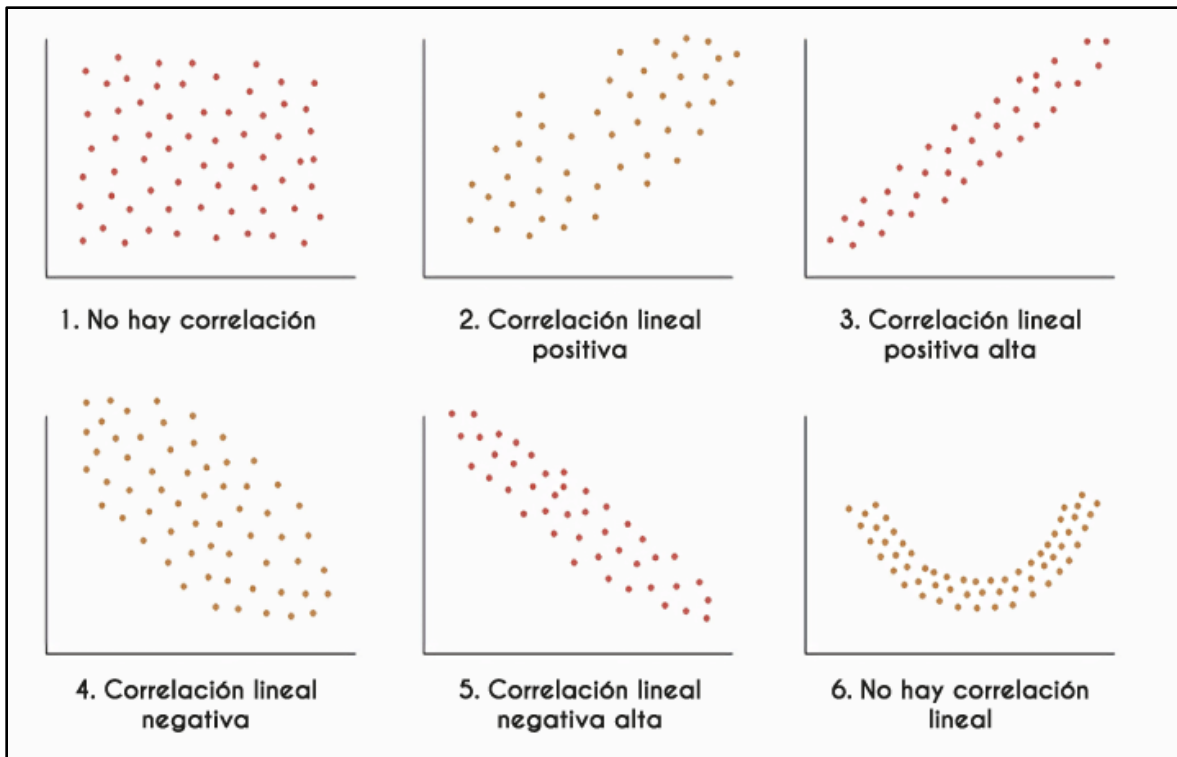


Figura 1. Nube de puntos.

Para conocer si las variables de estudio están relacionadas, se toma como base un comportamiento lineal en el agrupamiento de la nube de puntos, como se muestra en la figura 1. En nuestro ejemplo, podemos observar que la nube de puntos presenta una forma muy aproximada a una línea recta, lo que indica que existe una relación de grado o intensidad fuerte entre las variables peso y talla de $n = 6$ niñas.

El diagrama de dispersión es una herramienta que ayuda a identificar la posible relación entre dos variables. Representa la relación entre dos variables de forma gráfica, lo que hace más fácil visualizar e interpretar los datos.

2.3.2 Coeficiente de correlación lineal de Pearson.

Si tenemos dos variables cuantitativas y deseamos medir el grado de asociación, podemos utilizar el coeficiente de correlación lineal de Pearson. Este coeficiente se representa con la letra “r” y puede tomar valores entre -1 y +1, de modo que si el signo en el valor de “r” es positivo, nos indica que al aumentar el valor de la variable independiente “X” también aumenta el valor de la variable dependiente “Y”. En caso

contrario, si el signo en el valor de “r” es negativo, al aumentar el valor de la variable independiente “X”, disminuye el valor de la variable dependiente “Y”.

Para calcular el valor del coeficiente de correlación, se requiere que calcules los valores de cada una de las tres columnas que se agregaron a la Tabla 1., además de obtener las sumatorias de cada una de las 5 columnas como se presenta en la Tabla 2.9

Tabla 2.9 Cálculo de X^2 , Y^2 , $(X)(Y)$ y Sumatorias.

X Talla (cm)	Y Peso (kg)	X²	Y²	(X)(Y)	
86	12	7396	144	1032	
123	24	15129	576	2952	
74	9	5476	81	666	
106	17	11236	289	1802	
118	21	13924	441	2478	
99	16	9801	256	1584	
Σ	606	99	62962	1787	10514

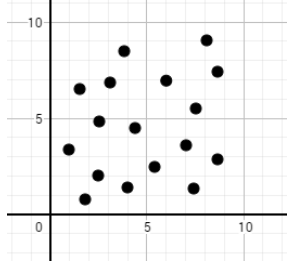

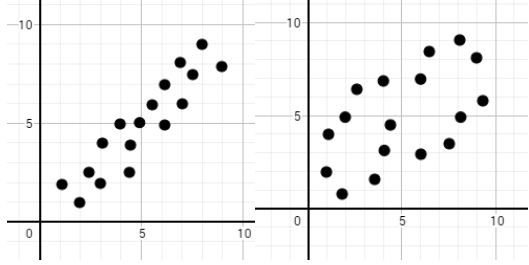
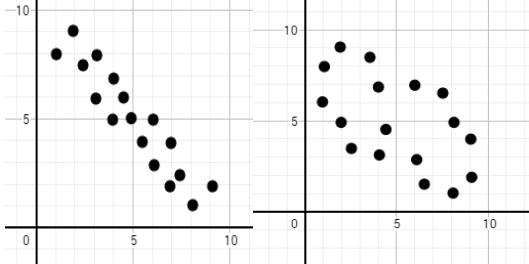
$$r = \frac{[(n)(\sum xy)] - [(\sum x)(\sum y)]}{\sqrt{[(n)(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$r = \frac{[(6)(10514)] - [(606)(99)]}{\sqrt{[(6)(62962) - (606)^2][n(1787) - (99)^2]}}$$

$$r = \frac{[63084] - [59994]}{\sqrt{[377772 - 367236][10722 - 9801]}}$$

$$r = \frac{3090}{\sqrt{(10536)(921)}} = \frac{3090}{\sqrt{9703656}} = \frac{3090}{3115.0692} = \mathbf{0.9920}$$

Por lo tanto, se trata de una correlación positiva de grado fuerte entre las variables talla y peso de 6 niñas. Al ser positivo el signo del coeficiente “r”, la variación es directamente proporcional, a medida que aumenta la talla aumenta el peso.

<p>Un valor cercano o igual a 0 indica respectivamente poca o ninguna relación lineal entre las variables.</p>	
<p>Cuanto más se acerque en valor absoluto a 1, mayor será el grado de asociación lineal entre las variables.</p>	
<p>Un coeficiente positivo indica asociación lineal positiva, es decir, tienden a variar en el mismo sentido.</p>	
<p>Un coeficiente negativo indica asociación lineal negativa, es decir, tienden a variar en sentido opuesto.</p>	

Respecto a la intensidad o grado de correlación, este se determina de acuerdo al valor numérico de “r “. La correlación será perfecta si $r = \pm 1$, esto se muestra en la Tabla 2.10.

Tabla 2.10 Tipo y Grado de correlación, de acuerdo al valor de “r”.

<i>r</i>	Tipo de correlación	Grado o intensidad de la correlación
- 1.0	Negativa	Perfecta
- 0.9		Fuerte
- 0.5		Moderada
- 0.1		Débil
0	No hay correlación	
0.1	Positiva	Débil
0.5		Moderada
0.9		Fuerte
1.0		Perfecta

2.3.3 Mínimos Cuadrados.

Es una técnica de análisis numérico enmarcada dentro de la optimización matemática, en la que, dados un conjunto de pares ordenados, variable independiente (X) y variable dependiente (Y), se intenta encontrar la función que mejor se aproxime a los datos (un “mejor ajuste”).

El principio de los mínimos cuadrados con el supuesto más sencillo y elemental con una sola ecuación y que ésta contiene sólo dos variables **X** y **Y**, consideramos que:

$$Y = f (X)$$

Este paso simplemente identifica a la variable **X**, la cual se considera que influye sobre la otra variable **Y**.

El segundo paso consiste en especificar la forma de la relación entre **Y** y **X**. La relación más simple entre dos variables es la línea recta, es decir,

$$Y = m X + b$$

donde:

m y **b** son parámetros desconocidos que indican la pendiente y la ordenada en el origen de la función, respectivamente.

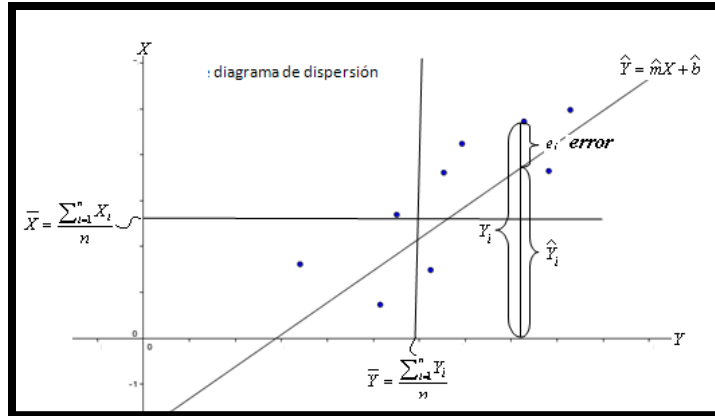


Figura 2.

Una **recta que mejor se ajusta** es una línea recta que es la mejor aproximación del conjunto de datos dado.

