GUÍA PARA PREPARAR EXAMEN EXTRAORDINARIO

ESTADÍSTICA PROBABILIDAD I

Becerril Montes Helios
Carrasco Licea Guadalupe
Escobar Cristiani Elsa Marlene
Mendoza Zaragoza Lilian

Agosto de 2019

INTRODUCCIÓN

Al estudiar Estadística y Probabilidad, construirás conocimientos y desarrollarás habilidades que te ayudarán a elaborar argumentaciones, interpretaciones y valoraciones de la información. A su vez, un correcto análisis de la información te permite llevar a cabo una toma eficiente de decisiones y contribuye a la formación de un pensamiento crítico propio.

Este tipo de conocimientos y habilidades resultarán fundamentales en tus estudios superiores y en cualquier actividad de una disciplina que maneje datos, gráficas, tablas o medidas, por ejemplo, actividades relacionadas con estudios económicos, médicos o demográficos, con mediciones de pobreza o desempleo, con desarrollo agropecuario, contaminación de aguas o producción petrolera, con intercambios comerciales, etcétera.

Propósitos generales

Al finalizar el trabajo recomendado en esta guía:

- Interpretarás formalmente resultados estadísticos, clarificando el papel del azar y valorando la variabilidad, con la finalidad de verificar la importancia de la estadística y la probabilidad en la construcción de conocimientos y en la evaluación de hechos en diversos campos del saber, a partir del diseño y aplicación de un proceso de investigación estadística que incluya la formulación de preguntas, el levantamiento y análisis de datos.

- Conformarás un pensamiento estadístico que te permitirá tomar decisiones sustentadas, juzgar críticamente la validez o pertinencia de la información estadística y la elaboración de inferencias formales.

INSTRUCCIONES

Usa esta guía de la siguiente manera para preparar tu examen extraordinario:

a) Lee los textos en los que se explican los conceptos.
b) Revisa los ejemplos resueltos de cada parte de la guía.
c) Resuelve los ejercicios propuestos en cada sección.
d) Compara tus resultados finales con los que vienen en el *Anexo de resultados* al final de la guía.
e) Si tienes alguna duda, relee los conceptos y/o acude a asesorías.

Es muy importante que tu trabajo de preparación sea sistemático y durante un tiempo razonable antes de presentar el examen.

CONTENIDO

UNIDAD 1. OBTENCIÓN, DESCRIPCIÓN E INTERPRETACIÓN DE INFORMACIÓN ESTADÍSTICA

Presentación

En esta unidad abordarás aspectos básicos de la Estadística descriptiva como la construcción de tablas y gráficas, y el cálculo de medidas. Usarás estas herramientas para detectar las principales características de los datos y para hacer observaciones útiles sobre ellos. Esto te permitirá convertir los datos en información valiosa que pueda ser aprovechada según el objetivo que se tenga en un estudio.

Propósito

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

1.1 CONCEPTOS BÁSICOS

A. ¿Qué es la Estadística?

Cuando se escucha la palabra *estadística*, la mayoría de las personas piensa en una gran colección de datos, tablas, gráficas, porcentajes y promedios. Sin embargo, no debemos reducir a esto la visión sobre la Estadística.

En la naturaleza existen fenómenos que no obedecen a leyes fijas y que dependen de circunstancias prácticamente incontrolables: fenómenos sociológicos, psicológicos, políticos, económicos, médicos, biológicos, industriales, meteorológicos, etcétera, los cuales presentan una gran variabilidad.

La investigación científica y la toma de decisiones en la vida diaria se enfrentan a esta presencia de la variabilidad, de modo que, para realizarlas de manera óptima, la información que se recolecta debe obtenerse con objetivos definidos, reflejar la realidad, resumirse de manera eficiente e interpretarse adecuadamente. De manera general, podemos decir que la razón principal del uso de la Estadística es la existencia de la variabilidad en los fenómenos que se estudian.

La Estadística es la rama de las matemáticas que estudia los métodos de recopilación, organización, descripción y análisis de datos así como la interpretación de la información, con el objetivo de tomar la decisión más eficaz ante alguna situación de incertidumbre.

B. Investigación estadística

En este tipo de investigaciones se analiza un problema y se utilizan procedimientos estadísticos que convierten datos en información útil para la toma de decisiones eficientes dirigidas a su solución.

Pueden identificarse seis etapas o fases durante su desarrollo, que se explican a continuación mediante un ejemplo.

Etapa1. Identificación del problema y del objetivo de la investigación.
Problema: Durante años se observó que un número muy grande de estudiantes del bachillerato de la UNAM, terminaba sus estudios con rezago o no lograba concluirlos.[1]
Objetivo: Identificar las principales causas que provocan el rezago de estudiantes de la UNAM en el nivel medio superior.

Etapa 2. Recolección de los datos. En el ejemplo, se requiere diseñar una encuesta y determinar un procedimiento para su aplicación, es decir, ¿qué se preguntará para identificar las causas del rezago? ¿a quiénes se les aplicará? ¿cuántas encuestas se requieren? ¿será presencial o en línea? etcétera.
De esta encuesta se puede obtener información sobre género, edad, ingreso familiar, situación de empleo, nivel de estudio de los padres, promedio en secundaria, calificación en el examen de admisión, calificación en el examen diagnóstico inicial, entre otros aspectos.

Etapa 3. Validación de los datos recabados. En toda encuesta, se suelen encontrar preguntas sin contestar, respuestas incomprensibles o ambiguas y otros problemas que dificultan la utilización de una parte de los datos. Se requiere limpiar los datos recabados para dejar los que sean válidos en el estudio.

Etapa 4. Exploración de los datos. El siguiente paso es ordenar los datos a través de tablas y gráficas para que su presentación permita que resalten las características más importantes de los mismos. A partir de una presentación ordenada, se puede iniciar la exploración de los datos haciendo observaciones, calculando medidas estadísticas, identificando grupos, revisando posibles relaciones entre dos o más de las características, etcétera.

Etapa 5. Uso de técnicas estadísticas para el análisis de los datos. La exploración descrita en la etapa anterior, se complementa mediante técnicas estadísticas que permiten inferir algún tipo de información sobre características de una población a partir de información parcial, es decir, información que no fue recabada en toda la población sino solo en una muestra representativa.

Etapa 6. Interpretación de la información y conclusiones. Determinar cuáles son las principales características que tiene en común la población de estudiantes

[1] Rezago en el bachillerato: hecho de que después de 3 años de estar inscrito de forma regular, un estudiante no haya concluido el 100% de las asignaturas del programa oficial de estudios.

rezagados, permite identificar lo que podría llamarse factores de riesgo para futuras generaciones. Por ejemplo, la cantidad de materias que no ha logrado acreditar un estudiante en los 4 primeros semestres, el promedio de secundaria y las calificaciones en los exámenes de admisión y de diagnóstico, así como el género, son factores que deben tomarse en cuenta para identificar con anticipación si un estudiante está en riesgo de no terminar en 3 años. Este análisis permite tomar medidas preventivas y remediales para reducir la presencia de este problema, como por ejemplo, la organización de tutorías y asesorías académicas, la aplicación de programas de información sobre las posibilidades psicológicas y académicas que ofrece la institución y otras.

C. Población y muestras

La *Población* es el conjunto de elementos (personas, animales, plantas u objetos) que tienen ciertas características de interés para un estudio estadístico. Una población debe definirse en términos de:

- Tipo de elementos que la constituyen.
- Lugar o espacio donde se encuentra.
- Período de tiempo en el que se desarrolla el estudio.

La mayor parte de las veces, la población de un estudio estadístico es muy grande, y algunas veces es hipotética (es decir, no se conoce en realidad).

Ejemplos de poblaciones

a) Tiendas de sillas con ventas mensuales mayores a 50,000 unidades en la Ciudad de Buenos Aires, Argentina durante los años de 2016 a 2018.
b) Murciélagos de la familia Phyllostomidae, que se distribuían desde la vertiente del Pacífico de Sinaloa hasta Chiapas, durante los meses de marzo a junio de 2018.
c) Adultos con edades de 65 años en adelante habitantes de la colonia Agrícola Oriental de la Alcaldía de Iztacalco, y que se encontraban en situación de extrema pobreza entre el 15 de mayo y el 15 de junio de 2018.

Se le llama *tamaño de la población* al número total de individuos que la componen y en estas notas lo denotaremos por *N*.

Generalmente el costo, el tiempo y los recursos que se requieren para hacer un estudio que abarque a toda una población, resultan muy elevados. Por ello, se suele recurrir al uso de muestras.

Una *Muestra* es cualquier subconjunto de la población seleccionado para la investigación. Una *Muestra aleatoria* es un subconjunto que ha sido seleccionado mediante un método azaroso o aleatorio. Para que una muestra aleatoria sea útil para una investigación, se requiere que sea representativa de la población, es decir,

que sus elementos recojan características esenciales de los elementos que componen la población.

En estas notas, el tamaño de la muestra será denotado por *n*.

Ejemplos de muestras

En la población a) del ejemplo anterior, una muestra puede ser: 100 tiendas de sillas con las características indicadas, seleccionadas aleatoriamente.

En la población b), una muestra es: En cada uno de los estados: Sinaloa, Nayarit, Jalisco, Colima, Michoacán, Guerrero, Oaxaca y Chiapas, se capturaron 127 murciélagos Phyllostomidae en el periodo indicado.

En la población c), una muestra es: 50 adultos con las características indicadas, elegidos al azar en la colonia Agrícola Oriental en el período señalado.

D. Variables y datos

Variables estadísticas

Las características de interés en una población o una muestra se llaman variables. Por ejemplo, son variables estadísticas:

- *El tiempo que te lleva trasladarte de tu casa a la escuela.*
- *El número de palabras que lees por minuto.*
- *Las carreras que elegirán los estudiantes de tu grupo de sexto semestre.*
- *El número de habitantes en el hogar de cada estudiante.*
- *La última película que vio en el cine cada estudiante de un grupo.*

Es claro que estas características toman distintos valores en los individuos de la población o muestra.

Datos

Son los valores que toma una variable de estudio en cada individuo de la muestra o de la población. En los ejemplos de variables mencionados en la parte anterior, se pueden presentar los siguientes datos:

- *26, 28, 32, 35, 25 minutos*
- *100, 105, 120, 110 palabras*
- *Medicina, Veterinaria, Sociología, Filosofía, Actuaría, etcétera*
- *2, 3, 4, 5, 6 personas*
- *Rapsodia Bohemia, Aquaman, etcétera.*

E. Recopilación de datos

Los datos estadísticos se obtienen por *levantamiento* o por *experimentación*.

El levantamiento de datos se puede llevar a cabo aplicando un cuestionario a través de entrevistas personales, telefónicas o mediante la web, mecanismos que se aplican cuando la población está formada por personas. También se puede recurrir a la observación directa, por ejemplo, tomar los signos vitales de pacientes o medir el tamaño promedio de las hojas de ciertas plantas.

En las situaciones descritas, el investigador registra o mide sistemáticamente, características y comportamientos que se presentan en el entorno, sin modificar a voluntad propia ninguno de los factores que intervienen en el proceso.

Si el levantamiento de información se aplica a todos los elementos de la población de estudio, se habla de un ***censo***.

La otra forma de recopilar datos es a través de estudios experimentales, es decir, cuando el investigador modifica por lo menos una de las variables del fenómeno en estudio. Por ejemplo, supongamos que se desea medir los efectos en la piel de los rayos ultravioleta emitidos por el sol. Se toma una muestra de 100 individuos que serán expuestos al sol durante 15 minutos, 50 individuos usarán un protector solar y la otra mitad no. Éste es un experimento porque se ha decidido modificar la variable uso de protector solar y se debe asignar aleatoriamente quiénes formarán parte de cada grupo.

La variabilidad de los datos estadísticos se debe a la presencia del azar en los fenómenos que se estudian o en la elección de la muestra.

F. Ramas de la estadística

Se pueden distinguir dos grandes ramas en la estadística: la Estadística descriptiva y la Estadística inferencial.

Estadística descriptiva

Esta rama incluye un conjunto de técnicas para recopilar, ordenar, organizar, resumir y presentar datos de manera que resalten sus características más importantes, lo cual permite hacer observaciones y extraer conclusiones. Utiliza tablas, gráficas y medidas estadísticas.

Aunque las técnicas de la estadística descriptiva son útiles para cualquier colección de datos (de hecho surgieron del tratamiento de información proveniente de censos), en la actualidad estas técnicas se suelen usar para trabajar con información proveniente de muestras.

Estadística Inferencial

Se trata de procedimientos que permiten obtener ciertas conclusiones acerca de una población con base exclusivamente en la información proporcionada por una muestra aleatoria representativa.

El azar interviene en la elección de la muestra por lo que debe ser tomado en cuenta al hacer inferencias, y es aquí donde la Estadística se relaciona con la *Probabilidad*, que es la rama de las matemáticas encargada de la toma de decisiones en condiciones de incertidumbre.

EJERCICIOS 1.1.1

1. Para organizar una campaña que promueva el deporte entre niños y adolescentes, el gobierno del estado de Chihuahua decidió realizar una encuesta entre alumnos de escuelas primarias y secundarias.
Se eligieron 300 escuelas al azar y se aplicó una encuesta a todos los estudiantes de esas escuelas.

a. ¿Cuál es la población de este estudio?
b. ¿Cuál es la muestra?
c. ¿Qué procedimiento se utilizó para recabar información?

2. Selecciona solo una opción que corresponda al concepto descrito en cada uno de los siguientes casos

2.1 El proceso de recoger, organizar y representar los datos demográficos de los estudiantes de un salón de clase corresponde a un estudio de Estadística

a) Inferencial b) Descriptiva c) Muestreo d) Probabilidad

2.2 El total de elementos bajo consideración al hacer una investigación estadística se llama

a) Población b) Muestra c) Selección d) Variable

2.3 La parte de la población escogida para hacer un análisis estadístico que permita obtener algunas conclusiones sobre la población se llama

a) Población b) Muestra c) Selección d) Variable

2.4 Los métodos que permiten obtener ciertas conclusiones sobre una población con base en la información que brinda una muestra, integran la Estadística

a) Inferencial b) Descriptiva c) Muestreo d) Probabilidad

2.5 Una característica de interés en un estudio estadístico es una

a) Población b) Muestra c) Selección d) Variable

3. Escribe de qué población fueron extraídas las muestras que se describen a continuación.

a. Se elige a 250 mujeres de 15 años y más que asisten a la clínica 22 del IMSS para un estudio sobre el número de hijos que tienen.
b. Se seleccionan al azar 600 estudiantes matriculados este semestre en el CCH Sur para un estudio sobre hábitos de tabaquismo.
c. Se aplica una encuesta telefónica (en teléfonos fijos) a 800 mexicanos con credencial del INE vigente para que califiquen el desarrollo de las anteriores elecciones presidenciales.

1.2 TIPOS DE VARIABLES

Las variables estadísticas pueden clasificarse de acuerdo a los valores que toman, en variables cuantitativas y variables cualitativas.

A. Variables cuantitativas o numéricas

Son aquellas variables que toman valores numéricos como resultado de un proceso de **conteo o medición**. Por ejemplo:

a) Peso de jóvenes mexicanos de 15 a 18 años de edad.
b) Cantidad de personas que viven en el hogar de cada estudiante del CCH.
c) Edad en años cumplidos de los estudiantes de un grupo.
d) Altura de los arbustos de cedro blanco que crecen en la CdMx.

Estas variables se subdividen en

Cuantitativas discretas

Son resultado de un proceso de conteo. Usualmente toman valores enteros no negativos. De los ejemplos mencionados antes, las variables de los incisos b y c son discretas.

Cuantitativas continuas

Son resultado de un proceso de medición. Toman valores en intervalos. De los ejemplos mencionados antes, las variables de los incisos a y d son continuas.

B. Variables cualitativas o categóricas

Son las variables que toman como valores categorías o nombres que identifican distintas cualidades o atributos de los elementos de la población o muestra. Por ejemplo:

a) Género de personas de una población.
b) Color de ojos de estudiantes del plantel.
c) Nivel máximo de estudios de trabajadores de una empresa.
d) Nivel socioeconómico de personas que habitan en cierta región.

Estas variables se subdividen en

Cualitativas nominales

Son variables cuyos valores no tienen un orden natural. De los ejemplos de variables cualitativas mencionados antes, las de los incisos a y b son nominales.

Cualitativas ordinales

Son variables cuyos valores sí tienen un orden natural. De los ejemplos anteriores, las variables de los incisos c y d son ordinales.

EJERCICIOS 1.2.1

En cada uno de los siguientes casos, escribe una lista de posibles valores de cada variables y determina de qué tipo de variable se trata.

Variable	Valores	Tipo de variables
a) Opinión sobre un maestro del CCH.		
b) Cantidad de café que sirve una máquina automática en una descarga si se anuncia que es de 300 ml.		
c) Cantidad de libros que un estudiante consulta en la biblioteca en un semestre.		
d) Carreras que eligen estudiantes de 6° semestre.		
e) Peso del contenido de las cajas de cereal que indican 800 gr.		
f) Tipo de medalla obtenida por los tres mejores deportistas de una prueba.		

1.3 TABLAS DE DISTRIBUCIÓN DE FRECUENCIAS

Ejemplo 1.

Calificaciones obtenidas por 48 alumnos en Inglés

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

Escribe dos observaciones sobre los datos anteriores.

_____

En la siguiente tabla, se escribió cada calificación junto al número de alumnos que obtuvo cada calificación.

Calif	*Alumnos*
5	*5*
6	*8*
7	*9*
8	*11*
9	*9*
10	*6*
Total	*48*

Ahora escribe 4 observaciones sobre la colección de calificaciones.

_____

_____

_____

_____

¿Cómo te resultó más sencillo analizar los datos anteriores: revisados en la lista original u ordenados en la tabla? _____

Una tabla de frecuencias es una tabla que muestra los valores que toma una variable, junto con el número de veces que se observa cada uno de ellos en una colección de datos. A este número se le llama ***frecuencia*** o ***frecuencia absoluta.***

El objetivo de construir esta tabla es obtener una presentación sencilla, ordenada y fácil de leer, que permita distinguir las características más evidentes de una colección de datos y proporcionar elementos para su análisis.

A. Datos no agrupados en intervalos

Este tipo de tablas se usan cuando la variable de estudio es cualitativa o bien cuantitativa discreta con pocos valores. La restricción de que los valores sean pocos se debe a que una tabla con muchos renglones no es fácil de leer.

La tabla de frecuencias más sencilla es la formada por dos columnas, una donde se escriben las categorías o valores de la variable y otra donde se escribe la frecuencia con que aparece cada valor, como la tabla del ejemplo inicial de esta sección.

Ejemplo 2.

Las calificaciones de dos grupos de Estadística y Probabilidad I, se resumen en las siguientes tablas.