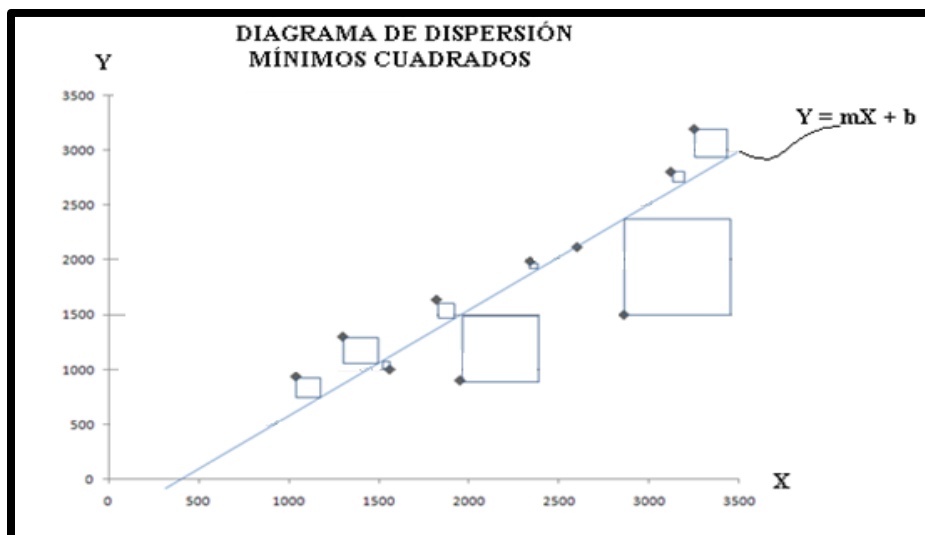
Es usada para estudiar la naturaleza de la relación entre dos variables.

Una recta que mejor se ajusta puede ser determinada aproximadamente usando el método visual al dibujar una línea recta en una **gráfica de dispersión** para que tanto el número de puntos arriba de la recta y debajo de la recta sean casi iguales (y la línea pasa a través de tantos puntos como sea posible, Figura 2).

$$X_1, X_2, \dots, X_n$$

$$Y_1, Y_2, \dots, Y_n$$

Si se tiene n puntos, $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, una recta $Y = mX + b$ recibe el nombre de recta de mínimos cuadrados.



La recta de mejor ajuste y por lo tanto la que da la mejor estimación es aquella en la suma del área de los cuadrados sea las más pequeña o mínima.

Las fórmulas para calcular la recta de mínimos cuadrados son las siguientes:

Pendiente.

$$m = \frac{[(n)(\sum xy)] - [(\sum x)(\sum y)]}{[(n)(\sum x^2) - (\sum x)^2]}$$

El signo del valor calculado de la pendiente, también nos da información respecto al comportamiento de las variables, como se muestra en la Figura 3.

Signo de la pendiente de la recta "m"	Tipo de variación entre las dos variables	
-	Variación inversamente proporcional	A medida que. aumentan los valores de la variable "X", los valores de la variable "Y" disminuyen.
+	Variación directamente proporcional	A medida que: aumentan los valores de la variable "X", los valores de la variable "Y" aumentan.

Figura 3. Tipo de variación de acuerdo al signo de la pendiente.

Ordenada en el origen

$$b = \frac{[(\sum y)] - [(m) (\sum x)]}{n}$$

Continuando con el ejercicio del peso y talla de 6 niñas, sustituyendo los valores correspondientes obtenemos:

Pendiente

$$m = \frac{[(6)(10514)] - [(606)(99)]}{[(6)(62962) - (606)^2]}$$

$$m = \frac{[63084] - [59994]}{[377772 - 367236]} = \frac{3090}{10536} = \mathbf{0.2933}$$

Ordenada en el origen

$$b = \frac{[(99)] - [(0.2933)(606)]}{6} = \frac{99 - 177.7398}{6}$$

$$b = \frac{-78.7398}{6} = -\mathbf{13.1233}$$

La ecuación de mejor ajuste será:

$$Y = \mathbf{0.2933 X + (-13.1233)}$$

Para trazarla en el gráfico de Dispersión, sustituiremos en la ecuación anterior dos valores opcionales de Talla (X) que no aparezcan en la Tabla 1; y así obtener el valor correspondiente de Y.

Tendremos dos puntos (x, y) que se requieren para dibujar la recta de mejor ajuste.

X = 80	X = 105
$Y = \mathbf{0.2933 (80) + (-13.1233)}$ $Y = \mathbf{23.4640 + (-13.1233)}$ $Y = \mathbf{10.3407}$	$Y = \mathbf{0.2933 (105) + (-13.1233)}$ $Y = \mathbf{30.7965 + (-13.1233)}$ $Y = \mathbf{17.6732}$
Punto A (80, 10.3407)	Punto B (105, 17.6732)



Gráfica 2. Diagrama de Dispersión con Recta de mejor ajuste.

Como vemos en la Gráfica 2, la recta de mejor ajuste intersecta al eje “Y” en un valor de – 13.1233 que es el mismo valor que obtuvimos de “b” al emplear las fórmulas correspondientes.

Respecto a la cercanía de los puntos de las parejas ordenadas de los datos originales con la recta de mejor ajuste, comprobamos que existe una correlación de tipo fuerte, ya que la distancia entre ellos es mínima.

2.3.4 Regresión lineal

Uno de los aspectos más relevantes de la Estadística es el análisis de la relación o dependencia entre variables. Frecuentemente, resulta de interés conocer el efecto que una variable puede causar sobre otra, e incluso predecir en mayor o menor grado valores en una variable a partir de otra. Para esto emplearemos la ecuación de mejor ajuste o modelo matemático.

$$Y = m X + b$$

Para nuestro ejemplo, prediciremos el peso que debería tener una niña cuya talla es de 92 cm.

$$Y = 0.2933 X + (-13.1233)$$

$$Y = 0.2933 (92) + (-13.1233)$$

$$Y = 26.9836 + (-13.1233)$$

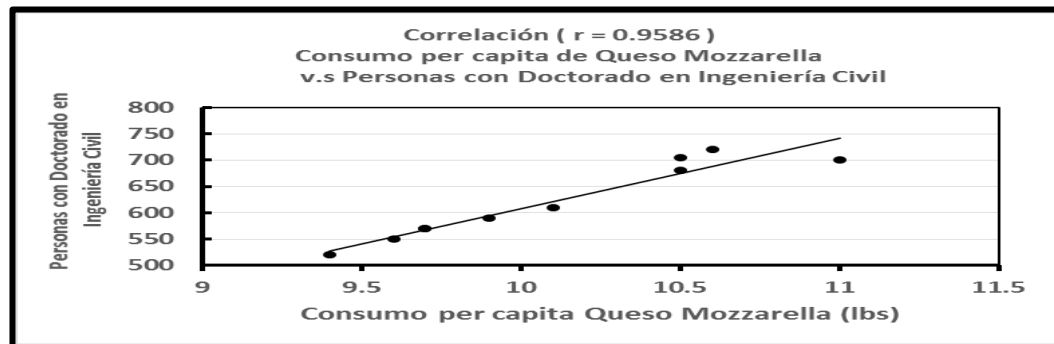
$$Y = 13.8603 Kg$$

El peso de una niña que tiene una talla de 92 cm, debería ser de 13.8303 kg.

Es importante considerar que este modelo tiene un límite de estimación de acuerdo al tipo de variables que se estén analizando.

Otro aspecto a destacar es que la relación de ciertas variables de tipo cuantitativo puede no ser producto de la causalidad.

En los Estados Unidos de Norteamérica la Fundación Nacional de Ciencia y el Departamento de Agricultura, generaron datos respecto al número de doctorados en Ingeniería civil del país y el consumo per cápita de queso mozzarella. Se consideró dicha información de los años 2000 a 2009. Aplicando el cálculo del coeficiente de correlación lineal de Pearson se obtuvo un valor de $r = 0.9586$, lo que indica que existe relación positiva entre las variables de grado fuerte. Sin embargo se aprecia a simple vista que el comportamiento de una no guarda relación real con la otra.



Fuente: <https://tylervigen.com/spurious-correlations>

Evaluación tema 2. Unidad 2. Asociación entre dos Variables Cuantitativas

I. Para cada pareja de datos en cada uno de los ejercicios siguientes:

- Calcula e interpreta el valor del Coeficiente de correlación lineal de Pearson (r)
- Obtén la ecuación de la recta de mejor ajuste
- Traza el Diagrama de Dispersión y la recta de mejor ajuste

Calcula el valor de la Variable "Y" de acuerdo al valor propuesto de "X"

# materias que se adeudan	15	0	8	1	3	12	5
Promedio	6.7	9.5	8.3	9.6	8.1	6.9	7.4

Máquinas empleadas	1.5	11	9	15	7
Productos elaborados	187	593	325	1024	280

II. En cada una de las siguientes tablas se presenta información de la relación entre dos variables cuantitativas. Describe el comportamiento de las variables.

1.

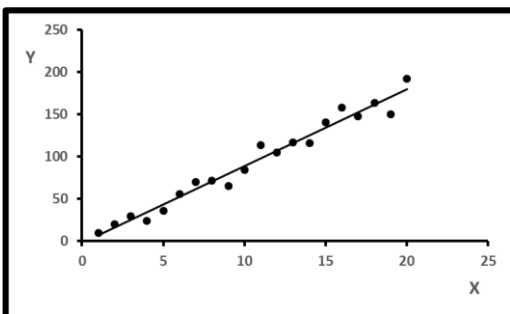
Coefficiente de correlación lineal de Pearson	$r = - 0.83$
Ecuación de la recta de mejor ajuste	$y = - 42.16 (x) + 9$

2.

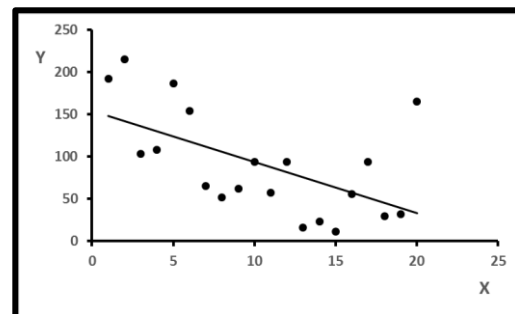
Coefficiente de correlación lineal de Pearson	$r = 0.91$
Ecuación de la recta de mejor ajuste	$y = 6.75 (x) - 4.8$

III. Coloca de lado derecho de cada diagrama de dispersión, el valor del Coeficiente de Correlación lineal (r) que corresponde de la siguiente tabla:

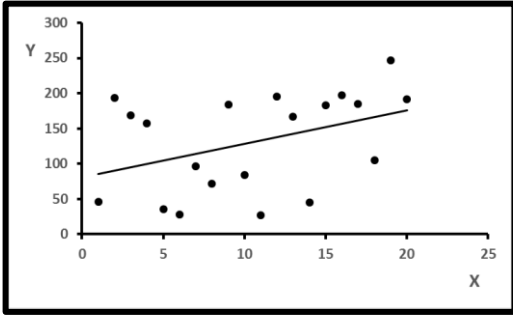
$r = - 0.57$	$r = - 0.72$	$r = 0.98$	$r = 0.40$
--------------	--------------	------------	------------



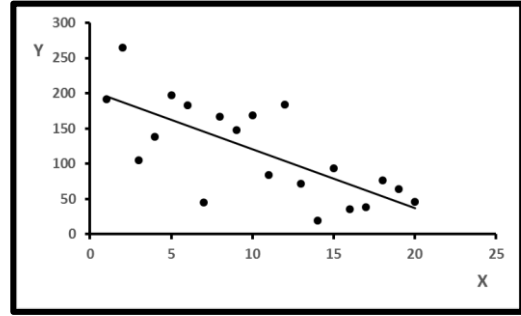
()



()



()



()

UNIDAD 3. Probabilidad Azar: modelación y toma de decisiones.