Grupo A

Calificación	*Frecuencia*
5	*8*
6	*9*
7	*8*
8	*13*
9	*7*
10	*5*
Total	*50*

Grupo B

Calificación	*Frecuencia*
5	*5*
6	*6*
7	*7*
8	*7*
9	*3*
10	*2*
Total	*30*

Un problema para comparar el desempeño académico de estos dos grupos con base en sus calificaciones, es que los grupos son de diferente tamaño. Por tanto, se requiere calcular otra cantidad que tome en cuenta el tamaño de cada grupo, por ejemplo, porcentajes.

Grupo A

Calificación	*Frecuencia*	*Porcentaje*
5	*8*	*16.00%*
6	*9*	*18.00%*
7	*8*	*16.00%*
8	*13*	*26.00%*
9	*7*	*14.00%*
10	*5*	*10.00%*
Total	*50*	*100.00%*

Grupo B

Calificación	*Frecuencia*	*Porcentaje*
5	*5*	*16.67%*
6	*6*	*20.00%*
7	*7*	*23.33%*
8	*7*	*23.33%*
9	*3*	*10.00%*
10	*2*	*6.67%*
Total	*30*	*100.00%*

Observa que el porcentaje de alumnos que no aprobaron, es similar en los dos grupos.

Respecto a los estudiantes que obtuvieron calificaciones aprobatorias bajas (6 o 7) el porcentaje en el grupo A es 18 + 16 = 34%, y en el grupo B es de 20 + 23.33 = 43%.

Por otro lado, considerando los estudiantes que obtuvieron calificaciones de 8, 9 o 10, en el grupo A el porcentaje es 26 + 14 + 10 = 50%, mientras que en el grupo B el porcentaje es 23.33 + 10 + 6.67 = 40%.

Así que podemos decir que las calificaciones fueron mejores en el grupo A que en el grupo B.

El cociente de la frecuencia entre el total de datos, se llama ***frecuencia relativa***.

En el siguiente ejemplo identificaremos todas las frecuencias que se incluyen en una tabla completa para facilitar el análisis de la información.

Ejemplo 3.

El número de hermanos que tienen los alumnos de un grupo del CCH Sur, se recoge en la siguiente lista.

1	*1*	*2*	*1*	*0*	*2*	*0*	*1*	*3*	*0*	*1*	*1*	*2*	*6*	*3*	*1*	*0*	*1*
2	*2*	*0*	*3*	*4*	*0*	*4*	*1*	*2*	*1*	*2*	*1*	*5*	*0*	*2*	*2*	*1*	*2*

*Los **valores** que toman los datos serán representados por $x_1, x_2, \ldots, x_r$.*

*La **frecuencia absoluta** del valor x_i, se denota por f_i.*

Hermanos (x_i)	*Frecuencia absoluta (f_i)*
0	*7*
1	*11*
2	*9*
3	*5*
4	*2*
5	*1*
6	*1*
Total	*36*

*La **frecuencia relativa** es la frecuencia absoluta dividida entre la cantidad de datos. Se representa con fr_i, y se calcula así:* $fr_i = \frac{f_i}{n}$. Se puede escribir como fracción, como decimal o como porcentaje

Hermanos (x_i)	*Frecuencia absoluta (f_i)*	*Frecuencia relativa (fr_i)*
0	*7*	*7/36 =0.194*
1	*11*	*0.3056*
2	*9*	*0.2500*
3	*5*	*0.1389*
4	*2*	*0.0556*
5	*1*	*0.0278*
6	*1*	*0.0278*
Total	*36*	*1*

Para tener más elementos de análisis, se calculan las frecuencias acumuladas.

*La **frecuencia absoluta acumulada** hasta un valor x_i es la suma de las frecuencias absolutas de todos los valores menores o iguales a x_i, y se representa por Fa_i.*

*La **frecuencia relativa acumulada** hasta un valor x_i es la suma de las frecuencias relativas de todos los valores menores o iguales a x_i, y se representa por Fra_i. También se puede calcular dividiendo las frecuencias absolutas acumuladas entre el total de datos.*

Hermanos x_i	*Frecuencia absoluta f_i*	*Frecuencia relativa fr_i*	*Frecuencia absoluta acumulada Fa_i*	*Frecuencia relativa acumulada Fra_i*
0	*7*	*7/36 =0.1944*	*7*	*7/36 =0.1944*
1	*11*	*0.3056*	*7 + 11=18*	*18/36 =0.500*
2	*9*	*0.2500*	*7 + 11 + 9 =27*	*27/36 = 0.750*
3	*5*	*0.1389*	*32*	*0.8888*
4	*2*	*0.0556*	*34*	*0.9444*
5	*1*	*0.0278*	*35*	*0.9722*
6	*1*	*0.0278*	*36*	*1.0000*
Total	*36*	*1.0000*		

Ahora, ya que tenemos la distribución de frecuencias, ¿qué información podemos obtener acerca de las estaturas de los alumnos?

Interpretemos algunos valores de cada columna:

f_3 Nueve estudiantes de 36, tienen 2 hermanos

fr_3 El 25% de los estudiantes encuestados tienen 2 hermanos

Fa_3 Treinta y dos de 36 estudiantes tienen a lo más 2 hermanos

Fra_3 El 75% de los estudiantes encuestados tienen 2 hermanos o menos.

Algunas observaciones sobre todos los datos de la tabla son:

- *Las mayores frecuencias se concentran en 0, 1 y 2 hermanos, y las menores frecuencias corresponden a 5 y 6 hermanos.*
- *Hay 7 encuestados que tienen 3 o 4 hermanos.*
- *El 50% de los estudiantes encuestados tienen uno o ningún hermano.*
- *La mayoría de los encuestados tienen 2 o menos hermanos.*
- *Sólo el 25% de los estudiantes encuestados tienen más de 2 hermanos.*

EJERCICIOS 1.3.1

1. La cuenta de la luz (en pesos) del mes de marzo de 30 familias escogidas aleatoriamente, se muestra a continuación.

230	560	340	485	870	560	370	340	560	450	230	340	340	485	450
450	450	870	560	450	340	230	970	70	870	485	560	1120	370	870

a. Completa la tabla de distribución de frecuencias siguiente.

Cantidad a pagar x_i	Frecuencia absoluta f_i	Frecuencia relativa fr_i	Frecuencia absoluta acumulada Fa_i	Frecuencia relativa acumulada Fra_i
70		0.033	1	
230	3	0.100		0.133
340			9	
	2	0.067		
	5	0.167		
485		0.100		
560			24	
	4	0.133		
	1			
1120		0.033		1.000
Total	30			

b. Escribe 4 observaciones sobre la información que resume la tabla.

1. _____
2. _____
3. _____
4. _____

A. Datos agrupados en intervalos

Es conveniente, e incluso necesario, agrupar datos en intervalos cuando:

- Se tiene una variable numérica discreta con una gran variedad de valores distintos, o bien
- Se tiene una variable numérica continua.

Por lo general la distribución de frecuencias debe tener como mínimo 5 intervalos, pero no más de 15, pues el objetivo es hacer una presentación resumida de la información que sea sencilla y permita distinguir características importantes.

Vamos a ir desarrollando los conceptos necesarios para construir una tabla de datos agrupados, usando un ejemplo. Consideremos los siguientes datos que corresponden a la edad de 55 personas.

27	23	41	38	44	29	35	26	18	22	24
25	36	22	52	31	30	22	45	28	18	20
18	28	44	25	29	28	24	36	21	23	32
26	33	25	27	25	34	32	23	54	38	23
31	23	26	48	16	27	27	33	29	29	28

Si se quisiera hacer una tabla de frecuencias sin agrupar los datos, serían necesarios 20 renglones diferentes, uno por cada valor de los datos. Es claro que conviene agrupar los datos en intervalos para reducir esa cantidad.

Cantidad de intervalos o clases

El número de intervalos o clases depende de la cantidad de datos que se tengan. Aunque no existe una regla única para determinar el número de intervalos, universalmente aceptada, hay algunas reglas empíricas que resultan útiles en esta decisión. En todos los casos, el número que se obtiene debe ser considerado como

una sugerencia de la cantidad de intervalos a ocupar, misma que puede modificarse un poco si eso es útil para que los intervalos resulten sencillos y fáciles de leer.

a) Una posibilidad es basar la decisión en la siguiente tabla:

Número de datos	Cantidad de intervalos
De 10 a 15	4 o 5
De 16 a 30	5 o 6
De 31 a 50	6 o 7
De 51 a 75	7 u 8
De 76 a 100	8 o 9
De 100 a 150	9 o 10
Más de 150	De 11 a 15

En el ejemplo tenemos 55 datos, así que la tabla anterior sugiere 7 u 8 intervalos.

b) Otra posibilidad es usar la **regla de Sturges** para estimar el número ideal de intervalos *(k).* Si la cantidad de datos es *n*, esta regla consiste en aplicar la fórmula:

$$k = 1 + 3.322Log(n)$$

donde *log(n)* es el logaritmo decimal del número de datos. En el ejemplo tenemos:

$$k = 1 + 3.322Log(55) = 1 + 3.322(1.74) = 1 + 5.78 = 6.78$$

Y redondeando obtenemos que la sugerencia es tomar 7 intervalos.

Longitud o amplitud de los intervalos o clases

Todos los intervalos deben tener la misma longitud para que reflejen cuántos datos caen en subintervalos del mismo tamaño. Sin embargo, es posible dejar abierto el primero o el último de los intervalos, criterio que se usa con frecuencia en estudios demográficos. Por ejemplo, se puede poner “80 y más” en una tabla sobre edades.

Para tener una referencia sobre la longitud de los intervalos, se requiere calcular el rango de los datos, definido como la diferencia del mayor menos el menor de los valores que toman.

Rango = dato máximo – dato mínimo

El rango se divide entre el número de intervalos para obtener la longitud sugerida *(c).*

$$c = \frac{Rango}{k}$$

En nuestro ejemplo, el dato máximo es 54 y el mínimo es 16, así que la amplitud de los intervalos es:

$$c = \frac{54 - 16}{7} = \frac{38}{7} = 5.43$$

Para que los intervalos sean fáciles de leer e interpretar, tomaremos una longitud entera, que puede ser c = 5.

Características de los intervalos o clases

Los intervalos deben cumplir que:

a) Son de la misma longitud.
b) Cubren todo el rango de los datos.
c) No se traslapan o enciman, es decir, no hay datos que puedan contarse en dos intervalos distintos.

Vamos a construir los intervalos del ejemplo considerando estas características. Sumando 5 unidades a partir del dato mínimo, el inicio de los intervalos quedará así:

Intervalo o Clase
16 –
21 –
26 –
31 –
36 –
41 –
46 –

Pero aumentando 5 unidades a 46, llegamos a 51 y el dato máximo es 54. Entonces, con 7 intervalos de amplitud 5 no alcanzamos a cubrir todos los datos.

Para resolver lo anterior, tenemos dos opciones:

• Tomar una amplitud de 6 unidades para cada intervalo (en lugar de 5). Entonces, los 7 intervalos abarcarían un total de 42 unidades, por lo que sí se cubre el rango que es de 38.

• Tomar 8 intervalos y no 7. En este caso quedarían 8 intervalos de amplitud 5, que da un total de 40, que también cubre el rango, que es de 38.

Vamos a adoptar la segunda de estas opciones ya que 40 es más cercano al rango que 42. Así que construiremos 8 intervalos de amplitud 5.

Ahora debemos considerar que un intervalo termina donde inicia el siguiente para no dejar huecos sin cubrir. Tenemos que el segundo intervalo comienza en 21, por lo que el primer intervalo en nuestro caso será de 16 a 21, mientras que el segundo intervalo será de 21 a 26, y así sucesivamente. Tendremos los siguientes intervalos

Intervalo o Clase
16 – 21
21 – 26
26 – 31
31 – 36
36 – 41
41 – 46
46 – 51
51 – 56

Pero estos intervalos no cumplen la característica c), porque los extremos como 21, 26, 31, etc quedan en dos intervalos. Entonces, si hay datos que tengan estos valores, ¿dónde los contabilizamos?

Hay varias formas de resolver esto. Algunas de ellas son:

- Usar intervalos abiertos por un lado y cerrados por el otro. Donde se coloca un corchete, indica que el intervalo incluye al extremo y se dice que es cerrado por ese lado. Donde se coloca un paréntesis, indica que el intervalo abarca números menores que el extremo, sin incluir a dicho extremo, y se dice que es abierto por ese lado. Así, por ejemplo, el intervalo

  $$[16, 21)$$

  indica que contamos desde el 16 y que no llegamos a contar el 21. Entonces, el 21 se contará únicamente en el siguiente intervalo, $[21, 26)$.

- Usar números decimales. Aunque los datos en este caso son enteros, para indicar que el extremo derecho no se incluye en cada intervalo, podemos escribir los intervalos así:

16 – 20.5
21 – 25.5
26 – 30.5
⋮

- En ejemplos como el anterior en el que los datos toman valores enteros, también se pueden tomar intervalos cuyo valor inicial sea una unidad mayor que el valor final del intervalo anterior.

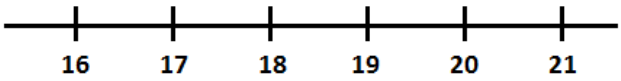
16 – 21
22 – 27
28 – 33
⋮

Aquí vamos a usar la primera opción, así que los intervalos quedan como sigue:

Intervalo o Clase
[16 – 21)
[21 – 26)
[26 – 31)
[31 – 36)
[36 – 41)
[41 – 46)
[46 – 51)
[51 – 56)

Y se cumple:

a) Los intervalos son de igual longitud porque hay cinco unidades en cada uno de ellos.



b) Cada intervalo comienza donde termina el anterior.

c) Se cubre todo el rango porque el dato mayor es 54 y el intervalo termina en 56.

Observa que en el último intervalo sobran 2 unidades. Podemos repartir ese sobrante entre el primero y el último intervalo, para obtener finalmente.

Intervalo o Clase
[15 – 20)
[20 – 25)
[25 – 30)
[30 – 35)
[35 – 40)
[40 – 45)
[45 – 50)
[50 – 55]

Se suele usar el último intervalo cerrado de ambos lados para que cuando este termine justo en el dato mayor, no sea necesario incluir un intervalo más.

Construcción de la tabla

Una vez construidos los intervalos, las demás columnas de la tabla se trabajan igual que en el caso de los datos no agrupados en intervalos. La frecuencia absoluta de un intervalo es el número de datos que caen dentro del intervalo.

En el ejemplo tenemos, los datos 16, 18, 18, 18 y 20 caen en el primer intervalo, así que la frecuencia absoluta de ese intervalo es 5. De igual manera, se calculan las demás frecuencias.

Intervalo o Clase	**Frecuencia**
$[15-20)$	**5**
$[20-25)$	**15**
$[25-30)$	**16**
$[30-35)$	**8**
$[35-40)$	**4**
$[40-45)$	**4**
$[45-50)$	**1**
$[50-55]$	**2**
Total	**55**

La frecuencia relativa de cada intervalo, es el cociente entre su frecuencia absoluta y el número de datos.

Intervalo o Clase	Frecuencia absoluta	**Frecuencia Relativa**
$[15-20)$	5	$\frac{5}{55} = \mathbf{0.091}$
$[20-25)$	15	$\frac{15}{55} = \mathbf{0.273}$
$[25-30)$	16	$\frac{16}{55} = \mathbf{0.291}$
$[30-35)$	8	$\frac{8}{55} = \mathbf{0.145}$
$[35-40)$	4	$\frac{4}{55} = \mathbf{0.073}$
$[40-45)$	4	$\frac{5}{55} = \mathbf{0.073}$
$[45-50)$	1	$\frac{1}{55} = \mathbf{0.018}$
$[50-55]$	2	$\frac{2}{55} = \mathbf{0.036}$
Total	55	**1**

La frecuencia absoluta acumulada se construye sumando las frecuencias absolutas hasta cada uno de los intervalos.

Intervalo o Clase	Frecuencia	Frecuencia Relativa	**Frecuencia Acumulada**
$[15-20)$	5	0.091	$\mathbf{5}$
$[20-25)$	15	0.273	$\mathbf{5+15=20}$
$[25-30)$	16	0.291	$\mathbf{5+15+16=36}$
$[30-35)$	8	0.145	$\mathbf{5+15+16+8=44}$
$[35-40)$	4	0.073	$\mathbf{5+15+16+8+4=48}$
$[40-45)$	4	0.073	$\mathbf{5+15+16+8+4+4=52}$
$[45-50)$	1	0.018	$\mathbf{5+15+16+8+4+4+1=53}$
$[50-55]$	2	0.036	$\mathbf{5+15+16+8+4+4+1+2=55}$
Total	55	1	

Y la frecuencia relativa acumulada se construye sumando las frecuencias relativas o dividiendo la frecuencia absoluta acumulada entre la cantidad de datos.

Intervalo o Clase	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	**Frecuencia Acumulada Relativa**
$[15-20)$	5	0.091	5	$\mathbf{0.091}$ **O bien** $\mathbf{\frac{5}{55}=0.091}$
$[20-25)$	15	0.273	20	$\mathbf{0.091+0.273=0.364}$ **O bien** $\mathbf{\frac{20}{55}=0.364}$
$[25-30)$	16	0.291	36	$\mathbf{0.091+0.273+0.291=0.655}$ **O bien** $\mathbf{\frac{36}{55}=0.655}$
$[30-35)$	8	0.145	44	$\mathbf{0.091+0.273+0.291+0.145=0.800}$ **O bien** $\mathbf{\frac{44}{55}=0.800}$
$[35-40)$	4	0.073	48	$\mathbf{0.091+0.273+0.291+0.145+0.073=0.873}$ **O bien** $\mathbf{\frac{48}{55}=0.873}$
$[40-45)$	4	0.073	52	$\mathbf{0.091+0.273+0.291+0.145+0.073+0.073=0.946}$ **O bien** $\mathbf{\frac{52}{55}=0.945}$
$[45-50)$	1	0.018	53	$\mathbf{0.091+0.273+0.291+0.145+0.073+0.073+0.018=0.964}$ **O bien** $\mathbf{\frac{53}{55}=0.964}$
$[50-55]$	2	0.036	55	$\mathbf{0.091+0.273+0.291+0.145+0.073+0.073+0.018+0.036=1}$ **O bien** $\mathbf{\frac{55}{55}=1}$
Total	55	1		

Así obtenemos finalmente la distribución de frecuencias de los datos del ejemplo, agrupándolos en intervalos:

Intervalo o Clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
$[15-20)$	5	0.091	5	0.091
$[20-25)$	15	0.273	20	0.364
$[25-30)$	16	0.291	36	0.655
$[30-35)$	8	0.145	44	0.800
$[35-40)$	4	0.073	48	0.873
$[40-45)$	4	0.073	52	0.945
$[45-50)$	1	0.018	53	0.964
$[50-55]$	2	0.036	55	1
Total	**55**	**1**		

Siempre debemos hacer un primer análisis de la información al terminar la construcción de una tabla de frecuencias. En este caso se observa que:

- Las edades más frecuentes están entre los 20 y los 30 años, pues en este rango se ubica más del 56% de los datos.
- Las edades menos frecuentes son las que se ubican entre los 45 y los 55 años, mismas que abarcan más o menos el 5% de los datos.
- De las 55 personas, hay 44 que tienen menos de 35 años, lo que representa un 80% de los encuestados.
- Casi el 95% de los encuestados son menores de 45 años, lo que representa a 52 personas.
- Solo el 9% son menores a 20 años, lo que corresponde a 5 personas.