Presentación

En esta unidad se pretende introducir al estudiante a comprender la idea de azar y manejar por ende el concepto de probabilidad. El propósito es tratar que el estudiante se introduzca poco a poco en este tema, que resulta, en algunas ocasiones difícil de asimilar y comprender. Se mencionan algunos elementos básicos de Teoría de Conjuntos y de Técnicas de Conteo. Se recomienda como estrategia de aprendizaje, métodos que propicien la socialización del trabajo y la discusión de las ideas probabilísticas con la resolución de problemas.

Propósito

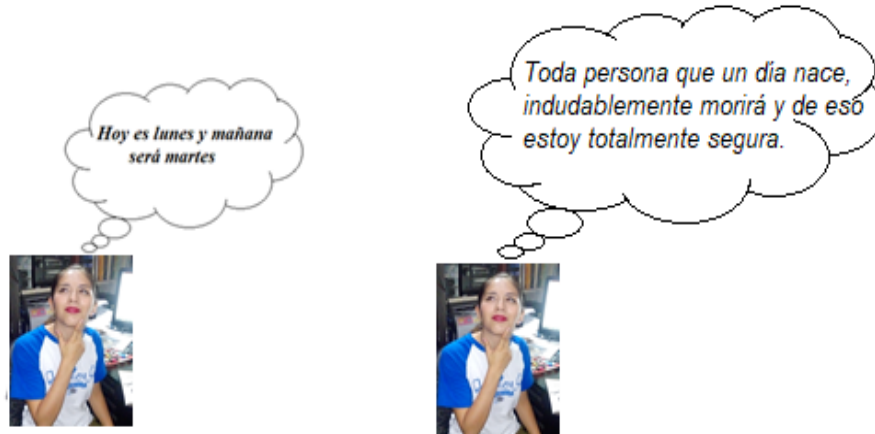
Al finalizar la unidad el alumno: Continuará el desarrollo de su pensamiento estadístico, a través del conocimiento y modelación de los fenómenos aleatorios, desde los tres enfoques de la probabilidad, incluyendo la toma de decisiones.

3.1 Fenómenos determinísticos y aleatorios

Pero, ¿qué es un fenómeno o experimento? Un fenómeno o experimento, es la acción que se realiza, con el propósito de analizar dicha acción y tiene como objetivo final, determinar la probabilidad de uno o varios resultados. Existen dos tipos de fenómenos o experimentos:



Fenómenos o Experimentos Deterministas en estadística, es aquel que bajo el mismo conjunto aparente de condiciones iniciales, puede presentar resultados iguales, es decir, se puede predecir el resultado, por ejemplo:



Fenómenos o Experimentos Aleatorios. En estadística, es aquel que bajo el mismo conjunto aparente de condiciones iniciales, puede presentar resultados diferentes, es decir, no se puede predecir el resultado.



3.2 Espacio muestral y diferentes tipos de eventos

Para el estudio de la probabilidad, es necesario tener en cuenta dos conceptos muy importantes, el **Espacio Muestral** que es denotado por Ω y se define como **el conjunto de todos los posibles resultados de interés de un experimento aleatorio y Evento o suceso** que se denota con las letras mayúsculas A, B, C , ... y **es un subconjunto de un espacio muestral**. Al utilizar diagramas de Venn Euler, el espacio muestral, es simbolizado por un rectángulo y se denota con la letra griega Ω (Figura. 3.1).



Figura 3.1 Espacio Muestral

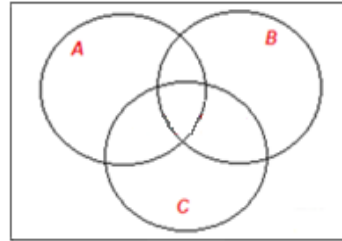


Figura 3.2 Eventos

Los eventos, se denotan con una serie de círculos y se les asignan las letras mayúsculas A,B,C,... y estos, son colocados dentro del Espacio Muestral (Figura.3.2), que como antes se mencionó corresponde al espacio muestral. Los eventos pueden ser vistos como:

a) **Evento Seguro:** Es aquel cuya **probabilidad** de ocurrencia es igual a 1.

Por ejemplo, ¿Cuál es la probabilidad de que al lanzar un dado, el número de los puntos de las caras que caen hacia arriba sea menor que siete? En este caso el espacio muestra es: $\Omega = \{1,2,3,4,5,6\}$ y si se define al evento A como el evento de que el número de puntos de la cara que cae hacia arriba sea menor que siete entonces:

$$A = \{1,2,3,4,5,6\}$$

por lo tanto, la probabilidad de que caigan 1, 2, 3, 4, 5, o 6 puntos al lanzar un dado es la certeza o 1.

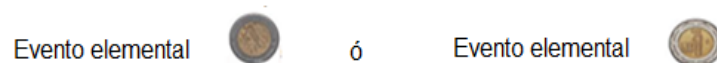
b) **Evento Imposible:** Es aquel del que se tiene la seguridad de no ocurrencia.

Por ejemplo, si definimos a B como el evento de los puntos que caen en la cara que cae hacia arriba, al lanzar un dado, sea mayor que siete. Es obvio que en el dado, la cara que tiene mayor número de puntos es seis, es imposible que caiga una cara donde el número de puntos sea mayor. Por lo tanto, la probabilidad sería igual a cero.

c) **Evento Elemental:** Es un subconjunto del espacio muestral que contiene un solo elemento y que no puede ser desglosado en otros resultados.

Es decir, si se definen los eventos elementales como:

- Sea A el evento de que al lanzar una moneda caiga águila. Y otro evento como que caiga sol.



- Sea B el evento de que al lanzar un dado, caiga cualquiera de las caras.



A cada evento se le llama evento elemental.

3.3 Enfoques de la probabilidad

También es muy importante hacer mención de los tres enfoques de probabilidad.

a) Probabilidad Subjetiva: Es la probabilidad que se tiene por experiencia; se puede definir como la probabilidad asignada a un evento por parte de un individuo, basada en la evidencia. Por ejemplo:

¿Qué probabilidad hay de que hoy llueva?

Cuándo nos hacen esta pregunta, comúnmente observamos el cielo y vemos si hay nubes negras o pensemos, quizás, si el mes en que estamos es lluvioso o no y probablemente de acuerdo a esa experiencia, podamos dar una posible probabilidad, pero no necesitamos realizar ningún calculo aritmético o matemático, es decir, la probabilidad asignada en este caso, sólo se da por la experiencia que tenemos.

b) Probabilidad Frecuencial Es la que se fundamenta en los datos obtenidos por una serie larga de realizaciones de un experimento. Para determinar la probabilidad frecuencial, se repite el experimento aleatorio un número determinado de veces y se registran los datos. Supongamos que se desconfía de la legalidad de una moneda, con la que se están jugando volados, por lo que se decide lanzarla 100 veces y ver con qué frecuencia aparece el sol y el águila (Figura. 3.3).

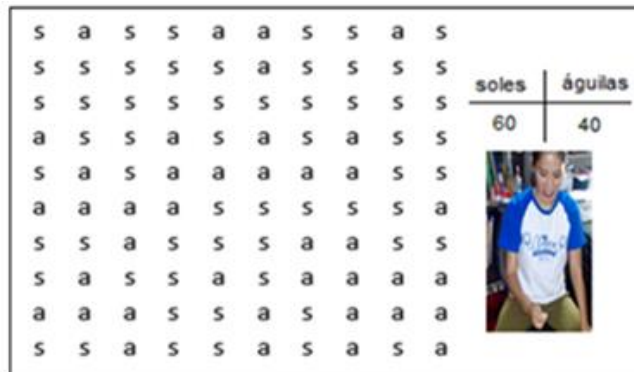


Figura 3.3 100 lanzamientos de una moneda

En la figura 3.3 se pueden observar los resultados obtenidos en el lanzamiento de una moneda 100 veces, cayeron 60 soles y 40 águilas, por lo que la probabilidad de que caiga sol es $P(sol) = \frac{60}{100} = \frac{3}{5}$ que es relativamente normal, pero si cayeran 20 soles y 80 águilas o viceversa, se tendría que desconfiar de la moneda y la probabilidad de sol sería $\frac{20}{100} = \frac{1}{5}$.

c) Probabilidad Clásica

Si todos los resultados en un espacio muestral Ω finito son igualmente probables, y A es un evento en ese espacio muestral, entonces la probabilidad clásica del evento A, expuesta por Pierre Laplace en su famosa Teoría analítica de la probabilidad publicada en 1812, está dada por la siguiente fórmula:

$$P(A) = \frac{\text{número de resultados favorables}}{\text{número total de posibles resultados}} = \frac{n(A)}{n(\Omega)}$$

Por ejemplo, en cierta rifa de un automóvil se venden 5000 boletos. ¿Cuál es la probabilidad de ganarse el automóvil?

a) Si se compran 20 boletos.

En este caso, el número de elementos del espacio muestral es $n(\Omega) = 5000$ y si definimos al evento A como, el evento de comprar de 20 boletos, entonces $n(A) = 20$, por lo tanto:

$$P(A) = \frac{n(A)}{n(\Omega)} = \frac{20}{5000} = \frac{1}{250} = 0.004 = 0.4\%$$

b) Si se compran todos los boletos.

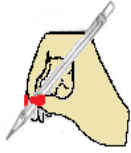
Si definimos el evento B como el de comprar todos los boletos, entonces:

$$P(B) = \frac{n(B)}{n(\Omega)} = \frac{5000}{5000} = 1 = 100\%$$

c) Si no se compran boletos (no se participa en la rifa).

Si C es el evento de no comprar boletos, entonces:

$$P(C) = \frac{n(C)}{n(\Omega)} = \frac{0}{5000} = 0$$



1. Con tus propias palabras ¿qué entiendes por probabilidad?

2. ¿Qué enfoques de probabilidad conoces? _____

3. ¿Qué se entiende por Espacio Muestral y cómo se simboliza? _____

4. ¿Qué se entiende por evento y cómo son simbolizados? _____

3.4 Cálculo de probabilidades de eventos simples y compuestos

Cuando se inicia el estudio de la probabilidad, es muy común utilizar juegos de azar, como el lanzamiento de monedas, de dados o utilizando barajas, sin embargo, los alcances de la probabilidad son muchísimo más, es utilizada, en empresas dedicadas a diversas actividades en la vida real, por ejemplo, las compañías de seguros, los aeropuertos, los bancos, etc. Además de que, en muchas empresas se realizan estudios de factibilidad o estudios de mercado, donde la Estadística y Probabilidad tienen una participación relevante. Comenzaremos con un experimento muy elemental, el lanzamiento de una moneda.

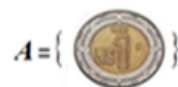
Ejemplo 3.1

Se lanza una moneda. Sea A el evento de que al lanzar una moneda caiga sol.

Solución

Aplicando la probabilidad clásica, debemos encontrar el número de elementos del Espacio Muestral y el número de elementos del evento A, que se conoce como la cardinalidad de Espacio Muestra y del evento A, respectivamente. Se tiene:

$\Omega = \{a, s\}$ Espacio Muestral, su cardinalidad es igual $n(\Omega) = 2$ y $A = \{s\}$ Evento A, su cardinalidad es igual $n(A) = 1$, es decir:



por lo tanto, la probabilidad del evento A es:

$$p(A) = \frac{n(A)}{n(\Omega)} = \frac{1}{2}$$

Ejemplo 3.2

Se lanza un dado. Sea A el evento de que al lanzar un dado, apareciera en la cara que cae hacia arriba 1 o 2 puntos. Cuál es la probabilidad A.

El espacio muestral es $\Omega = \{ \text{🎲}, \text{🎲}, \text{🎲}, \text{🎲}, \text{🎲}, \text{🎲} \}$ y el evento A $A = \{ \text{🎲}, \text{🎲} \}$

Por lo tanto:

$$p(A) = \frac{n(A)}{n(\Omega)} = \frac{2}{6}$$

En los ejemplos anteriores se han utilizado eventos elementales, sin embargo, una buena parte del cálculo de probabilidades consiste en encontrar la probabilidad de dos o más eventos elementales relacionados entre sí por medio de operaciones llamadas Unión, Intersección y Complemento. Se acostumbra denotar a los de dos maneras diferentes, por **extensión**, donde es tomado en cuenta a cada uno de los elementos que constituyen el evento. Por otra parte, por **compresión**, donde se toma en cuenta, la propiedad o el criterio de agrupación por el cual se establece el evento. Es decir, cuando tomado un elemento del evento, que llamaremos x, y se mencionan las características que debe cumplir dicho elemento para poder ser parte del evento. Por ejemplo, si se tienen dos eventos A y B, se definen la unión de A y B como AUB.

$$A \cup B = \{x | x \in A \text{ ó } x \in B\} \quad \text{por compresión}$$

El evento compuesto $A \cup B$ se muestra por comprensión, es decir, es el conjunto de elementos x tal que x pertenece a A ó x pertenece a B, y es posible expresarlo de dos maneras, cuando los eventos son mutuamente exclusivos (Figura. 3.4) o cuando no lo son (Figura. 3.5). Expresado en diagramas de Venn Euler.

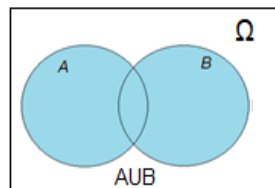


Figura 3.4 Eventos no Disjuntos

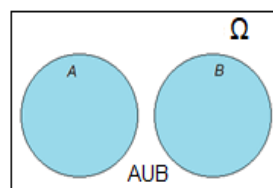


Figura 3.5 Eventos Disjuntos

La intersección de A y B se denota como $A \cap B$ es el evento compuesto de elementos x tal que el elemento x pertenece al evento A y también x pertenece al evento B.

$$A \cap B = \{x \mid x \in A \text{ y } x \in B\}$$

Y de igual manera puede expresarse, cuando los eventos son mutuamente exclusivos (Figura 3.6). y cuando no lo son (Figura 3.7). En diagramas de Venn Euler.

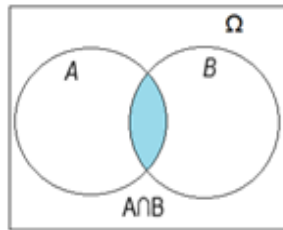
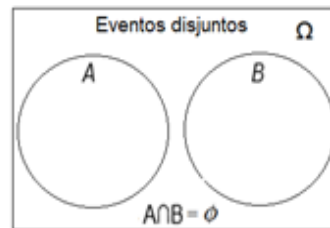


Figura 3.6 Eventos no Disjuntos



Eventos 3.7 Eventos Disjuntos

También el complemento de un evento A (Figura. 3.8), que es simbolizado como A^c y se expresa en diagramas de Venn Euler como:

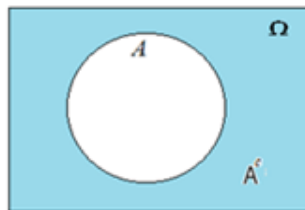


Figura 3.8 Complemento de A

$$A^c = \{x \mid x \notin A \text{ pero } x \in \Omega\}$$

Ejemplo 3.3

Supóngase un espacio muestral que consta de los números enteros del 1 al 20 y se definen los siguientes eventos:

- Sea A el evento de los números 1, 2, 3, 4, 5.
- Sea B el evento de los números 4,5,6,7,8,9,10
- Sea C el eventos de los números 13,14,15,16,17,18,19,20

Obtener:

La probabilidad de los eventos A, B, C y la probabilidad de A^c .

Solución

Se tiene que: $\Omega = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20\}$ su cardinalidad es

$$n(\Omega) = 20$$

ahora, para los eventos A, B y C

$A = \{1,2,3,4,5\}$ y su cardinalidad es $n(A) = 5$, en un diagrama de Veen

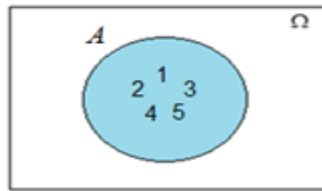


Figura 3.10 Evento A

y su probabilidad es:

$$P(A) = \frac{n(A)}{n(\Omega)} = \frac{5}{20}$$

$B = \{4,5,6,7,8,9,10\}$ y su cardinalidad es $n(B) = 7$, en un diagrama de Veen

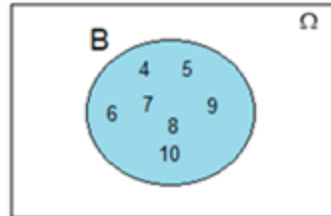


Figura 3.11 Evento B

y su probabilidad es:

$$P(B) = \frac{n(B)}{n(\Omega)} = \frac{7}{20}$$

$C = \{13,14,15,16,17,18,19,20\}$ y su cardinalidad es $n(C) = 8$ en un diagrama de Veen

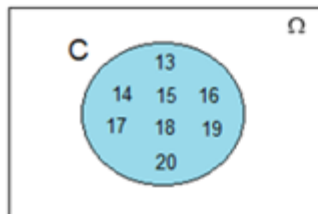


Figura 3.12 Evento C

y su probabilidad es:

$$P(C) = \frac{n(C)}{n(\Omega)} = \frac{8}{20}$$

El complemento de un evento se define, por comprensión, como:

$$A^c = \{x \mid x \notin A \text{ pero } x \in \Omega\}$$

A^c es el evento de elementos x tal que x no es elemento del evento A , pero x es elemento del Espacio Muestral. Es decir, como $A = \{1,2,3,4,5\}$ ningún de estos elementos puede pertenecer al complemento de A . Por lo que:

$$A^c = \{6,7,8,9,10,11,12,13,14,15,16,17,18,19,20\}$$

Nótese que:

$$P(\Omega) = P(A) + P(A^c)$$

y como $P(\Omega) = 1$ entonces:

$$1 = P(A) + P(A^c)$$

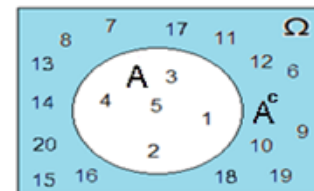


Figura 3.13 Complemento de A

despejando

$$P(A^c) = 1 - p(A) = 1 - \frac{5}{20} = \frac{20}{20} - \frac{5}{20} = \frac{15}{20} = \frac{3}{4}$$



1. ¿Cómo se expresa el espacio muestral con un diagrama de Venn Euler? _____

2. ¿Cómo son expresados los eventos con un diagrama de Venn Euler? _____

3. ¿Si se lanzan cuatro monedas de cuantas maneras pueden caer? _____

4. ¿Cómo representarías en un diagrama de Venn la unión y la intersección de dos eventos A y B?

Ejemplo 3.4

Una baraja tiene 20 cartas, ocho caballos y las restantes están en blanco.



Hallar la probabilidad de extraer una carta con un caballo.

Solución

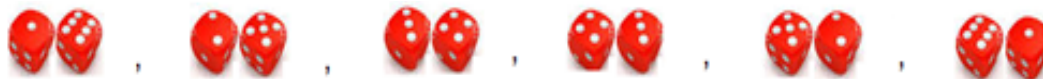
Sea A el evento de que al tomar una carta de las 20 se obtenga un caballo.

$$P(A) = \frac{n(A)}{n(\Omega)} = \frac{8}{20} = \frac{2}{5}$$

Ejemplo 3.5

Se lanza un par de dados. Sean

- A el evento de que la suma de los puntos de las caras que caen hacia arriba sea 7.



- B la suma de los puntos de las caras que caen hacia arriba sea a lo más 3.



Entonces $A \cup B$ representaría el evento compuesto de que al lanzar un par de dados, la suma de los puntos de las caras que caen hacia arriba sea 7 ó a lo más 3. Como A es el evento de que la suma de los puntos de las caras que caen hacia arriba sea 7 y B de que la suma de los puntos de las caras que caen hacia arriba sea a lo más 3. Entonces cualquier elemento x de $A \cup B$ debe cumplir con la condición del evento A o del evento B o de ambos cuando sea posible. Es decir:

$$A \cup B = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1), (1,2), (2,1), (1,1)\}.$$

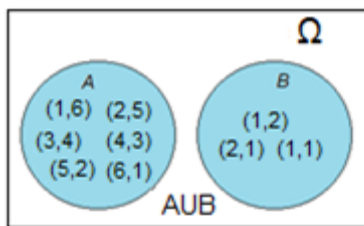


Figura 3.14 Eventos Disjuntos

Y representados en un diagrama de Venn, en este caso, los círculos están separados, debido a que no tienen elementos comunes, se representa como: (Figura. 3.14). Ahora ¿Cuál es la probabilidad $A \cup B$? , en este caso, aun cuando todavía no sería muy complicado tratar de calcular

la probabilidad, ya no es tan sencillo, sin embargo, cuando se comienza a calcular probabilidades, poco a poco esto, tiende a complicarse, debido a que entre más grande es el experimento, el espacio muestra también será grande y por ende los eventos, y saber el número de elemento del espacio muestral y el de los eventos es difícil o muy difícil, por lo que es necesario aprender técnica para poderlos contar. Esas técnicas son llamadas técnicas de conteo y se dividen para su estudio en dos. (Figura. 3.15).