EJERCICIOS 1.3.2

1. Los siguientes datos muestran el número de vuelos internacionales recibidos en el aeropuerto de la ciudad de México durante los dos meses anteriores. Construye la distribución de frecuencias.

71	47	66	67	73	38	63	67	29	54	62	70
63	37	68	50	59	60	45	48	52	49	48	56
70	62	61	65	62	45	62	56	63	39	36	43
49	50	39	41	57	49	73	47	38	61	48	31
55	57	72	53	42	70	56	58	39	60	53	36

Intervalo o clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada

2.- Escribe al menos 4 observaciones sobre la información que brinda la tabla anterior.

a. _____

b. _____

c. _____

d. _____

3.- Los datos siguientes corresponden a un estudio realizado con 40 personas para conocer la reacción sistémica a la picadura de abeja. Se toma el tiempo, en minutos y décimas de minuto, en el que aparecen las primeras reacciones a la picadura.

10.5	11.2	9.9	11.4	12.7	16.5	15.0	10.1
12.7	11.4	11.6	7.9	8.3	10.9	6.2	8.1
3.8	10.5	11.7	12.5	11.2	9.1	8.4	10.4
9.1	13.4	12.3	11.4	8.8	7.4	5.9	8.6
13.6	14.7	11.5	10.9	9.8	12.9	11.5	9.9

a. Construye una tabla de distribución de frecuencias en tu cuarderno.

b. Escribe los intervalos que usaste en tu tabla _____

_____

4. Escribe al menos 4 observaciones sobre la información resumida en la tabla del ejercicio 3.

a. _____

b. _____

c. _____

d. _____

5. La siguiente tabla muestra la distribución de frecuencias de los resultados obtenidos al entrevistar a 300 estudiantes de bachillerato que trabajan mientras estudian.

Intervalo de Clase (Ganancia Semanal)	**Frecuencia Simple**	**Frecuencia Relativa**	_____ _____	_____ _____ _____
$[300 - 500)$	105	____		
$[500 - 600)$	____	0.300		
$[600 - 700)$	45	____		
$[700 - 800]$	____	____		

Completa la tabla anterior, y con base en ella proporciona la información que falta:

a. La frecuencia absoluta del primer intervalo nos dice que: _____ _____

b. El 30% de los estudiantes ganan entre _____ y _____.

c. La frecuencia acumulada de la cuarta clase quiere decir que: _____ _____

d. El porcentaje de estudiantes que ganan máximo $699.5 es de: _____ _____.

1.4 REPRESENTACIÓN GRÁFICA

Además de la distribución de frecuencias, resulta conveniente construir alguna representación gráfica de los datos. De esta manera, se tiene una imagen que describe visualmente el comportamiento de los datos.

Toda gráfica debe tener un título descriptivo, el nombre de la variable que representa, las unidades de la variable, preferentemente la fuente de la cual se recaba la información y en su caso la escala utilizada.

La siguiente tabla muestra algunas recomendaciones para el uso de gráficas en función al tipo de variable que se analiza:

Tipo de variable	Gráfica recomendada
Cualitativa	Circulares De barras
Cuantitativas discretas con pocos valores	De líneas De barras De puntos
Series de tiempo	De líneas
Cuantitativas discretas con muchos valores Cuantitativas continuas	Histogramas Polígonos de frecuencia De puntos
Cualquier variable con frecuencias acumuladas	Ojiva

Vamos a ver cada uno de los tipos de gráfica mencionados en la tabla anterior.

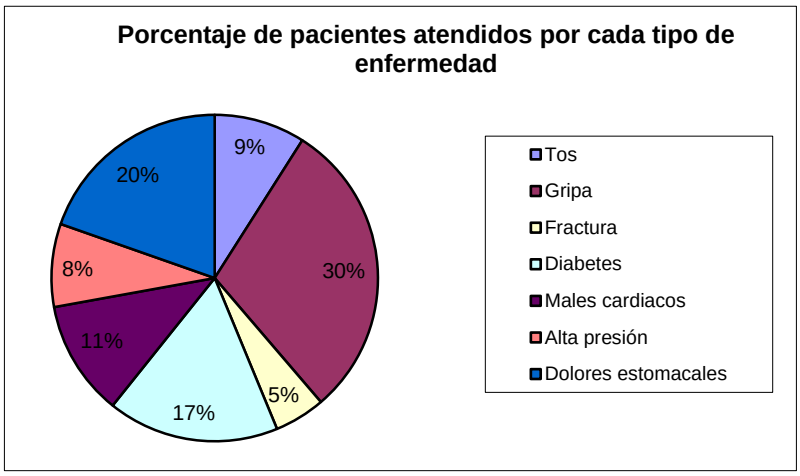
Gráfica Circular

Se conoce también como diagrama de pastel o de sectores. Ayuda a percibir la importancia relativa de cada categoría respecto al total.

Para determinar el ángulo central de cada sector, se divide 360º de manera proporcional a la frecuencia absoluta o relativa de cada valor (usando, por ejemplo, una regla de tres).

En cada sector circular, se suele escribir la frecuencia relativa dada en porcentaje.

Las gráficas circulares se acompañan de una leyenda en la que se indica la categoría que corresponde a cada uno de los sectores.



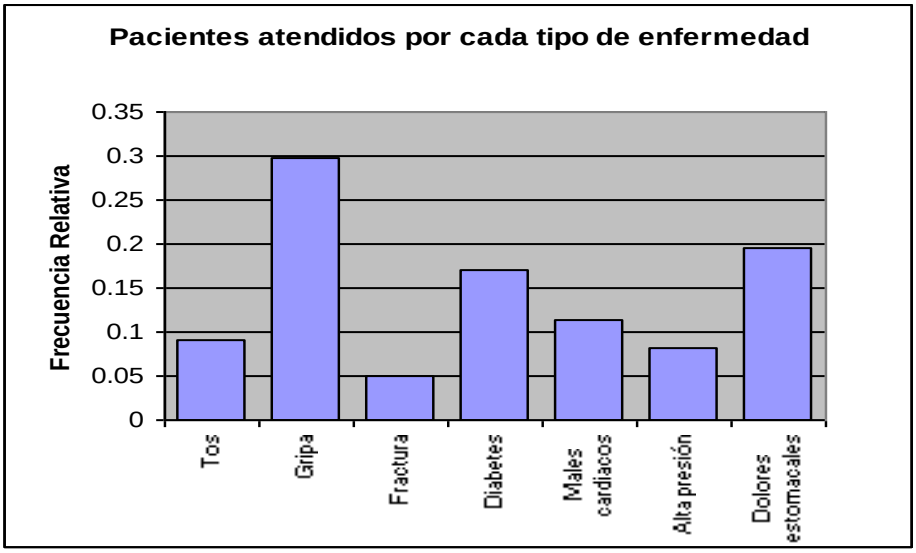
Gráfica de barras

En un sistema de ejes coordenados, se localizan en el eje horizontal los valores de la variable y en el eje vertical, la frecuencia absoluta o relativa que corresponde a cada valor. Con esa información, se construyen barras separadas, una para cada valor.

Las barras son rectángulos cuya altura es la frecuencia de cada valor o categoría y cuyo ancho es arbitrario, pero debe ser el mismo para todos los casos.

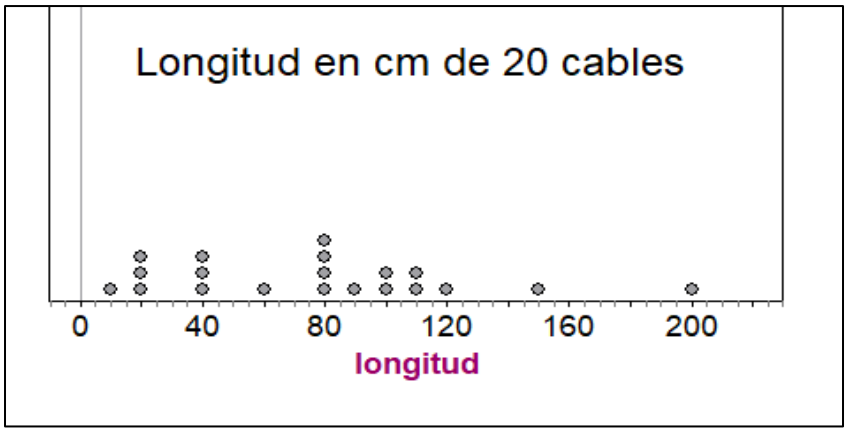
La separación de las barras es arbitraria, pero también debe ser la misma. Las bases de los rectángulos deben estar centrados sobre los valores de la variable.

Este tipo de gráfica se usa en variables cualitativas o cuantitativas discretas con pocos valores.



Gráfica de Puntos

En esta gráfica se identifica cada uno de los datos por un punto trazado sobre su valor a lo largo de una recta numérica, de manera que se observa cada valor individual. Si dos o más datos tienen el mismo valor se colocará un punto sobre otro como se puede observar en la siguiente gráfica:

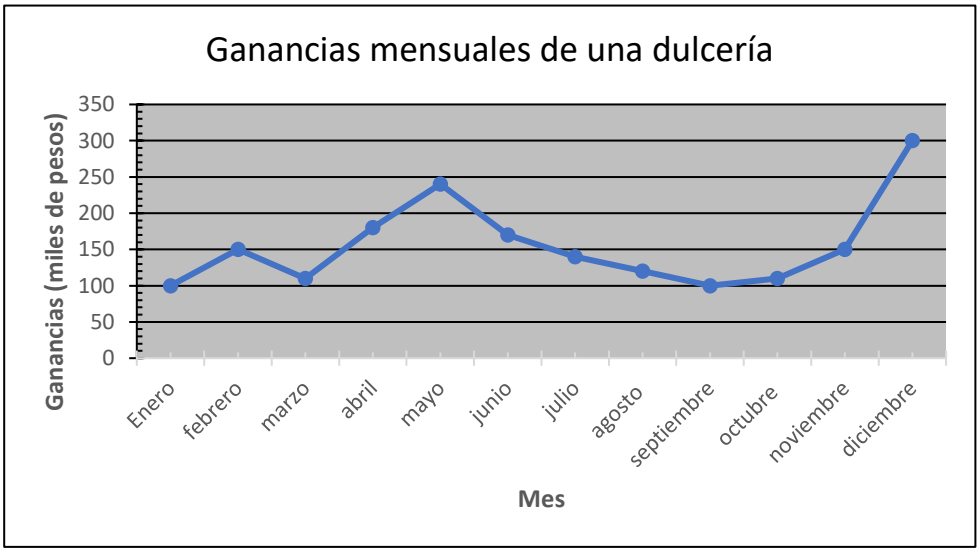


En estas gráficas se muestran la acumulación, variabilidad y la forma de la distribución de los datos. También es útil para comparar dos muestras y aunque es muy fácil de construirlo manualmente es recomendable usar un software cuando se tienen muestras numerosas.

Estas gráficas se usan en cualquier colección de datos cuantitativos.

Gráfica de líneas

Se usa en series de tiempo, es decir, datos que varían en el tiempo, y en datos de variables cuantitativas discretas con pocos valores.

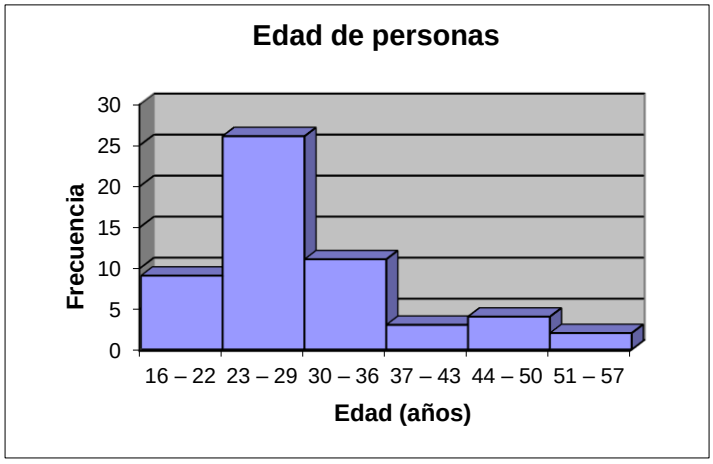


Se trata de una poligonal formada por segmentos de recta que unen una colección de vértices. Cada vértice tiene como abscisa el valor o la categoría y su ordenada es la frecuencia o el dato en el tiempo.

Histograma

Se usa principalmente para datos agrupados en intervalos. Se trata de un gráfico de barras en el que la barras se colocan pegadas, una junto a la otra, pues cada intervalo termina donde empieza el siguiente. Las alturas de las barras pueden ser las frecuencias absolutas o relativas.

El ancho de los rectángulos corresponde al tamaño de los intervalos. Las bases de las barras se encuentran centradas en el punto medio del intervalo, al que llamaremos *marca de clase.*

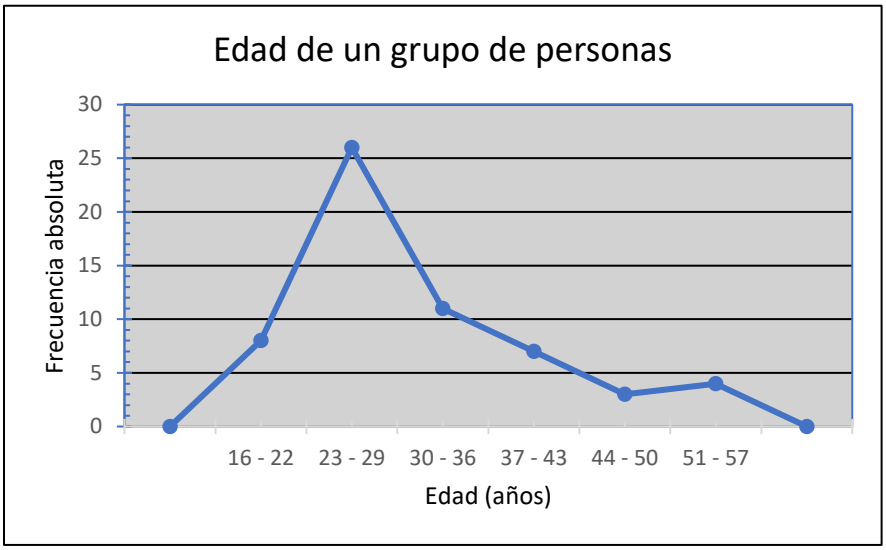


Polígono de Frecuencias

Es un gráfico de líneas que se cierra para formar un polígono trazando segmentos de recta que lo unen con el eje horizontal.

Se usan sobre todo en datos que se pueden agrupar en intervalos. Los vértices tienen como abscisas las marcas de clase o puntos medios de los intervalos, y como ordenadas las frecuencias correspondientes.

Se debe cerrar sobre el eje horizontal en dos puntos que corresponden a las marcas de clase de dos intervalos ficticios a los que se les asigna una frecuencia cero, uno anterior al primer intervalo real y el otro posterior al último intervalo real.

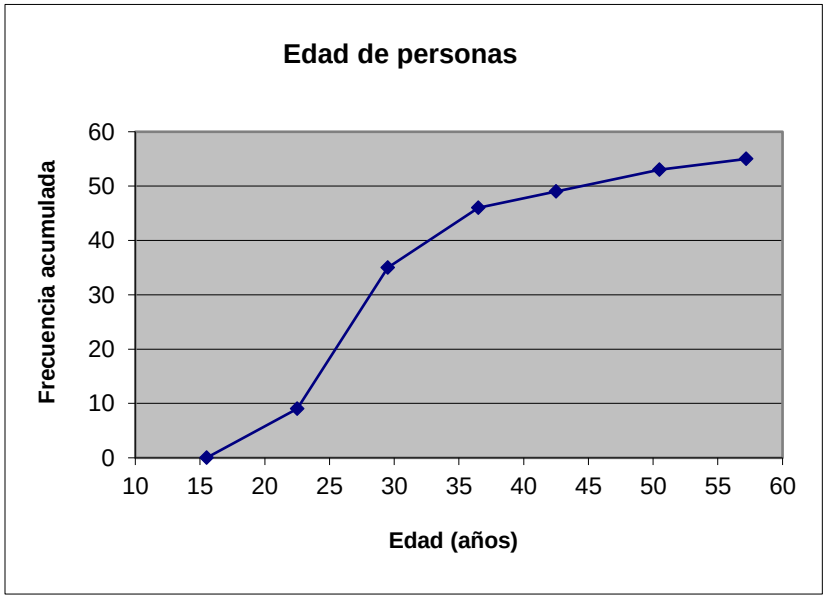


Ojiva

Consiste en una gráfica de líneas en la que la altura de los vértices corresponde a las frecuencias acumuladas. Por ello, la gráfica es ascendente. Siempre empieza en el eje horizontal.

Se usa en datos de cualquier variable que tenga frecuencias acumuladas.

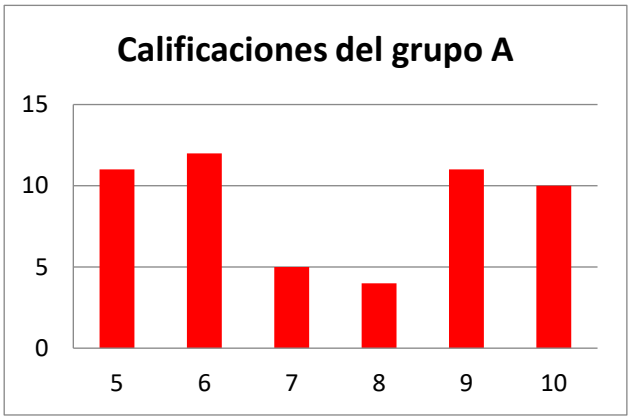
Para datos no agrupados en intervalos, se trazan los puntos que corresponden a los valores de la variable cuantitativa y la frecuencia acumulada (relativa o absoluta), a continuación se unen los puntos mediante segmentos de recta, el extremo derecho no se une con el eje horizontal.