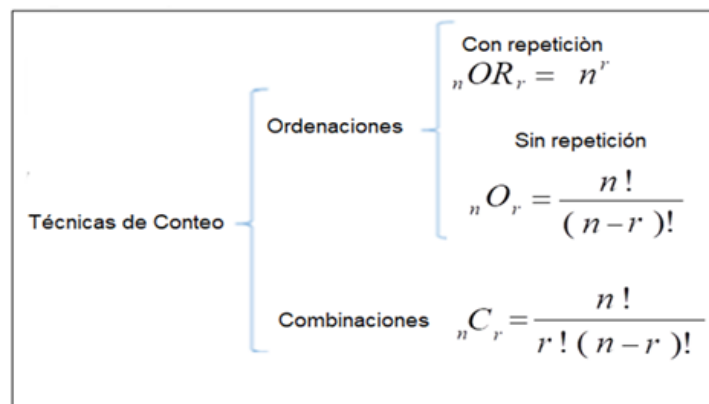


Figura 3.15 Técnicas de Conteo

Para aprender a utilizar estas técnicas, es conveniente tener en cuenta, que en las ordenaciones, los elementos que se desean contar, pueden ser manipulados y ordenados, además se debe observar si los elementos se pueden repetir (como sucede

con los números) y por lo tanto, se utilizan las ordenaciones con repetición ${}_nOR_r = n^r$; y cuando no es posible repetir los datos (como sucede con las personas) se utilizarán las ordenaciones sin repetición ${}_nO_r = \frac{n!}{(n-r)!}$, siempre y cuando, exista una razón para acomodarlo. Por ejemplo si en el grupo 564 de Estadística y Probabilidad I, se necesitan dos representantes de grupo de un total de 45 alumnos, no importaría el orden para poder elegirlos, puesto que sean cualesquiera los nombres de los alumnos, siempre será la misma representación. Sin embargo, si los representantes de grupo una fuera nombrado jefe de grupo y el otro su ayudante, sí importaría el orden de elección. En el primer caso se tendrían que utilizar las combinaciones de 45 alumnos tomados dos a la vez, que aunque es posible visualizarlos o manipularlos no importa el orden, mientras que en el segundo caso, se deben utilizar las ordenaciones sin repetición de 45 alumnos tomados dos a la vez, y aquí sí importa el orden, porque no es lo mismo ser el jefe de grupo que ser el ayudante. Por otra parte, en probabilidad cuando se utilizan urnas (cajas), estas se usan debido a que no es posible manipular los elementos o no es posible visualizarlos, es decir, no se puede o no se debe llevar control, por ejemplo, en el juego de la Lotería Nacional, no se puede llevar control, porque el juego perdería su legalidad.

En este caso se usan las combinaciones ${}_nC_r = \frac{n!}{r!(n-r)!}$.

Otra técnica muy usada es el **Principio Fundamental del Conteo** que dice:

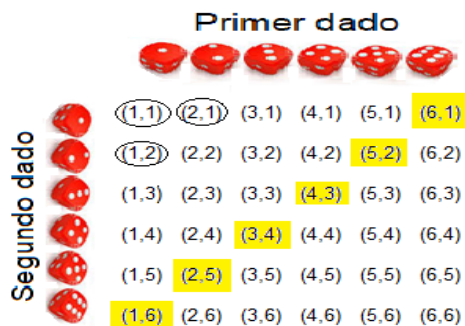
Si en un experimento un primer ensayo puede ocurrir de n_1 maneras diferentes y si continuando el experimento un segundo ensayo puede ocurrir de n_2 maneras diferentes y si continuando el experimento un tercer ensayo puede ocurrir de n_3 maneras diferentes, entonces el número de formas como puede ocurrir el experimento es:

$$\text{Número total de formas} = n_1 \times n_2 \times n_3 \times \dots$$

Continuando con el ejemplo 3.5, cuando se **lanzan dos dados**, cada uno de los dados, para el principio, significa un ensayo, es decir, hay dos ensayos; n_1 y n_2 las formas como pueden caer los dados, son 6 formas en cada uno. Por lo tanto, el número total de formas, en cómo pueden caer los dados, es igual a $6 \times 6 = 36$ formas totales. El espacio muestral correspondiente es:

$$\Omega = \left\{ \begin{array}{l} (1,1), (2,1), (3,1), (4,1), (5,1), (6,1) \\ (1,2), (2,2), (3,2), (4,2), (5,2), (6,2) \\ (1,3), (2,3), (3,3), (4,3), (5,3), (6,3) \\ (1,4), (2,4), (3,4), (4,4), (5,4), (6,4) \\ (1,5), (2,5), (3,5), (4,5), (5,5), (6,5) \\ (1,6), (2,6), (3,6), (4,6), (5,6), (6,6) \end{array} \right\}$$

Como se puede observar, en el siguiente cuadro, los elementos del evento A aparecen remarcados, mientras que los elementos del evento B se encuentran encerrados en círculos.



es claro, que los eventos son disjuntos, ya que no tienen elementos comunes y utilizando uno de los axiomas de probabilidad, se tiene que: “Para dos eventos mutuamente exclusivos A y B se tiene que:

$$P(A \cup B) = P(A) + P(B) \text{ si y sólo si } (A \cap B) = \emptyset.$$

Entonces, la probabilidad de que al lanzar un par de dados, la suma sea 7 o sea por lo menos igual a 3 es igual a la suma de las probabilidades, como $A = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$ y $B = \{(1,2), (2,1), (1,1)\}$ entonces:

$$P(A \cup B) = P(A) + P(B) = \frac{6}{36} + \frac{3}{36} = \frac{9}{36} = \frac{1}{4}$$

Utilizando el mismo experimento del ejemplo 3.4, se plantearan dos eventos que no sean disjuntos (Figura 3.16), con el fin de diferenciar entre eventos disjuntos y eventos no excluyentes.

- Sea C el evento de que al lanzar un par de dados, en el primer dado caiga 4.
- Sea D el evento de que al lanzar un par de dados, el segundo dado caiga 4.

¿Cuál es la probabilidad de $(C \cup D)$?

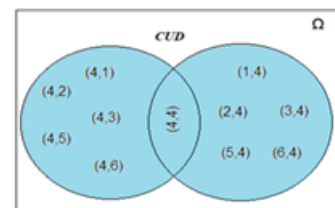


Figura 3.16 Unión de los eventos C y D

En la siguiente cuadro, los elementos del evento C están remarcados y los del evento D encerrados en un círculo.

		Primer dado					
Segundo dado		(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
		(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
		(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
		(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
		(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
		(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

Es claro ver que el elemento (4,4) es común para los dos eventos, esto significa que los eventos no son disjuntos. Y por lo tanto, se utilizará para calcular la probabilidad del evento compuesto $C \cup D$, el siguiente teorema: “**(regla de la adición)**: Para dos eventos cualquiera A y B , $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ ” (Lipschutz, L. y Lipso, M., 2000). Entonces

$$P(A \cup B) = 6/36 + 6/36 - 1/36$$

se debe restar una vez la probabilidad de la intersección, ya que si no lo hacemos, la estaríamos sumado dos veces.

Después de estudiar la unión de dos eventos y su probabilidad, se debe estudiar la probabilidad de la intersección de un evento compuesto, $A \cap B$, es decir, dados “dos eventos A y B , se define un nuevo evento llamado conjunción de A y B , que se denota con $A \cap B$, de la siguiente manera: $A \cap B$ ocurre siempre que ocurra A y ocurra B ; es decir, que ocurran ambos simultáneamente. A la probabilidad de $A \cap B$, que se simboliza $P(A \cap B)$, se llama probabilidad conjunta de A y B .” (Sánchez, E., Izunsa, S., Ramírez, G., 2009). Si el espacio muestral es equiprobable, la probabilidad conjunta se calcula mediante la ecuación:

$$P(A \cap B) = \frac{\text{Cardinalidad de } A \cap B}{\text{Cardinalidad } \Omega} = \frac{n(A \cap B)}{n(\Omega)}$$



1. Si se lanzan 4 monedas. ¿de cuántas formas pueden caer? _____

2. ¿Cuál sería la técnica de conteo que me puede ayudar en este caso?_____
3. Si se lanza una moneda y un dado, ¿de cuántas formas pueden caer?_____
4. Si a 10 alumnos les entregan tres premios por sus buenas calificaciones, el primero de \$20,000.00, el segundo de \$10,000.00 y el tercero de \$5,000.00, ¿de cuántas maneras se puede entregar?_____
5. ¿Se puede llevar orden en la entrega de los premios?_____
6. ¿Un mismo alumno/alumna, puede recibir más de un premio?_____
7. ¿Qué técnica me puede ayudar?_____

Problema 3.6

Se lanza una moneda tres veces.

Sea A el evento de que al lanzar una moneda tres veces caigan exactamente 2 soles y

Sea B el evento de que al lanzar una moneda tres veces caigan a lo más 2 soles.