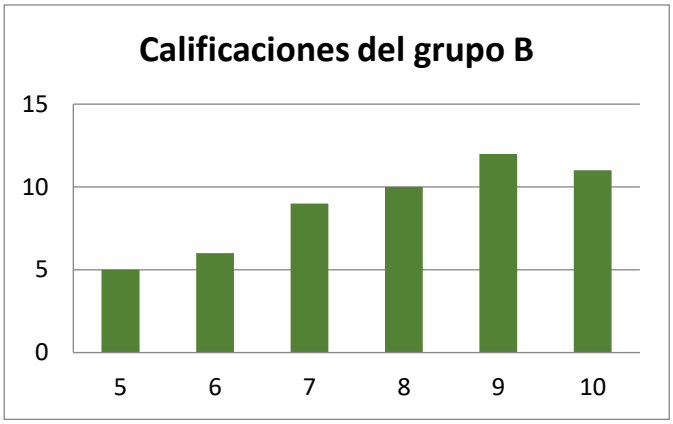


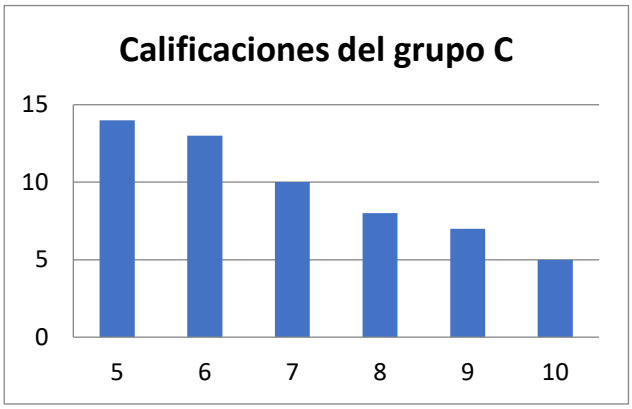
Para datos agrupados en intervalos, los vértices tienen como abscisa a las marcas de clase de cada intervalo.

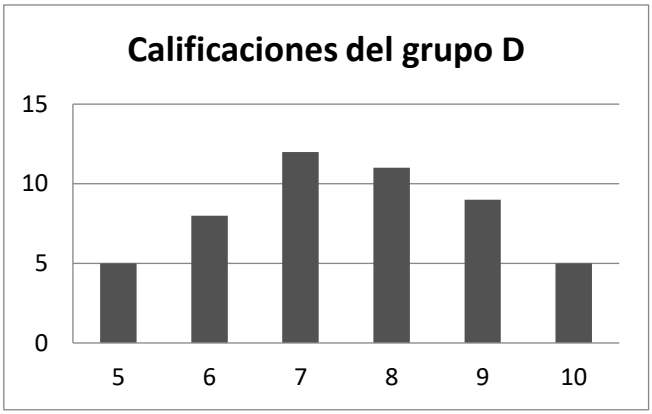
Ejemplo 1.

Analicemos la información que brindan las siguientes gráficas de barras sobre las calificaciones de 4 grupos en la misma asignatura.









Algunas observaciones sobre la información son:

- *En el grupo A las frecuencias mayores están en las calificaciones de los extremos (las más bajas y las más altas), mientras que las calificaciones medias tiene frecuencias bajas.*
- *En el grupo B, la frecuencia de cada calificación desde 5 hasta 9 va en aumento, y las calificaciones altas (8, 9 y 10) tiene las mayores frecuencias.*
- *En el grupo C, las frecuencias van disminuyendo al variar las calificaciones de 5 a 10. Las frecuencias más altas están en las calificaciones bajas (5 y 6).*
- *En el grupo D las calificciones con mayor frecuencia son las medias, y la frecuencia disminuye en las de los extremos.*

Ejemplo 2.

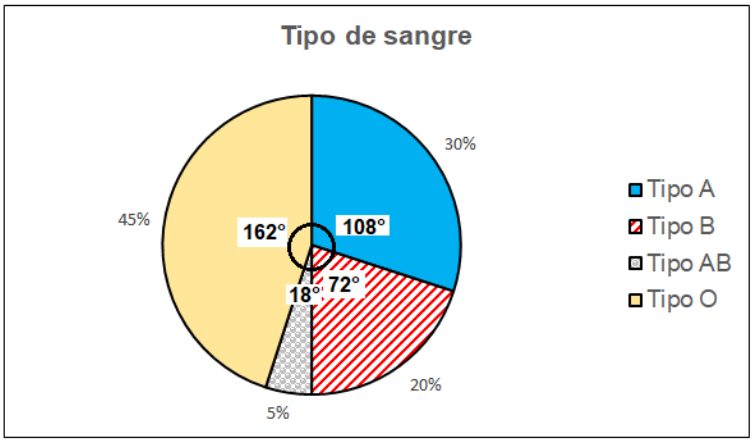
Se realizó un estudio con respecto al grupo sanguíneo de un grupo de asistentes a una clínica y se obtuvieron los siguientes resultados:

Grupo Sanguíneo	*Frecuencia absoluta*	*Frecuencia relativa*
A	6	30%
B	4	20%
AB	1	5%
O	9	45%
Total	20	100%

Vamos a trazar una gráfica circular. Para determinar los ángulos interiores de cada sector circular, calculamos:

30% de 360° = 108° *20% de 360° = 72°*

5% de 360° = 18° *45% de 360° = 162°*

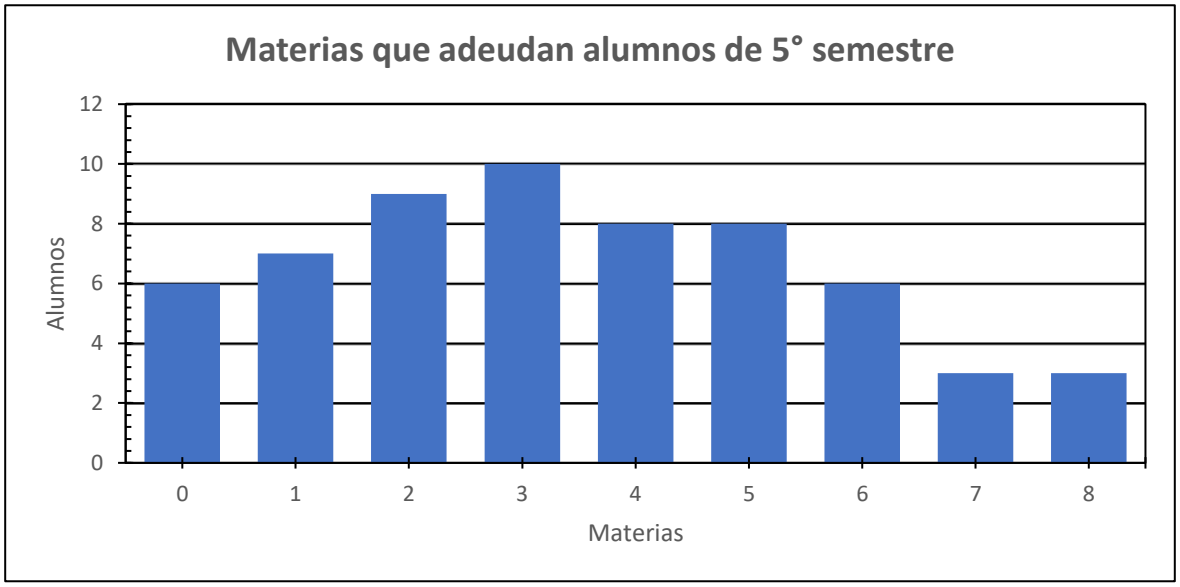


Ejemplo 3.

La siguiente información indica el número de materias que no han acreditado 60 estudiantes de 5° semestre del CCH.

Materias que adeudan	**Estudiantes**
0	6
1	7
2	9
3	10
4	8
5	8
6	6
7	3
8	3
Total	60

Para construir una gráfica de barras, se localizan los valores 0, 1, 2, …, 8 en el eje horizontal y las frecuencias indican la altura de las barras.



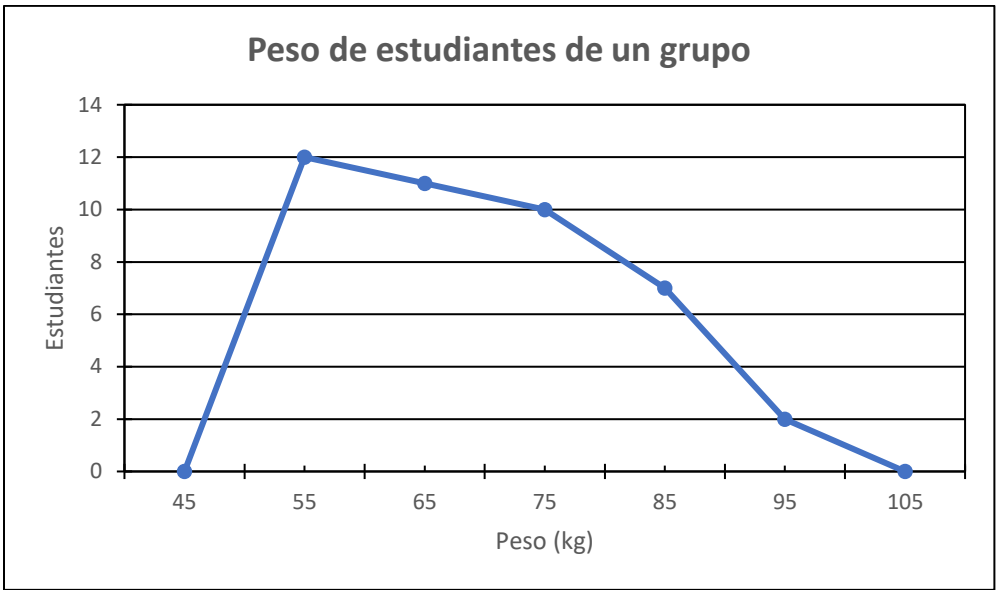
Ejemplo 4.

La siguiente información es el peso de 42 alumnos del CCH.

Peso (kg)	***Marca de clase (kg)***	***Frecuencia***	***Frecuencia acumulada***
[50,60)	*55*	*12*	*12*
[60,70)	*65*	*11*	*23*
[70,80)	*75*	*10*	*33*
[80,90)	*85*	*7*	*40*
[90,100)	*95*	*2*	*42*
Total		*42*	

Para construir el polígono de frecuencias, localizamos los valores de las marcas de clase en el eje horizontal y las frecuencias absolutas en el eje vertical. Luego unimos esos puntos por segmentos de recta y cerramos hacia el eje horizontal.

Para trazar la ojiva, primero localizamos los puntos dados por las marcas de clase y las frecuencias acumuladas. Luego unimos los puntos por segmentos iniciando con un segmento que parta del eje horizontal.



EJERCICIOS 1.4.1

1. Construye los gráficos que se indican a continuación.

a. Un gráfico de barras para el número de hermanos de los alumnos de un grupo del CCH. Los datos correspondientes son:

Hermanos x_i	*Frecuencia absoluta* f_i	*Frecuencia relativa* fr_i
0	*7*	*0.194*
1	*11*	*0.3056*
2	*9*	*0.2500*
3	*5*	*0.1389*
4	*2*	*0.0556*
5	*1*	*0.0278*
6	*1*	*0.0278*
Total	*36*	*1*

b. Un histograma y un polígono de frecuencias para las longitudes de las hojas caídas de un árbol. La información es la siguiente

Longitud x_i	*Marca de clase* m_i	*Frecuencia absoluta* f_i	*Frecuencia relativa* fr_i
[0, 4)	*2*	*7*	*23.33%*
[4, 8)	*6*	*10*	*33.33%*
[8, 12)	*10*	*6*	*20.00%*
[12,16)	*14*	*4*	*13.34%*
[16, 20)	*18*	*3*	*10.00%*
Total		*30*	*100.00%*

c. Una gráfica circular para los siguientes datos acerca la opinión de una muestra de ciudadanos sobre la gestión de un presidente municipal.

Opinión	*Frecuencia absoluta*	*Frecuencia relativa*
Excelente	*12*	*0.12*
Bueno	*10*	*0.10*
Regular	*23*	*0.23*
Malo	*30*	*0.30*
Pésimo	*25*	*0.25*
Total	*100*	*1.00*

2. *Analiza a cuál de las siguientes tablas corresponde cada una de las gráficas circulares de abajo. Escribe en cada gráfica el título que consideres adecuado y las categorías correspondientes en la leyenda.*

Deporte que más practican alumnos de un grupo del CCH

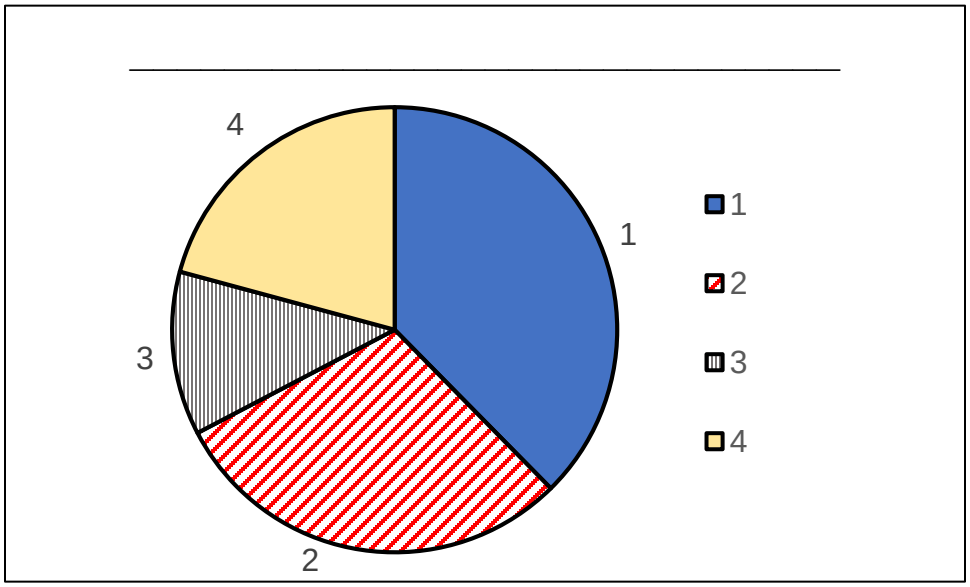
Deporte	**% alumnos**
Futbol	38%
Básquetbol	30%
Béisbol	12%
Otro	21%
Total	100%

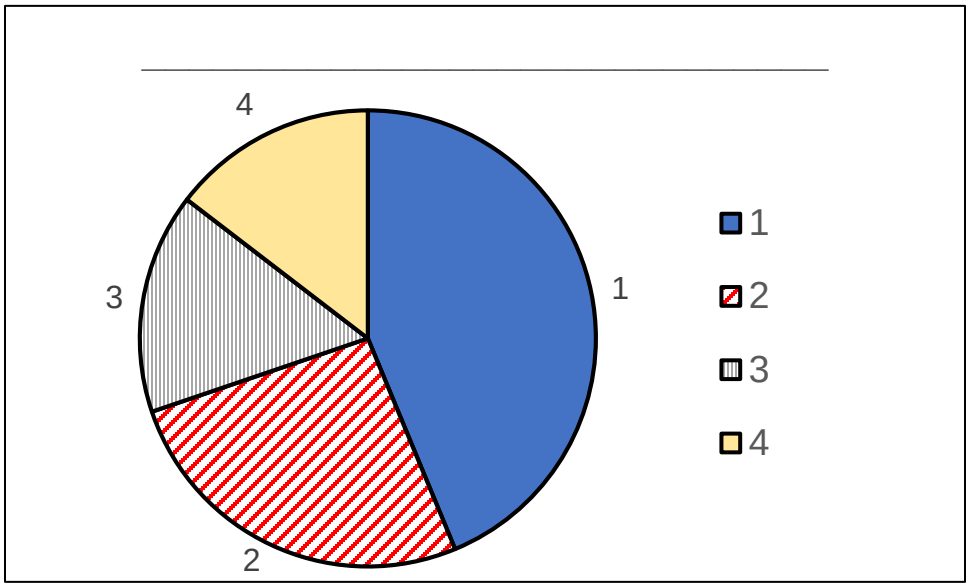
Tipo de lectura que más les gusta a alumnos de un grupo del CCH

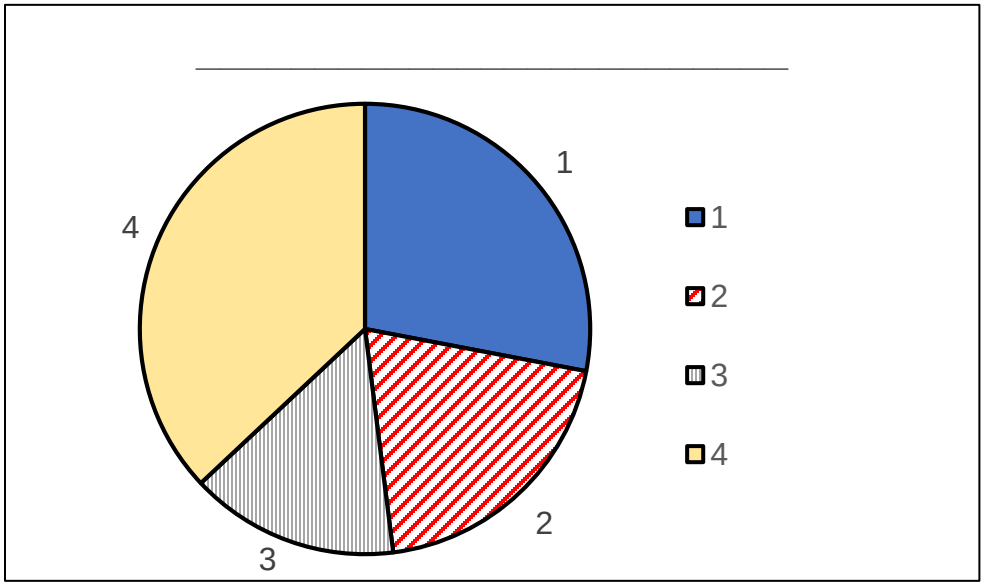
Tipo	**% alumnos**
Novela	42%
Cuento	25%
Poesía	15%
Otro	14%
Total	100%

Colonias donde viven alumnos de un grupo del CCH

Colonia	**% alumnos**
San Lorenzo	28%
Santa Úrsula	20%
Huipulco	15%
Otras	37%
Total	100%




4
1
3
2
1
2
3
4


1
4
2
3
1
2
3
4

1.5 MEDIDAS ESTADÍSTICAS

Además de ordenar los datos en tablas de frecuencias y de representarlos a través de gráficas, la exploración de los datos se enriquece calculando medidas estadísticas que nos permitan contestar preguntas como:

- Si ordenamos los datos de menor a mayor, ¿entre qué cantidades se encuentra la mitad menor de los datos? ¿Y la mitad mayor?
- ¿Hay algún valor que se repite más en los datos?
- ¿Hay algún número en torno al cual se agrupe una buena cantidad de los datos?
- En promedio, ¿qué tan lejos están los datos de ese número?
- ¿Qué tan lejos está el dato mayor del menor?

Estudiaremos aquí tres clases de medidas: las medidas de tendencia central, las medidas de dispersión y las medidas de posición.

A) Medidas de tendencia central

Buscamos cantidades que sean representativas de la colección de datos en el sentido de que todos los datos o la mayoría de ellos, se encuentren más o menos cerca de esas cantidades. A veces esto no es fácil de lograr porque hay algunos datos con valores notablemente diferentes a los demás, sea porque son mucho más grandes o porque son mucho más chicos. A los valores de estos datos notablemente diferentes, les llamaremos *valores extremos.*

Se les llaman medidas de tendencia central porque cuando no hay valores extremos, las cantidades representativas suelen ubicarse por el centro de los datos.

Estudiaremos tres medidas de tendencia central: la media aritmética o promedio, la mediana y la moda

a) Media aritmética o promedio

La media aritmética solo puede calcularse cuando la variable es cuantitativa y se obtiene sumando todos los datos y dividiendo el resultado entre la cantidad de datos. Cuando los datos son elementos de una muestra, la media se representa por $\overline{X}$, cuando los datos corresponden a toda la población se puede denotar por μ .

Ejemplo 1.

Las alturas en metros de 8 estudiantes elegidos al azar en un grupo son: 1.68, 1.65, 1.66, 1.70, 1.72, 1.73, 1.65 y 1.64. Entonces, la estatura media es

$$\overline{X} = \frac{1.68 + 1.65 + 1.66 + 1.70 + 1.72 + 1.73 + 1.65 + 1.64}{8} = 1.679$$

Ejemplo 2.

En la siguiente tabla se muestran las calificaciones en Estadística y Probabilidad I de una muestra de estudiantes del CCH-Sur que cursaron la asignatura en el semestre 2019-1.

Calificación	*Estudiantes*
5	*24*
6	*18*
7	*20*
8	*28*
9	*14*
10	*11*
Total	*115*

Para calcular la media de las calificaciones, es necesario recordar que lo que esta tabla indica es que, de los 115 estudiantes de la muestra, 24 obtuvieron la calificación 5, 18 estudiantes obtuvieron la calificación 6, 20 estudiantes obtuvieron 7 y así sucesivamente.

Por lo tanto,

$$\overline{X} = \frac{24(5) + 18(6) + 20(7) + 28(8) + 14(9) + 11(10)}{115}$$

La calificación media es $\overline{X} = 7.2$

Cuando los datos se presentan agrupados en intervalos, se debe elegir un representante de cada intervalo o clase para calcular la media. Este representante es la marca de clase o punto medio del intervalo. Si el intervalo es [a, b) su marca de clase es

$$m = \frac{a + b}{2}$$

Ejemplo 3.

Se le pide a una muestra de 100 pacientes que califiquen la calidad de la atención que recibieron en un hospital público. La escala de calificaciones es de 1 a 100, y los resultados se presentan en la siguiente tabla.

Puntaje	*Marca de clase*	*Pacientes*
$[20, 35)$	27.5	18
$[35, 50)$	42.5	27
$[50, 65)$	57.5	25
$[65, 80)$	72.5	14
$[80, 95)$	87.5	16
Total		100

El puntaje promedio que obtuvo el hospital, se aproxima de la siguiente manera

$$\overline{X} \approx \frac{27.5(18) + 42.5(27) + 57.5(25) + 72.5(14) + 87.5(16)}{100} = 54.95$$

Notación de sumatoria

El cálculo de la media aritmética, y el de otras medidas que veremos más adelante, requiere sumas. Para dar una fórmula general conviene conocer una forma de escribir brevemente una suma usando la letra griega *sigma* mayúscula (Σ), que corresponde a la *S* de nuestro abecedario.

Por ejemplo:

$$\sum_{i=1}^{12} i = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 11 + 12$$

Es decir, debajo de la sigma se pone el primer valor entero que toma la literal y arriba el último valor entero que toma. Para desarrollar la sumatoria, en cada sumando se sustituye la literal por uno de los valores enteros en el rango descrito.

Otros ejemplos:

$$\sum_{j=1}^{8} 2j = 2(1) + 2(2) + 2(3) + 2(4) + 2(5) + 2(6) + 2(7) + 2(8)$$
$$= 2 + 4 + 6 + 8 + 10 + 12 + 14 + 16$$

$$\sum_{k=1}^{6} k^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 = 1 + 4 + 9 + 16 + 25 + 36$$

Esta notación también se puede usar en sumas de valores generales representados por literales, por ejemplo

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

También se usa para un número general de sumandos (n), por ejemplo

$$\sum_{k=1}^{n} y_k = y_1 + y_2 + y_3 + \cdots + y_n$$

Fórmula general de la media

- Si se trata de *n* datos $x_1, x_2, x_3, \ldots, x_n$ no incluidos en tablas de frecuencias, la media aritmética es

$$\overline{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$

- Si los *n* datos se presentan en una tabla por valores (de frecuencias simples), donde los valores son $y_1, y_2, \ldots, y_k$ y tienen frecuencias absolutas $f_1, f_2, \ldots, f_k$ respectivamente, entonces la media es

$$\overline{X} = \frac{\sum_{j=1}^{k} y_j f_j}{n}$$

- Si los *n* datos se presentan en una tabla de datos agrupados en intervalos, donde las marcas de clase son $m_1, m_2, \ldots, m_k$ y los k intervalos tienen frecuencias absolutas $f_1, f_2, \ldots, f_k$ respectivamente, entonces la media se aproxima por

$$\overline{X} \approx \frac{\sum_{j=1}^{k} m_j f_j}{n}$$

Ventajas y limitaciones

La principal ventaja de la media es que es fácil de calcular y la fórmula para obtenerla se aplica de manera sencilla en cualquier presentación de los datos.

La limitación más notable de la media es que es muy sensible a valores extremos. Por ejemplo:

- Datos: 125, 128, 120, 122 y 129, la media es $\overline{X} = 124.8$ que resulta representativa de los 5 datos.

- Datos: 125, 128, 120, 122 y 582, la media es $\overline{X} = 215.4$ que no es representativa porque no se parece a los primeros 4 datos ni al último.

Cuando no hay datos con valores extremos, la media es útil para lo siguiente:

- Obtener estimaciones de medidas, por ejemplo, se toman 5 medidas del peso de un objeto ligero y se obtienen los datos: 3.1, 3.3, 2.9, 2.8, 3.2 gramos. El peso real del objeto se puede estimar por el promedio de las medidas anteriores, es decir, 3.06 g.

- Realizar repartos igualitarios, por ejemplo, cinco amigos acuerdan pagar el monto de la cuenta en un restaurante a partes iguales. Si los consumos fueron $130, $99, $132, $117 y $122, lo que cada uno de ellos debe pagar es el promedio, es decir $120.

- Representar una colección de datos para facilitar comparaciones, por ejemplo, para decidir si resulta conveniente utilizar una nueva técnica de entrenamiento en salto largo (sin impulso), se comparan las distancias alcanzadas por 5 estudiantes de bachillerato antes y después de aplicar la nueva técnica (cm).

	Ana	Ema	Laura	Rosa	Lila
Antes	206	207	212	214	205
Después	208	205	215	214	208

La longitud media antes de aplicar la nueva técnica es 208.8 cm y la longitud media después es 210 cm, por lo que la nueva técnica es útil.

b) Mediana

La mediana es un valor que se localiza en el centro de los datos ordenados (de menor a mayor o de mayor a menor). Puede determinarse cuando los datos corresponden a una variable cuantitativa y cuando se trata de una variable cualitativa ordinal.

Se Identifica con las letras *Mdn*. Si los datos corresponden a una muestra, la mediana se suele denotar por $\tilde{X}$.

Si la cantidad de datos es impar, la mediana es el dato que se encuentra en el centro, una vez que los datos han sido ordenados. Si la cantidad es par, se promedian los dos datos centrales, y la mediana no necesariamente es uno de los datos.

Ejemplo 4.

Datos: 162, 150, 149, 160, 156, 159, 147, 165, 162, 150, 164

Cantidad de datos: n = 11

Datos ordenados: 147, 149, 150, 150, 156, 159, 160, 162, 162, 164, 165

Mediana: Mdn = 159.

Ejemplo 5.

Datos: 0.54, 0.60, 0.57, 0.50, 0.61, 0.45, 0.50, 0.72, 0.61, 0.55

Cantidad de datos: n = 10

Datos ordenados:

0.45, 0.50, 0.50, 0.54, 0.55, 0.57, 0.60, 0.61, 0.61, 0.72

Mediana: $Mdn = \frac{0.55+0.57}{2} = 0.56$

Par datos de una variable cualitativa ordinal, puede solo indicarse entre qué datos se encuentra la mediana cuando la cantidad de observaciones es par.

Ejemplo 6.

Opinión sobre el desempeño de un presidente municipal. E significa excelente, B representa bueno, R indica regular, M simboliza malo y P es pésimo,

Datos ordenados:

E E B B R R R R M M M M M M P P

Así que la mediana de los datos está entre regular y malo.

Mediana para datos en tablas de frecuencias

Si la tabla es de datos por valores (no agrupados en intervalos), la mediana es el valor donde la frecuencia acumulada rebasa por primera vez la mitad de la cantidad de datos, o bien donde la frecuencia relativa acumulada rebasa por primera vez el 0.5 o 50% de los datos.

Ejemplo 7.

Un centro para personas de la tercera edad, realiza una encuesta a una muestra de

adultos mayores sobre el número de vasos de agua que toman al día. La información recabada se recoge en la siguiente tabla.

Vasos de agua al día	*Personas (f_i)*	*Porcentaje (fr_i)*	*Frecuencias acumuladas*	
			Absoluta (Fa_i)	*Relativa (Fra_i)*
2	*4*	*8%*	*4*	*8%*
3	*7*	*14%*	*11*	*22%*
4	*8*	*16%*	*19*	*38%*
5	*8*	*16%*	*27*	*54%*
6	*6*	*12%*	*33*	*66%*
7	*7*	*14%*	*40*	*80%*
8	*6*	*12%*	*46*	*92%*
9	*3*	*6%*	*49*	*98%*
10	*1*	*2%*	*50*	*100%*
Total	*50*	*100%*		

La mitad de la cantidad de datos es 50/2 = 25, y la frecuencia absoluta acumulada rebasa ese número en 5 vasos. Observa también que las frecuencias relativas acumuladas rebasan el 50% en ese mismo valor. Por tanto, la mediana es $\tilde{X} = 5$ vasos de agua al día.

Si la tabla es de datos agrupados en intervalos, se localiza el intervalo en el que se alcanza la mediana analizando de la misma forma que antes las frecuencias acumuladas. Hay dos procedimientos que se suelen usar para aproximar el valor que toma la mediana dentro del intervalo:

i) Tomar la marca de clase de ese intervalo como mediana.

ii) Supongamos que el intervalo donde cae la mediana es el número *j.* Se puede usar la fórmula

$$Mdn \approx L_{inf} + \left(\frac{\frac{n}{2} - Fa_{j-1}}{f_j} \right) c$$

Los elementos de la fórmula anterior son:

L_{inf} = límite inferior de la clase mediana.

$Fa_{j\text{-}1}$ = frecuencia absoluta acumulada de la clase que precede a la clase mediana.

f_j = frecuencia absoluta de la clase mediana.

c = amplitud de la clase mediana.

n = tamaño de muestra.

Ejemplo 8.

Se coloca un sensor de velocidades en una autopista para registrar las velocidades a las que pasan los vehículos por un punto. En una muestra se encontró que las velocidades registradas por el sensor fueron:

Velocidades (km/h)	*Marca de clase*	*Vehículos (f_i)*	*Porcentaje (fr_i)*	*Frecuencias acumuladas*	
				Absoluta (Fa_i)	*Relativa (Fra_i)*
[90, 95)	*92.5*	*8*	*12.31%*	*8*	*12.31%*
[95, 100)	*97.5*	*12*	*18.46%*	*20*	*30.77%*
[100, 105)	*102.5*	*14*	*21.54%*	*34*	*52.31%*
[105, 110)	*107.5*	*10*	*15.38%*	*44*	*67.69%*
[110, 115)	*112.5*	*12*	*18.46%*	*56*	*86.15%*
[115, 120)	*117.5*	*5*	*7.69%*	*61*	*93.85%*
[120, 125)	*122.5*	*4*	*6.15%*	*65*	*100.00%*
Total		*65*	*100%*		

El intervalo donde se encuentra la mediana es [100, 105), ya que su frecuencia relativa acumulada rebasa por primera vez el 50%. Así que, aplicando las estrategias antes mencionadas se obtiene:

i) $Mdn \approx 102.5$ *km/h, o bien*

ii) $Mdn \approx 100 + \left(\frac{\frac{65}{2} - 20}{14}\right)5 = 104.46$ *km/h*

Ventajas y limitaciones

La principal ventaja de la mediana es que no se ve alterada por valores extremos, lo que la hace útil como medida representativa cuando hay este tipo de valores. Por ejemplo:

- Datos: 18, 22, 35, 38 y 42, la mediana es $\tilde{X} = 35$.
- Datos: 18, 22, 35, 38 y 99, la mediana sigue siendo es $\tilde{X} = 35$.

Su limitación más fuerte es que no hay una fórmula para calcularla fácilmente en cualquier presentación de los datos.

c) Moda

La moda es el valor que más se repite en el conjunto de datos, es decir, es el dato con mayor frecuencia. Se puede identificar con las letras *Mo* y se suele usar el símbolo $\hat{X}$ para la moda de una muestra.

Esta medida es aplicable en cualquier tipo de variable y es la única medida de tendencia central posible cuando la variable es cualitativa nominal. Puede haber una moda, varias modas y también puede no haber ninguna.

Ejemplo 9.

- *Datos: 1.2, 1.5, 1.4, 1.6, 1.1, 1.5, 1.7, 1.0, 1.6, 1.5, 1.6*

  Hay 2 modas: 1.5 y 1.6.

- *Datos: 71, 75, 68, 66, 82, 67, 74, 65, 70, 76*

  No hay moda

- *Datos: 345, 352, 347, 356, 345, 352, 360, 350, 356, 345, 361*

  Mo = 345

En una tabla de frecuencias por valores, es fácil determinar si hay modas y cuáles son sus valores. Basta observar qué valores tienen mayor frecuencia (absoluta o relativa). En la tabla del ejemplo 7, es claro que las modas son 4 y 5 vasos de agua al día.

Si la tabla es de datos agrupados en intervalos, se localiza el intervalo de mayor frecuencia y de nuevo hay dos procedimientos válidos para aproximar el valor de la moda.

i) Tomar la marca de clase de ese intervalo como moda.

ii) Usar la fórmula

$$Mo = L_{inf} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) c$$

Los elementos de esta fórmula son:

L_{inf} = límite inferior de la clase modal.

Δ_1 = diferencia positiva entre la frecuencia de la clase modal y la clase que le precede.

Δ_2 = diferencia positiva entre la frecuencia de la clase modal y la clase que le sigue.

c = amplitud de la clase modal.

Ejemplo 10.

La siguiente tabla contiene la información acerca de la edad de los trabajadores de una empresa.

Edad (años)	*Marca de clase*	*Empleados (f_i)*	*Porcentaje (fr_i)*
De 16 a 25	*20.5*	*23*	*20.54%*
De 26 a 35	*30.5*	*34*	*30.36%*
De 36 a 45	*40.5*	*25*	*22.32%*
De 46 a 55	*50.6*	*18*	*16.07%*
De 56 a 65	*60.5*	*12*	*10.71%*
Total		*112*	*100%*

La frecuencia absoluta mayor es 34 y ocurre en el intervalo de 26 a 35 años. De acuerdo a lo señalado anteriormente, la moda puede aproximarse por:

i) *Mo = 30.5 años, o bien*

ii) $Mo = 26 + \left(\frac{11}{11+9}\right)9 = 30.95$ *años*

EJERCICIOS 1.5.1

1. En los siguientes casos, indica cuál de las medidas de tendencia central te parece más representativa de la colección.

   a. Seis estudiantes que midieron el diámetro de una tapa rosca de plástico, obtuvieron las siguientes medidas (mm): 25, 30, 28, 26, 48 y 27.

      Medida representativa: _____

   b. Las edades de 6 alumnos de quinto semestre son: 17, 18, 18, 17, 18, 18.

      Medida representativa: _____

   c. Un estudiante del CCH registró el tiempo que ocupó en trasladarse de su casa a la escuela los 5 días de la semana (en min): 26, 32, 30, 25, 30

Medida representativa: _____

2. Determina las tres medidas de tendencia central en cada una de las siguientes colecciones de datos:

   a. Longitud de 12 cuadras elegidas en calles de la Ciudad de México, dadas en kilómetros: 0.12, 0.17, 0.1, 0.22, 0.05, 0.11, 0.09, 0.12, 0.08, 0.15, 0.08.

      Media: _____ Mediana: _____ Moda(s): _____

   b. Se registró el número de hijos que tienen 120 mujeres de 15 años o más, que asisten a una clínica del IMSS en la CdMx. La información recabada es:

			Frecuencias acumuladas	
Hijos	Mujeres	Porcentaje	Absoluta	Relativa
0	31	25.83%	31	25.83%
1	24	20.00%	55	45.83%
2	28	23.33%	83	69.17%
3	15	12.50%	98	81.67%
4	10	8.33%	108	90.00%
5	8	6.67%	116	96.67%
6	4	3.33%	120	100.00%
Total	120	100.00%		

      Media: _____ Mediana: _____ Moda(s): _____

3. La distribución de edades de los empleados de una gran empresa, tiene las siguientes medidas (años cumplidos).

Edad mínima	16
Edad media	36.4
Edad mediana	30
Moda	28
Edad máxima	65

   a. ¿Entre qué edades se encuentra el 50% más joven de los empleados? _____

   b. ¿Entre qué edades se encuentra el 50% mayor de los empleados? _____

c. ¿Cuál es la edad que más se repite entre los empleados? _____

d. ¿Cuál es el promedio de edades? _____

4. En cada caso, construye una colección de 10 datos que tenga las características que se indican. Ningún valor puede repetirse más de 3 veces en las colecciones.

   a. Que tenga media 52

   Colección: _____

   b. Que tenga mediana 25

   Colección: _____

   c. Que tenga moda 157

   Colección: _____

   d. Que tanto su media como su moda, sean iguales a 20.

   Colección: _____

   e. Que tanto su media, como su mediana, como su moda, sean iguales a 200.