Calcular:

1. $P(A \cap B) =$
2. $P(A \cup B) =$
3. $(A \cup B)^c =$

Solución

El número total de formas como pueden caer 3 monedas al ser lanzadas es:

Por el principio fundamental del conteo.

$$\text{Número total de formas} = n(\Omega) = \underset{n_1}{2} \times \underset{n_2}{2} \times \underset{n_3}{2} = 2 \times 2 \times 2 = 8$$

pero, ¿cuáles son esas 8 formas?. Una manera de acomodarlo con mucha facilidad es por medio de divisiones sucesivas entre 2 (Figura 3.17), ya que las monedas pueden caer de 2 formas diferentes cada una. Primero se divide 8 entre 2, que es igual a 4, esto quiere decir, que para la primera moneda o el primer lanzamiento, debemos poner cuatro soles y cuatro águilas, posteriormente, se divide 4 entre 2 que es igual a 2, para el segundo lanzamiento, ponemos 2 soles, dos águilas hasta terminar la columna, por

último se divide 2 entre 2 y da como resultado 1, para el tercer lanzamiento, ponemos 1 sol, 1 águila, un sol, un águila, un sol, un águila, un sol, un águila (Figura 3.17). Quedando finalmente como:

$$\Omega = \{sss, ssa, sas, ass, saa, asa, aas, aaa\}$$

Por lo tanto

$$A = \{ssa, sas, ass\}$$

$$B = \{ssa, sas, ass, saa, asa, aas, aaa\}$$

y como:

$$P(A \cap B) = \{x \mid x \in A \text{ y } x \in B\}$$

$A \cap B = \{(ssa), (sas), (ass)\}$, son los elementos que están en A y que están en B y por lo tanto $n(A \cap B) = 3$, entonces:

$$P(A \cap B) = \frac{n(A \cap B)}{n(\Omega)} = \frac{3}{8}$$

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) = \frac{n(A)}{n(\Omega)} + \frac{n(B)}{n(\Omega)} - \frac{n(A \cap B)}{n(\Omega)} = \\ &= \frac{3}{8} + \frac{7}{8} - \frac{3}{8} = \frac{7}{8} \end{aligned}$$

$$P(A \cup B)^c = 1 - P(A \cup B) = 1 - \frac{7}{8} = \frac{1}{8}$$




Lanzamientos		
Primero	Segundo	Tercero
		
s	s	s
s	s	a
s	a	s
s	a	a
a	s	s
a	s	a
a	a	s
a	a	a

Figura 3.17 Tres lanzamientos de una moneda

Ejemplo 3.7

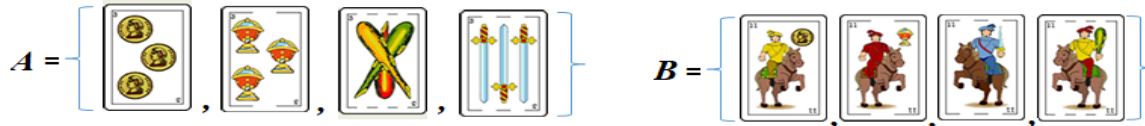
Se tiene una baraja española de 40 cartas.

- Sea A el evento de que al extraer un baraja, sea un caballo
- Sea B el evento de sacar un tres.

Si se sacan dos cartas, cual es la probabilidad de sacar un caballo y un tres.

Cuando se extrae más de una carta de la baraja, se puede hacer con o sin reemplazo, pero esto que quiere decir:

Con reemplazo	Sin reemplazo
Se toma de la baraja una carta, se observa y se apunta el resultado, pero se vuelve a introducir a la baraja.	Se toma de la baraja una carta y esta se queda fuera, ya no se devuelve a la baraja.



$$p(A) = \frac{4}{40} = \frac{1}{10} \quad \text{y} \quad p(B) = \frac{4}{40} = \frac{1}{10}$$

Si primero se hace sin reemplazo, cuál es la probabilidad de sacar un caballo y un tres. Sacar de la baraja un caballo, ya sabemos que es $1/10$ y como es sin reemplazo, sacar la segunda carta sería $4/39$, ya que habría una fuera del mazo. La probabilidad de sacar un caballo y posteriormente un tres, es:

$$p(A \cap B) = p(A)p(C) = \left(\frac{4}{40}\right)\left(\frac{39}{40}\right) = \left(\frac{1}{10}\right)\left(\frac{39}{40}\right) = \frac{39}{400} = 0.0975$$

Por otra parte, cuál sería la probabilidad si lo hiciéramos con reemplazo.

$$p(A \cap B) = p(A)p(B) = \left(\frac{4}{40}\right)\left(\frac{4}{40}\right) = \left(\frac{1}{10}\right)\left(\frac{1}{10}\right) = \frac{1}{100} = 0.01$$

Ejercicios 3.3

Ahora, se extraen dos cartas de la baraja. Cuál es la probabilidad de que sean:



- Las dos deoros.
- Una de copas y otra deoros.
- Al menos una deoros.
- La primera de copas y la segunda de oro.

Ejemplo 3.8

En una urna hay 6 bolas blancas y 4 negras (Figura 3. 18). ¿Qué probabilidad hay de que al extraer al azar dos bolas con reemplazo y sin reemplazo, de la urna:

- las dos sean blancas?
- las dos sea negras?
- haya una blanca y una negra?

En este problema a diferencia del anterior, debes extraer dos bolas de la caja, luego puede ocurrir que sean azules las dos, rojas las dos o que sean de diferentes colores, sin embargo, se resuelve tomando en cuenta si se extraen con o sin reemplazo. ¿Cómo se resuelve si se extraen las bolas con reemplazo?

Cuando es **con reemplazo**, la bola se saca de la urna se observa se anota el color de la bola y posteriormente se vuelve a poner en la urna, por lo cual, la bola puede volver a ser extraída, pero no se podría saber si al realizar otra extracción y sale del mismo color, fuera la misma bola. En otras palabras, no es posible llevar un orden de extracción, por lo tanto se deben usar combinaciones.

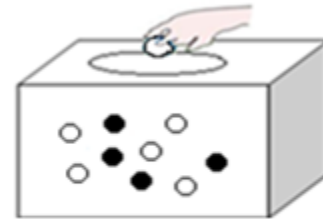


Figura 3.18
Urnas con bolas blancas y negras

- Sea A el evento de que al sacar dos bolas las dos sean blancas.
- Sea B el evento de que al sacar dos bolas las dos sean negras.
- Sea C el evento de que al sacar dos bolas sean de diferente color.

El espacio muestras sería:

$$\Omega = {}_{10}C_2 = \frac{10!}{2!(10-2)!} = \frac{10 \times 9 \times \cancel{8} \times \cancel{7} \times \cancel{6} \times \cancel{5} \times \cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1}}{(2 \times 1)(\cancel{8} \times \cancel{7} \times \cancel{6} \times \cancel{5} \times \cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1})} = \frac{10 \times 9}{2 \times 1} = \frac{90}{2} = 45$$

El número de elementos del evento A sería igual a:

$$n(A) = {}_6C_2 = \frac{6!}{2!(6-2)!} = \frac{6 \times 5 \times \cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1}}{(2 \times 1)(\cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1})} = \frac{6 \times 5}{2 \times 1} = \frac{30}{2} = 15$$

Por lo tanto, la probabilidad de que las dos bolas sean blancas es:

$$P(A) = \frac{n(A)}{n(\Omega)} = \frac{15}{45}$$

El número de elementos del evento B sería igual a:

$$n(B) = {}_4C_2 = \frac{4!}{2!(4-2)!} = \frac{4 \times 3 \times \cancel{2} \times \cancel{1}}{(2 \times 1)(\cancel{2} \times \cancel{1})} = \frac{4 \times 3}{2 \times 1} = \frac{12}{2} = 6$$

Por lo tanto la probabilidad de que las dos bolas sean negras es:

$$P(B) = \frac{n(B)}{n(\Omega)} = \frac{6}{45}$$

El número de elementos del evento C sería igual a:

$$n(C) = ({}_6C_1)({}_4C_1) = \left(\frac{6!}{1!(6-1)!} \right) \left(\frac{4!}{1!(4-1)!} \right) = 6 \times 4 = 24$$

Por lo que la probabilidad de que al sacar dos bolas salgan de diferente color, como aquí no importa el orden, es decir, puedo sacar primero una bola blanca y después una negra o viceversa (es por tal motivo que estamos utilizando combinaciones), es:

$$P(C) = \frac{n(C)}{n(\Omega)} = \frac{24}{45}$$

Ahora si lo calculamos **sin reemplazo**, como al extraer la primera bola la dejamos fuera, la extracción de la segunda se realizará ya sabiendo cual es el color de la primera. Y podemos llevar un orden relativo y por lo tanto utilizaremos el principio fundamental del conteo, el número de elementos del espacio muestral sería:

$$n(\Omega) = \begin{matrix} \boxed{10} & \boxed{9} \\ n_1 & n_2 \end{matrix} = 90$$

Y el número de elementos del evento A es:

$$n(A) = \begin{matrix} \boxed{6} & \boxed{5} \\ n_1 & n_2 \end{matrix} = 30$$

Por lo tanto, la probabilidad de sacar dos bolas blancas es:

$$P(A) = \frac{n(A)}{n(\Omega)} = \frac{30}{90} = \frac{1}{3}$$

Ahora el número de elementos del evento B es:

$$n(B) = \begin{matrix} \boxed{4} & \boxed{3} \\ n_1 & n_2 \end{matrix} = 12$$

Por lo que, la probabilidad de sacar dos bolas negras es:

$$P(B) = \frac{n(B)}{n(\Omega)} = \frac{12}{90} = \frac{2}{15}$$

Por último, es posible sacar primero una bola blanca y después una negra o bien, sacar primero una negra y después una blanca, por lo que:

Número de elementos del evento C es:

$$n(C) = \begin{matrix} \boxed{6} & \boxed{4} \\ n_1 & n_2 \end{matrix} = 24 \quad \text{ó} \quad n(C) = \begin{matrix} \boxed{4} & \boxed{6} \\ n_1 & n_2 \end{matrix} = 24$$

Por lo tanto, la probabilidad pedida es:

$$P(C) = \frac{n(C)}{n(\Omega)} = \frac{24}{90} = \frac{4}{15}$$



1. ¿Cuántos enfoques de probabilidad hay y cuáles son? _____
2. ¿Cómo se define la probabilidad clásica? _____
3. Si se lanza una moneda, ¿cuál es la probabilidad de caiga sol? _____
4. Si se lanza un dado, ¿cuál es la probabilidad de que el número de puntos que caen en la cara que cae hacia arriba, sea par? _____
5. Si 10 alumnos decidieran apostar \$10.00 cada uno y pusieran su nombre en un papel y lo metieran a una caja y se sacara un papel al azar. Si tú fueras uno/una de los 10, ¿cuál sería la probabilidad de que ganaras la apuesta? _____

Problema 3.9

Si se tiene telas con los colores rojo, blanco, verde, amarillo, azul, negro, naranja y gris, y se formaran banderas horizontales de 3 colores continuos, como se muestra en la siguiente figura:



Ejercicios 3.5



1. ¿Importa el orden como se acomodan los colores para formar las banderas? _____

2. ¿Se pueden repetir los colores? _____
3. ¿Es la misma bandera una bandera verde, blanco y rojo que una bandera rojo, blanco y verde o una bandera, blanco, rojo y verde? _____
4. ¿Cuál sería la probabilidad de tomara una bandera, tuviera| los colores de la bandera mexicana? (Figura.3.21) _____



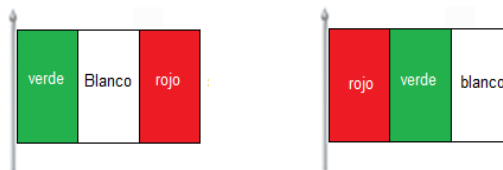
Figura 3.21
Colores de la bandera mexicana

Solución

Como una bandera tiene sentido de orientación y el número de colores de las telas es 8, por el principio fundamental del conteo se tiene:

$$\begin{array}{|c|c|c|} \hline 8 & 7 & 6 \\ \hline \end{array} = 8 \times 7 \times 6 = 336$$

se pueden formar 336 banderas diferentes de tres colores horizontales. El orden como sean acomodados los colores sí importa y además no se pueden repetir, ya que no es lo mismo una bandera verde, blanca y roja, como es el caso de la bandera mexicana, que si pusiéramos rojo, verde y blanco.



la probabilidad de formar una bandera con los colores de la bandera mexicana es:

Como dentro de los colores de las telas sólo hay un verde, un blanco y un rojo, la bandera mexicana por el principio fundamental del conteo se puede formar de:

$$\begin{array}{|c|c|c|} \hline \text{verde} & \text{blanco} & \text{rojo} \\ \hline 1 & 1 & 1 \\ \hline \end{array} = 1 \times 1 \times 1 = 1$$

Para la bandera mexicana solo hay una forma de formarla, con los colores y las características del planteamiento del problema. Por lo tanto, la probabilidad de hacer la

bandera mexicana es $\frac{1}{336}$.