   Colección: _____

B) Medidas de dispersión

Estas medidas estadísticas indican qué tan cercanos o alejados están los valores que toma la variable de estudio.

a) Rango

El rango de una colección de datos es el resultado de restar el valor máximo menos el mínimo, y es un indicador de qué tan alejados están los datos entre sí.

Ejemplo 11.

Analicemos la dispersión de las siguientes colecciones de datos a través del rango

- *Datos: 7.51, 2.35, 1.27, 4.16, 6.51, 2.95, 5.17, 7.04, 3.36, 5.77, 6.33*

  Rango = 7.51 – 1.27 = 6.24.

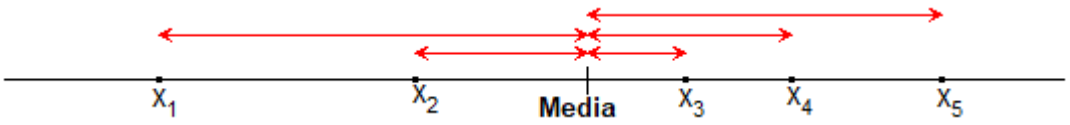
- *Datos: 764, 765, 768, 766, 765, 767, 764, 765, 766, 763*

  Rango = 768 – 763 = 5

Por lo que hay mayor dispersión en la primera colección.

b) Varianza y Desviación estándar

Ahora buscamos una medida que indique qué tan lejos están los valores que toma una variable de su media aritmética o promedio. En el siguiente dibujo, esas distancias entre los valores y la media están representadas por las flechas sobre la recta numérica.



Si promediamos las diferencias entre cada x_i y a la media $\overline{X}$ usando la expresión

$$\frac{(x_1-\overline{X})+(x_2-\overline{X})+\cdots+(x_n-\overline{X})}{n},$$

las diferencias positivas se compensarían con las negativas y no obtendríamos una medida de la cercanía o la lejanía respecto a la media.

Para evitar lo anterior, se puede promediar el valor absoluto de esas diferencias o bien el cuadrado de las mismas, para que todas sean positivas. En estudios más avanzados, el uso del valor absoluto complica cálculos y dificulta estrategias sencillas. Por ello, optaremos por elevar al cuadrado las diferencias anteriores para obtener la medida llamada *Varianza.*

$$\frac{(x_1-\overline{X})^2+(x_2-\overline{X})^2+\cdots+(x_n-\overline{X})^2}{n}$$

Sin embargo, esto da lugar a un nuevo problema. Los valores que toma una variable estadística tienen unidades. Por ejemplo, si la variable indica el diámetro de taparroscas de plástico, sus valores están en mm. Pero al hacer los cálculos anteriores, obtendríamos una medida dada en mm^2, lo que impide una adecuada comparación de valores pues una es medida de longitud y la otra de área.

Para volver a las unidades originales de los valores de la variable estadística, aplicamos raíz cuadrada y obtenemos una buena medida de la dispersión respecto a la media, llamada Desviación Estándar (o Desviación Típica).

$$\sqrt{\frac{\left(x_1-\overline{X}\right)^2+\left(x_2-\overline{X}\right)^2+\cdots+\left(x_n-\overline{X}\right)^2}{n}}$$

Cuando se trabaja con una muestra, se ha visto que dividir la expresión anterior entre n – 1 en lugar de hacerlo entre n, ofrece una cantidad que estima de mejor manera el valor correspondiente de la población.

En resumen:

- Para una muestra de tamaño n cuyos elementos son x_1, x_2, …, x_n, tenemos:

Varianza:

$$S^2=\frac{\left(x_1-\overline{X}\right)^2+\left(x_2-\overline{X}\right)^2+\cdots+\left(x_n-\overline{X}\right)^2}{n-1}=\frac{\sum_1^n\left(x_j-\overline{X}\right)^2}{n-1}$$

Desviación Estándar:

$$S=\sqrt{\frac{\sum_1^n\left(x_j-\overline{X}\right)^2}{n-1}}$$

- Para una muestra dada en tablas de frecuencias, cuyos valores o marcas de clase son $y_1, y_2, \ldots, y_k$ y tienen frecuencias absolutas $f_1, f_2, \ldots, f_k$ respectivamente, tenemos.

Varianza:

$$S^2=\frac{\left(y_1-\overline{X}\right)^2f_1+\left(y_2-\overline{X}\right)^2f_2+\cdots+\left(x_k-\overline{X}\right)^2f_k}{n-1}=\frac{\sum_1^k\left(x_j-\overline{X}\right)^2f_j}{n-1}$$

Desviación Estándar:

$$S=\sqrt{\frac{\sum_1^k\left(y_j-\overline{X}\right)^2f_j}{n-1}}$$

Ejemplo 12.

En un grupo de Estadística I del CCH Sur, se observó la estatura de 16 alumnos y se obtuvieron los siguientes datos (ya ordenados):

1.52 1.52 1.53 1.53 1.57 1.58 1.58 1.60 1.64 1.64 1.64 1.66 1.66 1.74 1.76 1.79

Para realizar los cálculos de la varianza, resulta conveniente construir una tabla como la siguiente

Estatura x_i	Frecuencia f_i	$x_i - \overline{X}$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^2 f_i$
1.52	2	-0.1025	0.01051	0.02101
1.53	2	-0.0925	0.00856	0.01711
1.57	1	-0.0525	0.00276	0.00276
1.58	2	-0.0425	0.00181	0.00361
1.6	1	-0.0225	0.00051	0.00051
1.64	3	0.0175	0.00031	0.00092
1.66	2	0.0375	0.00141	0.00281
1.74	1	0.1175	0.01381	0.01381
1.76	1	0.1375	0.01891	0.01891
1.79	1	0.1675	0.02806	0.02806
Total	16			$\Sigma = 0.1095$

$\overline{X} = 1.6225$

Así se obtiene:

Varianza: $S^2 = \frac{0.1095}{15} = 0.0073$

Desviación Estándar: $S = \sqrt{0.0073} = 0.08544$

Interpretación: Las alturas de los 16 estudiantes de la muestra, se encuentran en promedio a una distancia de 0.085 m = 8.5 cm de su media 1.6225 m

C) Medida de dispersión relativa

En un zoológico, el peso promedio de los elefantes es de 4800 kg y su desviación estándar es de 2200 kg. En cambio, el peso promedio de los conejos es de 4.3 kg y su desviación estándar es de 1.8 kg. ¿Cuáles pesos están más dispersos, los de los elefantes o los de los conejos?

Evidentemente, en términos absolutos la dispersión en el peso de los elefantes es mucho más grande. Sin embargo, para darnos una idea más precisa debemos tomar en cuenta el peso promedio de cada grupo de animales.

Una buena forma de comparar estas colecciones tan diferentes, es analizar qué parte de la media es la desviación estándar en cada caso. A esto se le llama *Coeficiente de Variación (CV)* y se expresa en porcentaje.

$$CV = \left(\frac{S}{\overline{X}}\right) 100\%$$

Ejemplo 13.

Vamos a calcular el coeficiente de variación correspondiente a cada una de las colecciones de animales que hemos descrito en párrafos anteriores.

Elefantes: $CV = \left(\frac{2200}{4800}\right) 100\% = 45.83\%$

Conejos: $CV = \left(\frac{1.8}{4.3}\right) 100\% = 41.86\%$

Por tanto, está un poco más dispersa la colección de pesos de los elefantes que la de los conejos.

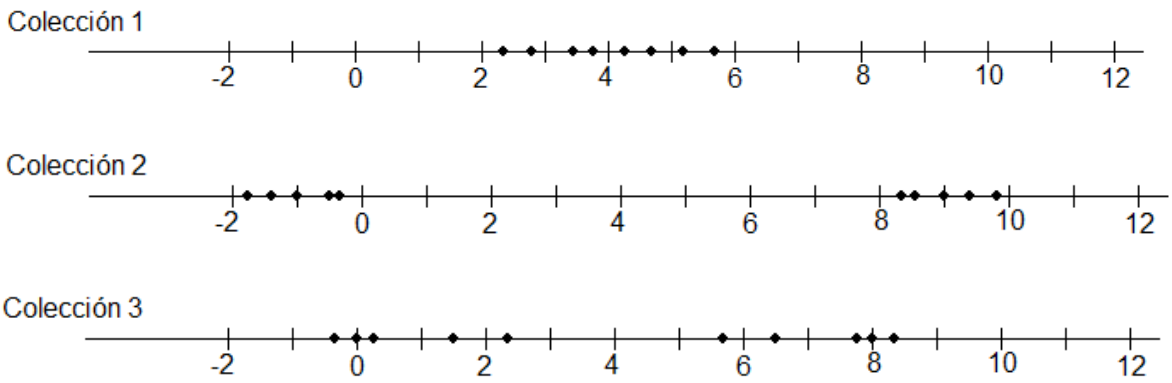
EJERCICIOS 1.5.2

1. Calcula el rango y la desviación estándar de las siguientes colecciones de datos y compara los resultados.

   Colección A: 14, 16, 10, 11, 18, 9 Colección B: 18, 18, 17, 9, 9, 10

   ¿Cuál de las dos colecciones está más dispersa? _____ Explica en qué basaste tu respuesta _____

2. En los siguientes casos, identifica en qué intervalo crees que se encuentra la desviación estándar de las colecciones de datos representadas por los puntos en la recta numérica. Todas las colecciones tienen media igual a 4.



   Desviación estándar entre 2 y 4 unidades: colección _____

   Desviación estándar menor a 2 unidades: colección _____

   Desviación estándar mayor a 4 unidades: Colección _____

3. En una empresa A, el salario medio de los trabajadores es de $4 500 mensuales con una desviación estándar de $2 000. En una empresa B, el salario medio es de $40 000 mensuales con una desviación estándar también de $2 000.

    a. ¿En cuál de las dos empresas los salarios son más homogéneos, es decir, más parecidos entre sí? _____ Explica tu respuesta. _____

    b. Calcula el coeficiente de variación en ambos casos e interpreta los resultados que se obtienen.

    CV_A = _____ CV_B = _____

    Interpretación _____

D) Medidas de posición

Las medidas de posición son cantidades que dividen una colección ordenada de datos, en 4, 10 o 100 partes iguales.

a) Cuartiles

Los cuartiles son 3 números que dividen a la colección ordenada de datos en 4 partes iguales. Sus valores pueden ser datos, o bien cantidades obtenidas al promediar dos datos consecutivos de la colección ordenada.

Cuartil 1 (Q_1): es un número tal que el 25% de los datos son menores o iguales a él y el 75% de los datos son mayores o iguales a él. Su posición en la colección ordenada se obtiene dividiendo el total de datos (n) entre 4. Si $\frac{n}{4}$ es un entero, se promedian los datos ubicados en las posiciones $\frac{n}{4}$ y $\frac{n}{4}+1$. Si $\frac{n}{4}$ es un número fraccionario, se toma como cuartil 1 el dato que está en la siguiente posición.

Cuartil 2 (Q_2): coincide con la mediana porque es una cantidad tal que el 50% de los datos son menores o iguales a ella y el otro 50% son mayores o iguales.

Cuartil 3 (Q_3): es un número tal que el 75% de los datos son menores o iguales a él y el 25% son mayores o iguales. Su posición en la colección ordenada se obtiene multiplicando el total de datos (n) por $\frac{3}{4}$. Si $\frac{3}{4}n$ es un entero, se promedian los datos ubicados en las posiciones $\frac{3}{4}n$ y $\frac{3}{4}n+1$. Si $\frac{n}{4}$ es un número fraccionario, se toma como cuartil 1 el dato que está en la siguiente posición

Ejemplo 14.

Se tomó el tiempo (en minutos) que tardaron 16 niños en resolver un test psicológico. Estos son los datos:

28, 20, 21, 20, 23, 25, 26, 19, 29, 22, 18, 30, 26, 29, 24, 25

Datos ordenados:

18, 19, 20, 20, 21, 22, 23, 24, 25, 25, 26, 26, 28, 29, 29, 30

Cuartil	*Posición*	*Valor*
Q_1	$\frac{1}{4}(16) = 4$, *promedio 4° y 5°*	*20.5*
Q_2 = *Mdn*	$\frac{2}{4}(16) = 8$, *promedio 8° y 9°*	*24.5*
Q_3	$\frac{3}{4}(16) = 12$, *promedio 12° y 13°*	*27*

Así que los cuartiles son: $Q_1 = 20.5$, $Q_2 = 24.5$ *y* $Q_3 = 27$.

Ejemplo 15.

Las notas obtenidas por 21 estudiantes del CCH en un examen de Historia, son:

8.5, 10, 7, 7.5, 6, 5, 6.5, 8.5, 9, 7.5, 8, 10, 9.5, 9, 4, 5.5, 6, 7.5, 7, 7.5, 8

Datos ordenados:

4, 5, 5.5, 6, 6, 6.5, 7, 7, 7.5, 7.5, 7.5, 7.5, 8, 8, 8.5, 8.5, 9, 9, 9.5, 10, 10

Cuartil	*Posición*	*Valor*
Q_1	$\frac{1}{4}(21) = 5.25$, *6° lugar*	*6.5*
Q_2 = *Mdn*	$\frac{2}{4}(21) = 10.5$, *11° lugar*	*7.5*
Q_3	$\frac{3}{4}(21) = 15.75$, *16° lugar*	*8.5*

Así que los cuartiles son: $Q_1 = 6.5$, $Q_2 = 7.5$ *y* $Q_3 = 8.5$

Los cuartiles permiten crear un esquema, conocido como esquema de caja con brazos, que proporciona una buena imagen gráfica de la concentración de los datos en cada una de las 4 partes. También muestran qué tan lejos están los valores extremos de la mayoría de los datos.

Este esquema se construye usando cinco valores: el valor mínimo, los tres cuartiles y el valor máximo. Sobre una recta numérica horizontal se marca una escala convencional y en ella se localizan los 5 números.

El primer brazo es un segmento horizontal que va del mínimo al primer cuartil. Se agrega una caja rectangular que va del primer cuartil al tercero. El segundo brazo es otro segmento horizontal que va del tercer cuartil al valor máximo. La mediana se marca como un segmento vertical dentro de la caja.

Ejemplo 16.

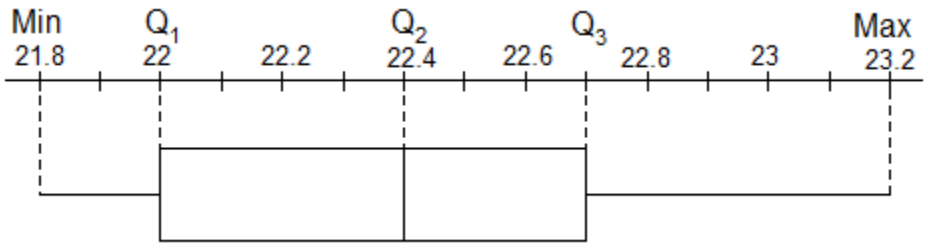
Un atleta entrena para la carrera de 200 metros planos y toma el tiempo que le lleva correrlos durante 15 días consecutivos. Los tiempos que hace, en segundos y décimas de segundo, son los siguientes, ordenados de menor a mayor

21.8 21.8 21.9 22 22 22.2 22.4 22.4 22.5 22.6 22.6 22.7 22.9 23 23.2

Los 5 números que se requieren para el esquema son:

$$Mín = 21.8, \quad Q_1 = 22, \quad Q_2 = 22.4, \quad Q_3 = 22.7, \quad Max = 23.2$$

La siguiente imagen muestra el esquema de caja con brazos correspondiente.



A partir de la gráfica se pueden hacer las siguientes observaciones:

- Existe mayor variación entre Q_3 y el máximo, es decir, esa es la cuarta parte más dispersa de los tiempos que hizo el atleta.
- Existe menos variación entre el dato mínimo y Q_1, es decir, la primera cuarta parte de los tiempos es la más concentrada.
- La mitad central de los datos va de 22 a 22.7 segundos, es decir, en este rango se localiza la mitad de los tiempos que se obtiene eliminando la cuarta parte mayor y la cuarta parte menor.
- La mitad más pequeña de los tiempos va de 21.8 a 22.4 segundos y la mitad mayor va de 22.4 a 23.2 segundos

Es importante comenzar el diagrama de caja con brazos trazando una recta numérica con una escala adecuada a los datos.

b) **Deciles**

Se trata de 9 números que dividen a la colección ordenada de datos en 10 partes iguales. Se denotan por D_1, D_2, D_3, …, D_9. Estas medidas se usan cuando la cantidad de datos es grande. Sus valores se calculan de manera similar a la forma en que se calcularon los cuartiles.

El 10% de los datos son menores o iguales que D_1 y el 90% mayores o iguales.

El 20% de los datos son menores o iguales que D_2 y el 80% mayores o iguales.

El 30% de los datos son menores o iguales que D_3 y el 70% mayores o iguales.

Y así sucesivamente.

c) Percentiles

Son 99 números que dividen a la colección ordenada de datos en 100 partes iguales. Se denotan por P_1, P_2, P_3, … P_{99}. Estos permiten determinar rangos de variación de cualquier porcentaje de los datos. Por ejemplo, el 22% de los datos son menores o iguales a P_{22} y el 78% son mayores o iguales. El 65% de los datos son menores o iguales que P_{65} y el 35% es mayor o igual.

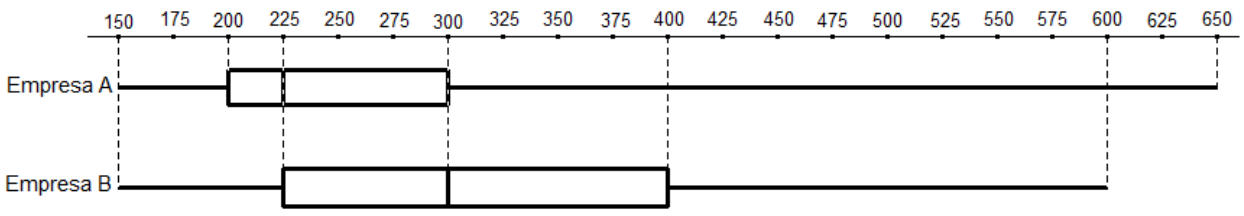
EJERCICIOS 1.5.3

1. Los pesos (en kg) de un grupo de 26 hombres adultos son:

   85, 72.5, 68.5, 70, 88, 72, 65, 66.5, 68, 86, 71, 70.5, 75, 71.5, 70, 72, 66, 65.2, 68, 68.5, 71, 69, 76.5, 80.5, 73.5, 75

   Determina los valores de los tres cuartiles _____

2. Los siguientes diagramas muestran información sobre los salarios diarios de los empleados de 2 empresas (en pesos).



   Contesta:

   a. ¿En cuál de las dos empresas los salarios están más concentrados hacia montos bajos? _____ Explica tu respuesta _____
   _____

b. ¿Entre qué montos está la mitad más baja de los salarios en cada empresa?

Empresa A: _____

Empresa B: _____

c. ¿Entre que montos está la cuarta parte más alta de salarios en cada empresa?

Empresa A: _____

Empresa B: _____

d. ¿Cuál es la cuarta parte más concentrada en cada empresa?

Empresa A: _____

Empresa B: _____

1.6 REGLA EMPÍRICA

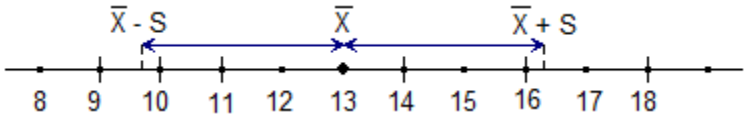
De acuerdo a lo que se ha estudiado acerca de las medidas estadísticas, si no hay datos extremos o atípicos en una colección, sabemos que una buena parte de los datos debe estar entre la media menos la desviación estándar y la media más la desviación estándar, es decir, entre $\overline{X} - S$ y $\overline{X} + S$. La pregunta que ahora queremos responder es ¿qué porcentaje de los datos caen en este rango?

La respuesta no es siempre la misma.

Ejemplo 1.

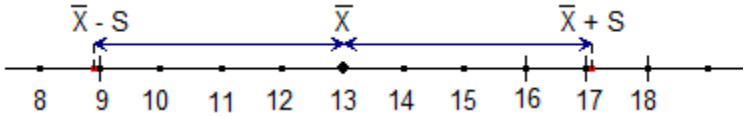
Colección A: 14, 16, 10, 11, 18, 9	*Colección B: 18, 16, 17, 9, 9, 9*
$\bar{X} = 13$	$\bar{X} = 13$
Rango: 9	*Rango: 9*
S = 3.27	*s= 4.04*

En la colección A, $\overline{X} - S = 9.73$ *y* $\overline{X} + S = 16.27$.



Entre esos valores quedan los datos 10, 11, 14 y 16 que representan $\frac{4}{6} = 0.6667 \rightarrow 66.67\%$ *de los datos.*

En la colección B, $\overline{X} - S = 8.96$ *y* $\overline{X} + S = 17.04$*.*



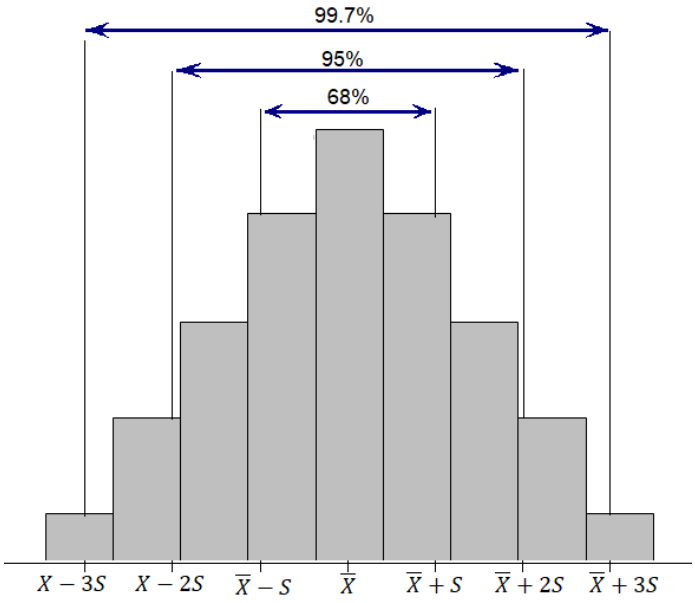
Entre esos valores quedan los datos 9, 9, 9, 16 y 17 que representan $\frac{5}{6} = 0.8333 \rightarrow 88.33\%$ *de los datos*

Hay una regla, conocida como *Regla Empírica* que describe aproximadamente qué porcentaje de los datos quedan alrededor de la media sumando y restando una, dos o tres veces la desviación estándar, siempre y cuando la distribución de frecuencias sea aproximadamente simétrica o tenga forma de campana de Gauss.

La regla establece que

- Entre $\overline{X} - S$ y $\overline{X} + S$ queda el 68% de los datos.
- Entre $\overline{X} - 2S$ y $\overline{X} + 2S$ queda el 95% de los datos.
- Entre $\overline{X} - 3S$ y $\overline{X} + 3S$ queda el 99.7% de los datos.

Se le conoce también como *Regla 68-95-99.7* por los porcentajes a que se refiere.



Cómo su nombre lo indica, la regla empírica es producto de la experiencia práctica de los estudios de muchos investigadores en diferentes ramas del conocimiento que han encontrado comportamientos bastante similares.

1.7 EJERCICIOS COMPLEMENTARIOS DE LA UNIDAD 1

1. Escribe a qué concepto se refieren los siguientes enunciados.
   a. Conjunto de elementos de interés en un estudio estadístico.
   b. Un subconjunto representativo de la población de estudio.
   c. La rama de la estadística que consiste en métodos para determinar alguna información acerca de la población con base en la información que brinda una muestra.
   d. La rama de la estadística que incluye una serie de técnicas para recopilar, ordenar, organizar, resumir y presentar datos de manera que resalten sus características más importantes, para analizar la información que aportan.
   e. Una variable que toma valores numéricos y es el resultado de contar.
   f. Gráfica de barras que se usa cuando los datos están agrupados en intervalos.
   g. El número de veces que aparece un valor en una colección de datos.
   h. El único gráfico que se construye con frecuencias acumuladas.
   i. Fórmula que permite calcular la cantidad de intervalos para una muestra de n datos.
   j. Medida que divide a los datos ordenados en dos partes de igual tamaño.
   k. El valor o valores que tiene la mayor frecuencia en los datos.
   l. El promedio de los cuadrados de las desviaciones de los datos respecto a su media.
   m. Regla que permite aproximar qué porcentaje de los datos se encuentran alrededor de la media, a una distancia de 1, 2 o 3 veces la desviación estándar.

2. Se registró el tiempo (en minutos) que tarda en hacer efecto un medicamento en una muestra de 60 pacientes. Esta es la información.

15	18	32	19	16	21	22	16	17	25	30	22	27	30	32
35	15	18	19	20	17	26	23	20	21	20	16	18	16	30
20	28	21	19	23	28	21	19	18	20	27	23	21	22	29
18	17	26	15	18	29	14	32	29	16	36	19	18	16	22

   a. ¿De qué tipo de variable se trata?
   b. Construye una tabla de frecuencias por intervalos.

c. Escribe 4 observaciones acerca de la información que brinda la tabla.
d. Traza una gráfica adecuada al tipo de variable. No olvides escribir títulos y la escala de los ejes.
e. Determina la media, la mediana y la moda.
f. Escribe 4 observaciones de la información que brindan las medidas de tendencia central.
g. Determina los cuartiles y traza el esquema de caja con brazos.
h. Escribe 4 observaciones acerca de la información que brinda el esquema.

3. Las distribuciones de edades en dos comunidades, tienen las siguientes medidas. En ambas comunidades la menor edad es 0 años.

	Comunidad A	Comunidad B
Edad media	35	35
Edad mediana	28	34
Moda	21	30
Rango	80	60
Desviación estándar	5.8	3.2

a. ¿En qué comunidad las edades están más concentradas?

b. ¿Cuál es la edad máxima en la comunidad A?, ¿Y en la comunidad B?

c. ¿Entre qué edades de encuentra la mitad más joven de la población de la comunidad A? ¿Y la mitad de mayor edad?

d. ¿Entre qué edades de encuentra la mitad más joven de la población de la comunidad B? ¿Y la mitad de mayor edad?

e. ¿Cuál es la edad más frecuente en cada comunidad?

f. ¿Cómo explicas que la edad promedio sea igual en ambas comunidades?

UNIDAD II. OBTENCIÓN E INTERPRETACIÓN DE INFORMACIÓN ESTADÍSITCA EN DATOS BIVARIADOS

Presentación

Considera que en cada elemento de la población o muestra ahora se observan dos características, representadas por dos variables, por ejemplo, situación de empleo y género de personas adultas, o longitud del antebrazo y altura de una persona. La pregunta de interés en esta situación es: ¿Las dos variables están relacionadas de alguna forma?

Los métodos para buscar una respuesta a esta pregunta dependen del tipo de las variables involucradas. En esta unidad estudiarás métodos que aportan información al respecto cuando las dos variables son cualitativas y cuando ambas son cuantitativas.

Propósito

Al terminar esta unidad 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 para interpretar críticamente la información estadística.

2.1 INTRODUCCIÓN

Cuando en cada elemento de la población o muestra se observan los valores de dos variables estadísticas distintas, lo que se obtiene son datos bivariados.

Ejemplo 1.

a) *Se selecciona un estudiante al azar de 5° semestre en el plantel y se registra la carrera que piensa estudiar y su género. En este caso, son ejemplos de datos bivariados (medicina, mujer), (veterinaria, hombre) y (economía, mujer).*

b) *En una muestra de camiones de carga pesada que llegan a una caseta de la autopista, se mide el ancho y la altura en metros. Son ejemplos de datos bivariados que se pueden obtener en este caso: (2.55, 4), (2.6, 4.1) y (2.4, 3.8)*

c) *A cada trabajador de una empresa, se le pide que indique el nivel máximo de estudios que tiene, y su ingreso mensual en pesos. Son ejemplos de datos bivariados (licenciatura, 15000), (bachillerato, 8000) y (secundaria, 6000)*

Así que cuando se habla de una muestra de *n* datos bivariados, es porque hay *n* parejas de datos.

En el primer inciso del ejemplo anterior, se trata de dos variables cualitativas, en el segundo inciso las dos variables son cuantitativas y en el tercero se tiene una variable cualitativa y una cuantitativa.

Al estudiar estos casos, el objetivo es determinar qué tan asociadas están las variables involucradas. En esta unidad estudiarás algunos métodos para analizar lo anterior cuando las dos variables son cualitativas y cuando ambas son cuantitativas.

2.2 DOS VARIABLES CUALITATIVAS

A. Tablas de contingencia

Cuando los datos bivariados son dos variables cualitativas, resulta conveniente organizarlos en una tabla de doble entrada donde los renglones representan a las categorías de la variable 1, mientras que las columnas representan a las categorías de la variable 2. A esta tabla se le llama *Tabla de Contingencia*,

En cada celda interior, se escribe la frecuencia absoluta de los elementos que simultáneamente tienen el valor del renglón y el de la columna correspondientes.

Ejemplo 2.

La siguiente tabla muestra el número de pacientes hospitalizados por la misma enfermedad, según el género y el hospital en el que fueron tratados, en los últimos 6 meses. Los dos primeros hospitales (Ángeles del Pedregal y Médica Sur) son privados. Los otros dos (20 de Noviembre y López Mateos) son públicos.

	Hospital			
Género	**Ángeles del Pedregal**	**Médica Sur**	**20 de Noviembre**	**López Mateos**
Hombre	45	**29**	223	148
Mujer	30	42	**212**	153

Las variables son:

- *Género, que toma los valores hombre y mujer.*
- *Hospital, que toma los valores Ángeles del Pedregal, Médica Sur, 20 de Noviembre y López Mateos*

La cantidad de pacientes hombres que fueron tratados en Médica Sur es 29; y 212 es la cantidad de pacientes mujeres que fueron tratadas en el 20 de Noviembre.

Al sumar las frecuencias absolutas de cada fila y de cada columna, se obtienen las ***frecuencias absolutas marginales***.

Ejemplo 3.

	Hospital				
Género	**Ángeles del Pedregal**	**Médica Sur**	**20 de Noviembre**	**López Mateos**	**Total**
Hombres	45	29	223	148	
Mujeres	30	42	212	153	**437**
Total	**75**		435		

Las sumas calculadas indican que:

- *75 pacientes estuvieron hospitalizados en el hospital Ángeles del Pedregal por esta enfermedad.*
- *437 pacientes eran mujeres.*

Completa las sumas de la tabla y escribe lo que falta a continuación.

- _____ *pacientes estuvieron hospitalizados en el 20 de Noviembre.*
- _____ *pacientes eran hombres.*
- *301 pacientes* _____.
- *El total de pacientes en los que se hizo el estudio es* _____.

La organización de los datos bivariados en tablas de contingencia permite observar algunas de las características de los mismos. Por ejemplo, en los datos anteriores podemos observar que:

- La cantidad de hombres es similar a la de mujeres.
- En los dos hospitales privados se atendió a un número mucho más chico de pacientes que en los dos hospitales públicos.
- El hospital que más pacientes atendió por esta enfermedad fue el 20 de Noviembre.

Para poder profundizar el análisis, conviene construir otras tres tablas: de frecuencias relativas, de porcentajes por renglón y de porcentajes por columna.

B. Tablas de frecuencias relativas

Si dividimos las cantidades de las celdas de la tabla entre el número total de datos, obtenemos las frecuencias relativas de las distintas parejas de categorías. Las vamos a escribir en porcentajes.

Ejemplo 4.

	Hospital				
Género	***Ángeles del Pedregal***	***Médica Sur***	***20 de Noviembre***	***López Mateos***	***Total***
Hombres	5.10%	3.29%	25.28%	16.78%	50.45%
Mujeres	3.40%	4.76%	24.04%	17.35%	49.55%
Total	8.50%	8.05%	49.32%	34.13%	100.00%

Algunos ejemplos de las operaciones que se hicieron son:

$$\frac{45}{882} \times 100 = 5.10\% \qquad \frac{29}{882} \times 100 = 3.29\%$$

De la misma forma se calcularon los demás porcentajes.

La descripción de los porcentajes obtenidos debe incluir las dos categorías que corresponden a cada celda. Por ejemplo:

- *16.78% de los pacientes eran hombres y fueron hospitalizados en el López Mateos.*
- *3.40% de los pacientes eran mujeres y fueron hospitalizadas en el Ángeles del pedregal.*

Las frecuencias relativas marginales son las que corresponden a los porcentajes de los totales, por ejemplo:

- *34.13% de los pacientes fueron hospitalizados en el López Mateos.*
- *50.45% de los pacientes eran hombres.*

Completa los siguientes enunciados.

- *_____ de los pacientes eran hombres y estuvieron hospitalizados en el Ángeles del Pedregal.*
- *_____ de los pacientes eran mujeres y estuvieron hospitalizados en Médica Sur.*
- *24.04% de los pacientes _____.*

Observa que el cálculo de las frecuencias relativas permite hacer otras observaciones globales sobre la información que brindan los datos. Por ejemplo:

- Se confirma que aproximadamente la mitad de los pacientes son hombres y la mitad mujeres.
- Los porcentajes de pacientes de cada género tratados en los hospitales privados, son semejantes: Más o menos entre el 3% y el 5% cada uno de ellos.
- En el 20 de Noviembre se trató a prácticamente la mitad de los pacientes.
- En los dos hospitales públicos fue tratado el 49.32 + 34.13 = 83.45% de los pacientes.

C. Tablas de porcentajes por renglón

Otra tabla que ayuda al análisis es la que indica qué porcentaje del total de cada renglón, corresponde a cada una de las categorías de las columnas. Es decir, en este caso, se toma como 100% el total de cada renglón.

Ejemplo 5.

	Hospital				
Género	***Ángeles del Pedregal***	***Médica Sur***	***20 de Noviembre***	***López Mateos***	***Total***
Hombres	10.11%	6.52%	50.11%	33.26%	100%
Mujeres	6.86%	9.61%	48.51%	35.01%	100%
Total	8.50%	8.05%	49.32%	34.13%	100%

Estos son ejemplos de los cálculos realizados en el primer renglón:

$$\frac{45}{445} \times 100 = 10.11\% \qquad \frac{29}{445} \times 100 = 6.52\% \, .$$

En el segundo renglón, el total o 100% es 437.

Ahora la descripción de los porcentajes es por género. Por ejemplo:

- *10.11% de los hombres, fueron hospitalizados en el Ángeles del Pedregal.*
- *9.61% de las mujeres, fueron hospitalizadas en el Médica Sur.*

Escribe lo que falta a continuación.

- _____ *de las mujeres, estuvieron hospitalizadas en el López Mateos.*
- _____ *de los hombres, estuvieron hospitalizados en el 20 de Noviembre.*
- *6.86% de* _____.

- *33.26% de _____.*

D. Tablas de porcentajes por columna

En esta tabla se indica qué porcentaje del total de cada columna corresponde a cada una de las categorías de los renglones. Es decir, aquí se toma como 100% el total de cada columna.

Ejemplo 6.

Los porcentajes de la primera columna se calculan sobre el total de pacientes hospitalizados en el Ángeles del Pedregal, es decir, 75. Por ejemplo,

$\frac{45}{75} \times 100 = 60\%$ $\frac{30}{75} \times 100 = 40\%.$

	Hospital			
Género	***Ángeles del Pedregal***	***Médica Sur***	***20 de Noviembre***	***López Mateos***
Hombres	*60.00%*	*40.85%*	*51.26%*	*49.17%*
Mujeres	*40.00%*	*59.15%*	*48.74%*	*50.83%*
Total	*100%*	*100%*	*100%*	*100%*

En la segunda columna, el total o 100% es 71.

En esta tabla, la descripción de los porcentajes es por hospital, por ejemplo:

- *De los pacientes que estuvieron en el Ángeles del Pedregal, el 60% eran hombres y el 40% mujeres*
- *El 48.74% de los pacientes que estuvieron en el hospital 20 de Noviembre eran _____, y el _____ eran _____.*

En el siguiente ejemplo, construiremos las 4 tablas que hemos visto e iremos haciendo notar algunas observaciones.

Ejemplo 7.

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.*

Las dos variables son: Problemas respiratorios que toma los valores: No tiene y Sí tiene. La segunda variable es Nivel de tabaquismo cuyas categorías son: Mucho, Moderado, Poco y Nada.

a) *Tabla de contingencia*

El total de encuestados es 400, y los porcentajes que se mencionan en la información son:

$$55.25\%\ de\ 400 = 221 \qquad 30\%\ de\ 400 = 120$$

$$25\%\ de\ 400 = 100 \qquad 20\%\ de\ 400 = 80$$

Usando los enunciados que nos dan, se pueden ir escribiendo cantidades en las distintas celdas de la tabla.

- *102 debe estar en la celda que corresponde a "Mucho" y "Sí tiene".*
- *35 debe estar en "Moderado" y "No tiene"*
- *El total de "No tiene" debe ser 221.*
- *Las frecuencias marginales de nivel de tabaquismo son: De "Mucho" → 120, de "Moderado" → 100, de "Poco" → 100 y de "Nada" → 80.*
- *77 debe estar en "Nada" y "No tiene"*
-

		Problemas Respiratorios		
		No tiene	*Sí tiene*	*Total*
Nivel de tabaquismo	*Mucho*		*102*	*120*
	Moderado	*35*		*100*
	Poco			*100*
	Nada	*77*		*80*
	Total	*221*		*400*

Con estas cantidades, podemos completar la tabla. Analiza qué operaciones debes hacer para determinar las cantidades que faltan en cada columna.