Ejemplo 3.10.

Se tienen los números 1, 2, 3, 4, 5, 6, 7, 8, 9 y se quieren numerar casas con dos números (Figura. 3.22). Se sabe que las casas que terminan en 2, tienen calentador solar. Como el señor Pedro va a comprar una casa, desea saber cuál es la probabilidad de que le toque un calentador solar. Lo primero es saber:



Figura 3.22 Casas



1. ¿Cuántas casas podemos numerar? _____
2. ¿Nos podría ayudar alguna técnica de conteo? _____
3. ¿Cuál sería? _____
4. ¿Se pueden manipular los números? _____
5. ¿Puedo formar los números? _____
6. ¿Se pueden repetir? _____

Solución

Utilizando el Principio Fundamental del Conteo o las ordenaciones con repetición, podemos encontrar el número total de números que podemos formar, que es igual a $n(\Omega)$, utilizando las ordenaciones con repetición

$$nORr = {}_9OR_2 = 9^2 = 81$$

11	21	31	41	51	61	71	81	91
12	22	32	42	52	62	72	82	92
13	23	33	43	53	63	73	83	93
14	24	34	44	54	64	74	84	94
15	25	35	45	55	65	75	85	95
16	26	36	46	56	66	76	86	96
17	27	37	47	57	67	77	87	97
18	28	38	48	58	68	78	88	98
19	29	39	49	59	69	79	89	99

Se pueden formar 81 números diferentes ¿Cuántas casas con calentador hay? Para poder saber el número de casas con calentador, que son las que terminan en 2, hay una

restricción, es decir, los números formados pueden comenzar con cualquier de los nueve números que se tienen, sin embargo, sólo hay una forma como pueden terminar y es con el número dos. Utilizando el Principio Fundamental del Conteo. Nótese en la tabla anterior, que el número podrá comenzar con cualquiera de los nueve números, pero sólo las casas terminadas en dos, tiene calentador solar, porque:

$$n(\text{casas que termina en 2}) = \begin{array}{|c|c|} \hline 9 & 1 \\ \hline n_1 & n_2 \\ \hline \end{array} = 9 \times 1 = 9$$

Ahora cual sería la probabilidad de que el señor Pedro, tuviera en su casa un calentador solar

$$P(\text{casa con calentador solar}) = \frac{n(\text{casas que terminan en 2})}{n(\text{total de casas})} = \frac{9}{81} = \frac{1}{9}$$

Ejercicios 3.7



1. Si nos pidieran numerar casas de tres números, con los mismos números. ¿Cuántas casas podríamos numerar? _____
2. ¿Tendríamos que cambiar las técnicas usadas si nos piden aumentar el número de dígitos? _____
3. ¿Cuál sería la probabilidad de que tomada una casa al azar, comenzara con 1 o 2 y fuera par? _____
4. Si las casas que tienen un estacionamiento para dos vehículos, son aquellas que tienen los tres números iguales. ¿Cuál es la probabilidad de comprar una casa que tenga un estacionamiento para dos vehículos? _____

Ejemplo 3.11

En el grupo 564 de Estadística y Probabilidad del CCH Vallejo, hay 47 estudiantes, de los cuales 13 son nadadores, 18 juegan futbol, 8 practican ambos deportes, 2 juegan tenis y 20 no hacen ningún deporte (Figura 3.23).

Cuál es la probabilidad de que tomado un alumno al azar.

- a. No practique ningún deporte
- b. Juegue futbol y nade
- c. Que solamente juegue futbol o practique tenis
- d. Practique tenis

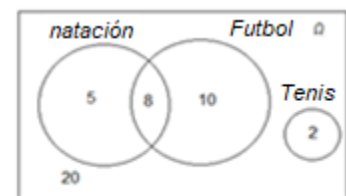


Figura 3.23

Solución

Sean B nadan, F juegue futbol, A practique tenis y N no practique ningún deporte, entonces:

Los elemento que pertenecen al espacio muestral es $n(\Omega) = 47$

El número de elementos que pertenecen al evento de que no practiquen ningún deporte es $n(N) = 20$, por lo tanto,

$$P(N) = \frac{n(N)}{n(\Omega)} = \frac{20}{47} .$$

Que un alumno juegue futbol y nadan, se simboliza como $(F \cap B)$ y representa los elemento que están en F y que también están B. Los elementos que cumplen esta característica son:

$$n(F \cap B) = 8 \quad \text{por lo tanto} \quad P(F \cap B) = \frac{n(F \cap B)}{n(\Omega)} = \frac{8}{47}$$

Para poder resolver el inciso c, es necesario definir la diferencia de dos eventos, sean F y B, dos eventos que representan jueguen futbol y nadan, respectivamente. Se define la diferencia de F y B, que se representa por $(F-B)$, como el conjunto formado por todos los elementos que están en F, pero no están en A. Es decir:

$$F-B = \{x | x \in F \text{ pero } x \notin B\}$$

en un diagrama de Venn (Figura 3.24)

$$n(F - B) = 5$$

que son los que sólo nadan y sabemos que 2 alumnos practican tenis, por lo tanto.

$$\begin{aligned} P[(F - B) \cup A] &= P(F - B) + P(A) \\ &= \frac{10}{47} + \frac{2}{47} = \frac{12}{47} \end{aligned}$$

Por último, la probabilidad de que un estudiante sólo juegue tenis es:

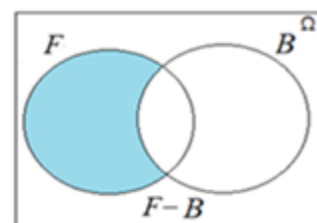


Figura 3.24 Diferencia de eventos

$$P(\text{juegue tenis}) = \frac{2}{47}$$

Ejemplo 3.12

Una clase consta de 10 hombres y 20 mujeres; la mitad de los hombres y la mitad de las mujeres son morenos. Determinar la probabilidad de que una persona elegida al azar sea un hombre o moreno y la probabilidad de que sea mujer y sea morena.

Solución

Sean H el evento de ser hombre; M el evento de ser mujer; A el evento de ser moreno y O el evento de tener la piel de otro color.

	Hombre	Mujer	total
Moreno	5	10	15
Otro color de piel	5	10	15
total	10	20	30

$$P(H \cup A) = \frac{10}{30} + \frac{15}{30} - \frac{5}{30} = \frac{20}{30} = \frac{2}{3} \quad \text{y} \quad P(M \cap A) = \frac{10}{30}$$

Ejemplo 3.13

Considérese a una pareja con 2 hijos.

- Sea A el evento de que los dos sean hombres.
- Sea B el evento de al menos un hombre.

Solución

¿Cuál sería la probabilidad de que los dos fueran hombres, dado que sabemos que por lo menos uno de ellos es hombre, es decir, la probabilidad de A dado B o la probabilidad condicional de que suceda A dado que B ya sucedió. Es decir, la probabilidad de A depende de B . El resultado es:

$$\Omega = \{hh, hm, mh, mm\}$$

y los eventos A y B

$$A = \{hh\} \quad \text{y} \quad B = \{hh, hm, mh\}$$

si se calculan las probabilidades de A y de B . Entonces:

$$P(A) = \frac{1}{4} \quad \text{probabilidad de un hombre}$$

y $P(B) = \frac{3}{4} \quad \text{probabilidad de al menos un hombre}$

Ejemplo 3.14

Un alumno que no fue a la escuela por estar enfermo, debe entregar tres tareas a tres profesores diferentes.

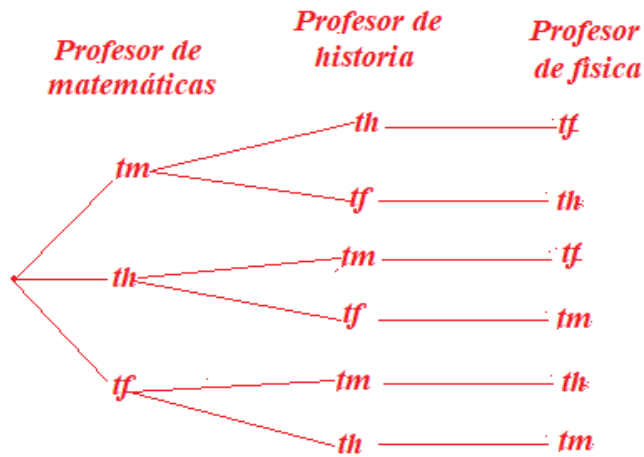


Figura 3.25 Tareas

Si su mamá entrega las tareas al azar a los profesores, ¿cuál es la probabilidad de que al menos una de las tareas la haya recibido el profesor adecuado?

Solución

Sean M el profesor de matemáticas, H el profesor de historia y F de física; y a las tareas correspondientes *tm*, *th* y *tf*. Realizando un diagrama de árbol:



Las formas totales son:

Profesor			
Matemáticas	Historia	Física	
tarea de matemáticas	tarea de historia	tarea de física	<i>Las tres correctas</i>
tarea de matemáticas	tarea de física	tarea de historia	<i>Sólo una correcta</i>
tarea de historia	tarea de matemáticas	tarea de física	<i>Sólo una correcta</i>
tarea de historia	tarea de física	tarea de matemáticas	<i>Ninguna correcta</i>
tarea de física	tarea de matemáticas	tarea de historia	<i>Ninguna correcta</i>
tarea de física	tarea de historia	tarea de matemáticas	<i>Sólo una correcta</i>

Por lo menos 1

Profesor			
Matemáticas	Historia	Física	
tarea de matemáticas	tarea de historia	tarea de física	<i>Las tres correctas</i>
tarea de matemáticas	tarea de física	tarea de historia	<i>Sólo una correcta</i>
tarea de historia	tarea de matemáticas	tarea de física	<i>Sólo una correcta</i>
tarea de física	tarea de historia	tarea de matemáticas	<i>Sólo una correcta</i>

Se observa que hay seis posibles ordenaciones y que en cuatro de ellas hay al menos una entrega correcta. Por tanto, la probabilidad pedida será:

$$P(\text{al menos una tarea se entregue bien}) = \frac{4}{6} = \frac{2}{3}$$

3.5 Probabilidad Condicional y eventos independientes

La probabilidad de que ocurra el evento A si ya ha ocurrido el evento B se denomina *probabilidad condicionada* y se define

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \quad \text{si } p(B) \neq 0$$

Se refiere a la probabilidad de ciertos eventos (A) que dependen o se ven influidas por la ocurrencia de otros (B). La probabilidad condicional se representa $P(A|B)$, y se pronuncia como "la probabilidad de A dado en B". Esta definición es consistente, es decir cumple los axiomas de probabilidad. En términos generales consideramos que dos eventos A y B son independientes si la probabilidad de uno de ellos no depende de la ocurrencia del otro evento, es decir:

$$P(B) = P(B|A)$$

La probabilidad de que ocurra B es igual a la probabilidad de que ocurra B sabiendo que ocurrió A.

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

de donde, si los eventos A y B son independientes, entonces $P(B | A) = P(B)$ y obtenemos:

$$P(A \cap B) = P(A)P(B)$$

Recíprocamente, si $P(A \cap B) = P(A)P(B)$ entonces:

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

Entonces podemos decir, que dos eventos A y B son independientes, si la ocurrencia de uno de ellos no afecta la ocurrencia del otro, es decir, cuando los eventos A y B no están relacionados. Para eventos independientes, la regla de la multiplicación establece que:

$$P(A \cap B) = P(A) \times P(B)$$

En otras palabras, dos eventos son independientes sí cuando la probabilidad de cada uno de ellos no está influida porque el otro evento ocurra o no, es decir, cuando ambos eventos no están relacionados. Para que dos eventos sean independientes tienen que verificar al menos una de las siguientes condiciones:

$$P(A|B) = P(A) \quad \text{y} \quad P(B|A) = P(B)$$

NOTA: Eventos Disjuntos o mutuamente excluyentes \neq Eventos independientes

A y B son eventos independientes $\Leftrightarrow P(A \cap B) = P(A)P(B)$ ya que por definición de independencia $P(A|B) = P(A)$ y por definición de probabilidad condicional

$P(A|B) = \frac{P(A \cap B)}{P(B)}$ entonces si A y B son independientes se deberá cumplir

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

Se tratará de probar esto con un ejemplo.

Sean $\Omega = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20\}$

$$A = \{1,2,3,4,5\} \quad \text{y} \quad B = \{9,10,11,12,13\}$$

entonces $A \cap B = \phi$

ahora
$$P(B) = \frac{n(B)}{n(\Omega)} = \frac{5}{20} = \frac{1}{4}$$

además
$$P(A \cap B) = \frac{n(A \cap B)}{n(\Omega)} = \frac{0}{20} = 0$$

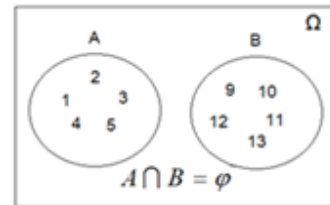


Figura 3.26

si dos eventos son independientes, se debe cumplir que:

$$P(A \cap B) = P(A) \cdot P(B)$$

entonces

$$P(A \cap B) = 0 \quad \text{y} \quad P(A) \cdot P(B) = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$$

por lo tanto, $P(A \cap B) \neq P(A)P(B)$

y no se cumple la igualdad, por lo que podemos concluir que A y B no son eventos independientes.

Ejemplo 3.15:

Una persona desea saber ¿cuál es la probabilidad de que su segundo hijo sea hombre, ya que el primero fue mujer?

Solución

El espacio muestral

$$\Omega = \{mm, mh, hm, hh\}.$$

Sea A el evento de que el segundo hijo se hombre $A = \{mh, hh\}$, entonces el evento B sería el primer hijo fue mujer $B = \{mm, mh\}$. Por lo tanto, la probabilidad de que suceda A dado que B ya sucedió es, como $A \cap B = \{mh\}$ entonces:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{4}}{\frac{2}{4}} = \frac{4}{8} = \frac{1}{2}$$

Ejemplo 3.16

¿Cuál sería la probabilidad de al lanzar un dado, caiga un número impar, sabiendo que anteriormente el dado cayo por lo menos 3?

Solución

Sea A el evento de que al lanzar un dado caiga en la cara que cae hacia arriba, la suma de los puntos sea non y sea B el evento de que al lanzar un dado, la suma de los puntos de las caras que caen hacia arriba se a lo más 3, se tiene que:

$$A = \{1,3,5\} \text{ y } B = \{1,2,3\} \text{ por lo tanto, } (A \cap B) = \{3\}$$

El resultado es:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Ejemplo 3.17

En el CCH Vallejo, el 25% de los alumnos reprobaban matemáticas, el 15% química y el 10% reprobaban las dos materias. Si se toma al azar un estudiante, y se sabe que reprobó química, ¿cuál es la probabilidad de que también haya reprobado matemáticas?

Solución

- Sea A el evento de reprobado matemáticas.
- Sea B el evento de reprobado química.

Entonces $P(A) = 0.25$, $P(B) = 0.15$ y $P(A \cap B) = 0.10$ Por lo tanto, la probabilidad de que el alumno reprobado matemáticas dado que reprobó química, es:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.10}{0.15} = \frac{2}{3} = 0.66$$

Si reprobado matemáticas, ¿cuál es la probabilidad de que también reprobado química?

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{0.10}{0.25} = \frac{2}{5} = 0.40$$

Propuesta de evaluación para la unidad 3

Se recomienda al profesor, que al inicio de la unidad, explore el nivel de conocimientos previos, “el factor más importante que influye en el aprendizaje, es lo que el alumno ya sabe. Determinar esto y enseñarle en consecuencia” (Ausubel, 1968). Con ello el profesor podría tener una mejor idea del nivel de conocimientos que los alumnos poseen. Los siguientes ejercicios, sólo son una propuesta y el profesor podrá utilizarlos si así lo desea.

1. Se lanzan 4 monedas. ¿Cuál es la probabilidad de que caigan:
 - a) Tres soles?
 - b) Exactamente caiga una vez sol?
 - c) Por lo menos una caiga una vez sol?
2. Se lanzan dos dados. ¿Cuál es la probabilidad que la suma de los puntos de las caras que caen hacia arriba sea:
 - a) por lo menos 10?
 - b) 2 o 4?
 - c) pares o impares?
 - d) Impares y pares?
3. Una urna contiene 4 canicas blancas y 6 negras. Se extraen una canicas al azar sin reemplazo, ¿cuál es la probabilidad de que sea blanca?
4. Se tiene un grupo de 10 personas. y se quiere formar un comité de 5 personas. ¿Cuántas maneras hay de formarlo?
5. Se tiene un grupo de 10 personas. y se quiere repartir en ellos tres premios, uno de \$10,000.00, otro de \$5,000.00 y el último de \$1,000.00 ¿Cuántas maneras hay de formarlo?
6. Se tienen los dígitos (0.1.2.3.4.5.6.7.8.9) y se forman números de 4 cifras. ¿Cuál es la probabilidad de:
 - a. Los números sean pares?
 - b. Los números comiencen con 1,2,o 3 y sean nones?
 - c. Qué los números terminen en cero?
 - d. Todos los números sean iguales?

7. Sean A y B dos eventos aleatorios, con $P(A) = \frac{1}{2}$ y la $P(B) = \frac{1}{3}$. Si además se sabe que $P(A \cap B) = \frac{1}{4}$. Obtener:
- $P(A \cup B) =$
 - $P(A|B) =$
 - $P(B|A) =$
 - $P(A \cup B)^C =$
 - $P(A \cap B)^C =$
8. Un hombre visita a un matrimonio que tiene dos hijos. Uno de los hijos, el mayor, se sabe que es un niño. Hallar la probabilidad de que el otro sea también niño.
9. Se extraen dos cartas de una baraja española (de cuarenta cartas). Calcular la probabilidad de que sean:
- Las dos de ases.
 - Una de espadas y otra de bastos.
 - Al menos una de copas.
10. En una urna existe 10 fichas numeradas con los dígitos (0,1,2,3,4,5,6,7,8,9).
- ¿Qué probabilidad hay de que al sacar una ficha, salga un múltiplo de 2?
- ¿Qué probabilidad hay de que al sacar una ficha, salga un número menor que 3?

Bibliografía

Recomendada para el alumno

- Ávila, A. Hernández, H. Becerril, H. Cifuentes, M. Domínguez, M. Sánchez, A. y Santos, R. (2006). *Paquete Didáctico de Estadística y Probabilidad I*. México: Colegio de Ciencias y Humanidades.
- Castillo, J. y Gómez, J. (1998). *Estadística Inferencial Básica*. México: Grupo Editorial Iberoamericana.
- de Oteyza, E., Lam, E., Hernández, C. y Carrillo, A. (2015). *Estadística y Probabilidad*. México: Pearson.
- Johnson R. (1990). *Estadística Elemental*. México: Grupo Editorial Iberoamericana.
- Moore, D. (2005). *Estadística Aplicada Básica*. Madrid, España: Antoni Bosch.
- Ross, S. (2008). *Introducción a la Estadística*. Barcelona, España: Reverté.
- Sánchez, E., Inzunza, S. y Ávila, R. (2015). *Probabilidad y Estadística I*. México: Patria.
- Spiegel, M. y Stephens, L. (2009). *Estadística*. 4ra Edición. México: McGraw Hill.
- Triola, M. (2009). *Estadística*. Décima edición. México: Pearson Addison Wesley

Recomendada para el profesor.

- Batanero, C (1998). Recursos para la educación estadística en Internet. *UNO*, 15, 13-25.
- Batanero, C. (2000). Significado y comprensión de las medidas de tendencia central. *UNO*, 25, 41-58.
- Batanero, C. y Borovcnik, M. (2016). *Statistics and Probability in High School*. Rotterdam, Países Bajos: Sense Publishers.
- Batanero, Carmen y Godino, Juan D. (2001). *Análisis de datos y su didáctica*. Departamento de Didáctica de la Matemáticas, Universidad de Granada.
- Sánchez, E. y Orta, J. (2015). Levels of Reasoning of Middle School Students about Data Dispersion in Risk Contexts. *The Mathematics Enthusiast*, 12(1), 275 – 289.
- Wild, C. y Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, 67(3), 223-248.