		Problemas Respiratorios		
		No tiene	*Sí tiene*	*Total*
Nivel de tabaquismo	*Mucho*	*18*	*102*	*120*
	Moderado	*35*	*65*	*100*
	Poco	*91*	*9*	*100*
	Nada	*77*	*3*	*80*
	Total	*221*	*179*	*400*

b) Tabla de frecuencias relativas

		Problemas Respiratorios		
		No tiene	*Sí tiene*	*Total*
Nivel de tabaquismo	*Mucho*	4.50%	25.50%	30%
	Moderado	8.75%	16.25%	25%
	Poco	22.75%	2.25%	25%
	Nada	19.25%	0.75%	20%
	Total	55.25%	44.75%	100%

Algunas observaciones:

- *55% de los encuestados, fuman mucho o moderado y 45% fuman poco o nada.*
- *Más de la cuarta parte de los encuestados, fuma mucho y sí tienen problemas respiratorios.*
- *Solo 3% de los encuestados fuman poco o nada y sí tienen problemas respiratorios.*

c) Tabla de porcentajes por renglón

		Problemas Respiratorios		
		No tiene	*Sí tiene*	*Total*
Nivel de tabaquismo	*Mucho*	15.00%	85.00%	100%
	Moderado	35.00%	65.00%	100%
	Poco	91.00%	9.00%	100%
	Nada	96.25%	3.75%	100%
	Total	55.25%	44.75%	100%

Ejemplos de observaciones:

- *De los encuestados que fuman mucho, el 85% tiene problemas respiratorios, y ese porcentaje baja a 65% entre los que fuman de manera moderada.*
- *De los que no fuman, más del 95% no tienen problemas respiratorios y ese porcentaje baja a 91% entre los que fuman poco.*

d) Tabla de porcentajes por columna

		Problemas Respiratorios		
		No tiene	*Sí tiene*	*Total*
Nivel de tabaquismo	*Mucho*	8.14%	56.98%	30%
	Moderado	15.84%	36.31%	25%
	Poco	41.18%	5.03%	25%
	Nada	34.84%	1.68%	20%
	Total	100%	100%	100%

Algunas observaciones

- *De los encuestados que no tienen problemas respiratorios, alrededor de 76% fuman poco o nada.*
- *De los que sí tienen problemas respiratorios, más del 93% fuman mucho o moderado.*

Después de las observaciones anteriores, ¿qué dirías sobre la relación entre el nivel de tabaquismo y los problemas respiratorios? _____

_____

EJERCICIOS 2.2.1

1. La tabla de contingencia siguiente representa el Estado Civil de un grupo de personas y su preferencia por ciertos periódicos.

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

	Periódico preferido				
Estado Civil	El Universal	Excélsior	Reforma	La Jornada	Total
Soltero	11	9	10	14	
Casado	10	6	10	8	
Viudo	6	4	5	5	
Separado	10	8	5	9	
Total					

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

- _____ personas prefieren leer el Excélsior.
- Se entrevistó a _____ personas viudas.
- ¿Cuántas personas son solteras y prefieren el periódico la Jornada? _____
- ¿Qué porcentaje de personas son casadas y prefieren el periódico Reforma? _____
- ¿Qué porcentaje de personas no son casadas? _____
- ¿Qué porcentaje de personas no leen Excelsior? _____
- De las personas separadas, el _____ % prefiere leer la Jornada
- De las personas viudas, ¿qué porcentaje prefiere leer el Reforma? _____
- ¿Qué estado civil tiene el mayor porcentaje de lectores de La Jornada? ___ _____
- De las personas que prefieren el Reforma, el _____ % son separadas
- De las personas que prefieren el Universal, ¿qué porcentaje son solteros? _____
- ¿Cuál de los periódicos tiene entre sus lectores el mayor porcentaje de casados? _____

b. ¿Qué opinas de la relación entre las variables Estado civil y Periódico que se lee? _____

2.3 RELACIÓN LINEAL ENTRE DOS VARIABLES CUNTITATIVAS

La relación entre dos variables numéricas puede ser funcional, si se trata de una relación matemática exacta entre ellas, o puede ser estadística si la relación entre ellas es sólo aproximada a una relación matemática. Aquí se va a abordar la relación estadística entre dos variables cuantitativas.

El modelo al que una colección de datos bivariados se aproxima, puede ser de muy distinta naturaleza, ya sea lineal, cuadrático, exponencial, logarítmico, trigonométrico, etc. En el curso de Estadística y Probabilidad I solo se aborda el modelo de relación lineal. La pregunta a responder es: ¿Qué tan lineal es la relación entre dos variables cuantitativas involucradas en una colección de datos bivariados?

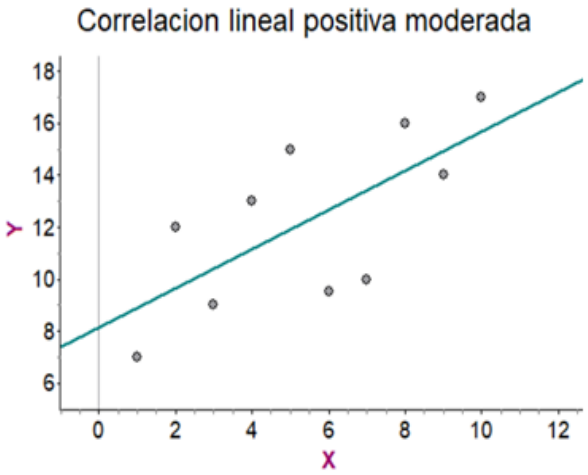
Para responder esta pregunta se harán tres pasos o procedimientos: la elaboración de un diagrama de dispersión, la determinación de una recta de regresión lineal y el cálculo del coeficiente de correlación lineal.

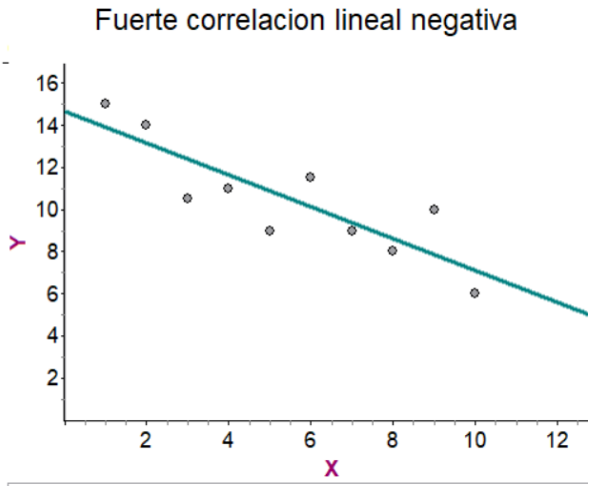
A. Diagrama de dispersión

En un plano cartesiano se localiza cada dato bivariado como un punto. La nube de puntos que se forma se llama diagrama de dispersión.

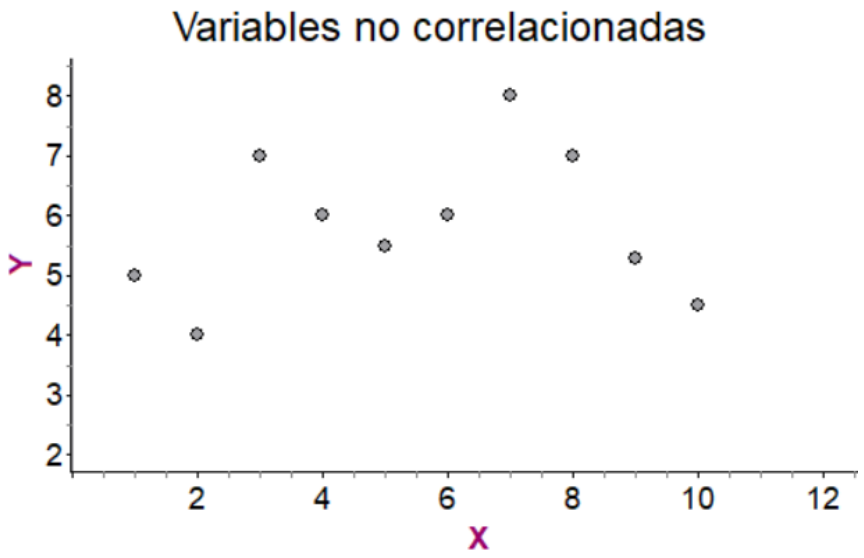
Se debe identificar cuál de las dos variables se usará como variable independiente (también llamada explicativa, regresora o exógena) y cuál será la dependiente (respuesta o endógena). La variable independiente se localiza en el eje horizontal y la dependiente en el eje vertical.

A veces, el diagrama da una idea de qué tan cerca o lejos están los datos de una relación lineal. Se habla de correlación lineal negativa cuando el modelo lineal aproximado tiene pendiente negativa. Análogamente, se habla de correlación lineal positiva cuando la recta aproximada tiene pendiente positiva.





Un diagrama de dispersión típico de datos que provienen de variables **no correlacionadas** es:

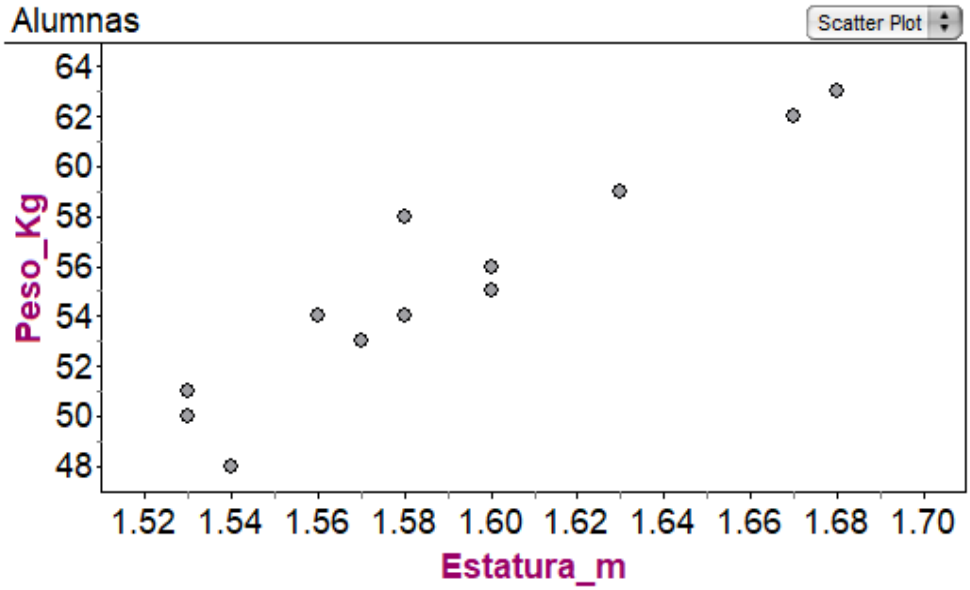


Ejemplo 8.

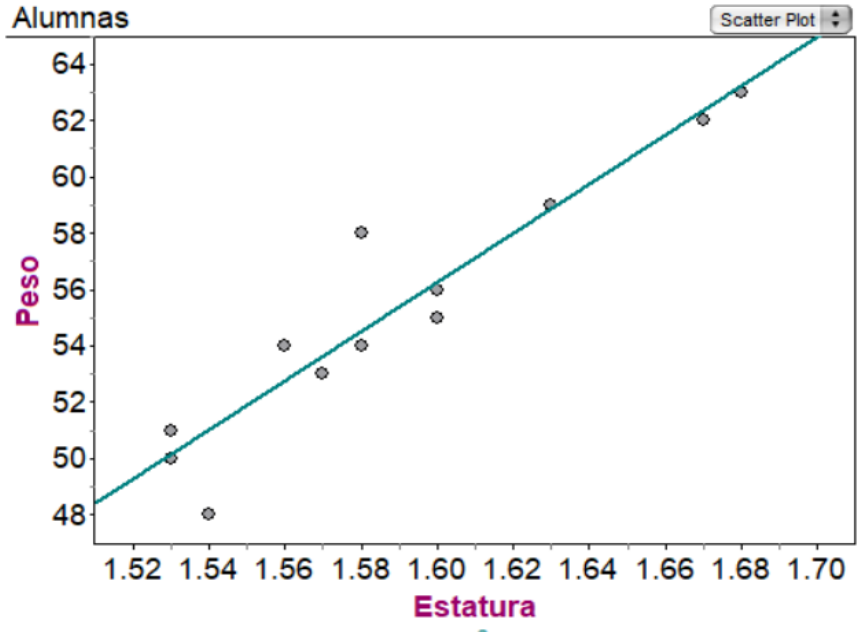
Para examinar la relación entre la estatura (en metros), y el peso (en kilogramos) de mujeres adolescentes, se tomó una muestra de 12 estudiantes. Los datos se muestran en la siguiente tabla.

Alumna	*Estatura (m.)*	*Peso (kg.)*
1	*1.60*	*56*
2	*1.63*	*59*
3	*1.68*	*63*
4	*1.67*	*62*
5	*1.53*	*50*
6	*1.58*	*54*
7	*1.57*	*53*
8	*1.58*	*58*
9	*1.54*	*48*
10	*1.60*	*55*
11	*1.56*	*54*
12	*1.53*	*51*

Localizamos en un plano cartesiano los puntos (1.60, 56), (1.63, 59), etcétera. Se obtiene el siguiente diagrama de dispersión.



Se puede ver que los puntos se encuentran más o menos cercanos a una recta. Pero para visualizar qué tan cercanos están, conviene trazar una recta.



Ahora se puede apreciar que hay dos puntos un poco lejanos pero todos los demás están muy cerca de una recta.

El diagrama de dispersión ofrece una representación visual que puede ser útil pero tiene muchas limitaciones para llegar a una conclusión sobre la linealidad de la

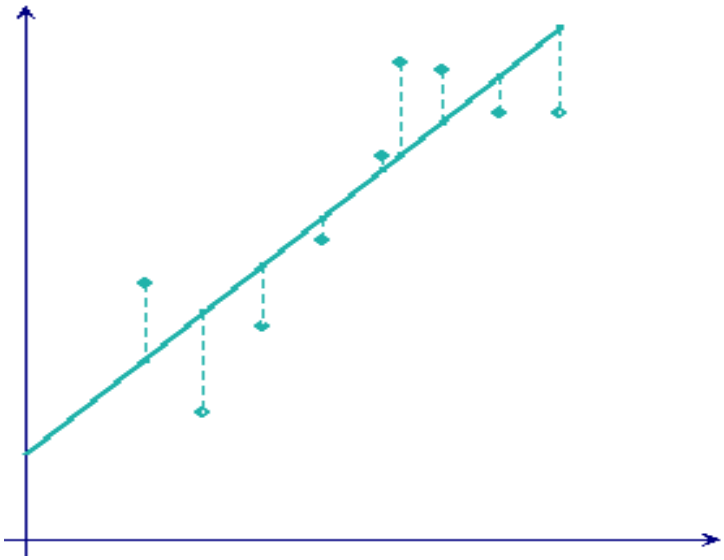
relación entre dos variables. Por ejemplo, los puntos se pueden ver más cercanos o más lejanos según la escala que se use en el eje vertical.

Por tanto, se requiere una medida de qué tan cerca están de una recta los puntos de una colección de datos bivariados.

B. Recta de regresión (mínimos cuadrados)

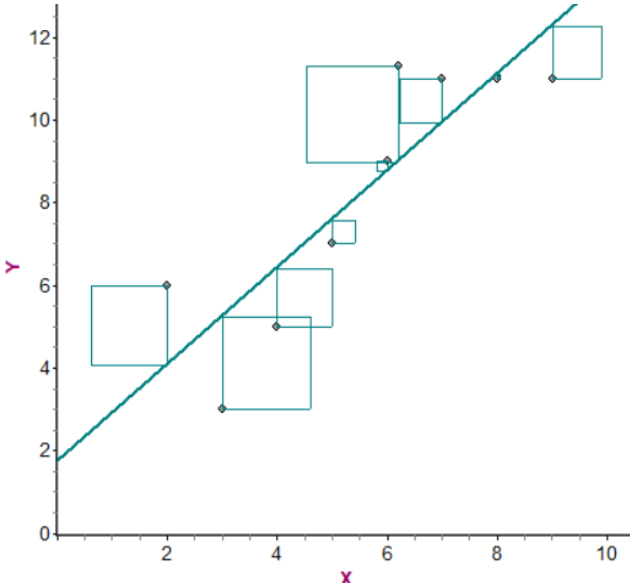
Antes de medir qué tan cerca o lejos de una recta está una colección de puntos, se requiere encontrar la recta con la que se va a comparar. Por ello, el propósito en esta parte es encontrar la recta que más se parece a una nube de puntos.

Un criterio puede ser que sea la recta para la cual la suma de distancias verticales sea lo más chica posible.



Pero algunas de estas diferencias serían positivas y otras negativas. Por eso, aunque la suma de las diferencias sea muy pequeña, los puntos pueden estar lejos de la recta, porque números positivos y negativos se pueden cancelar.

Una alternativa es buscar la recta para la cual la suma de los cuadrados de estas distancias es lo más chica posible.



Si los *n* datos bivariados son (x_1, y_1), (x_2, y_2), (x_3, y_3), …, (x_n, y_n), la recta que cumple la condición descrita tiene los siguientes parámetros.

Pendiente:

$$m = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2}$$

Ordenada al origen:

$$b = \frac{\sum_{i=1}^{n} y_i}{n} - (m)\frac{\sum_{i=1}^{n} x_i}{n}$$

A la recta que se construye se le llama ***recta de regresión*** por mínimos cuadrados.

Ejemplo 9.

Usando los datos del ejemplo 8, vamos a determinar la recta de regresión, Para ello, extendemos la tabla inicial de la siguiente manera.

Alumna	Estatura x_i	Peso y_i	$(x_i)(y_i)$	${x_i}^2$
1	1.60	56	89.60	2.5600
2	1.63	59	96.17	2.6569
3	1.68	63	105.84	2.8224
4	1.67	62	103.54	2.7889
5	1.53	50	76.50	2.3409
6	1.58	54	85.32	2.4964
7	1.57	53	83.21	2.4649
8	1.58	58	91.64	2.4964
9	1.54	48	73.92	2.3716
10	1.60	55	88.00	2.5600
11	1.56	54	84.24	2.4336
n = 12	1.53	51	78.03	2.3409
Total	$\sum x_i$ =19.07	$\sum y_i$ =663	$\sum x_i y_i$ =1056.01	$\sum x_i^2$ =30.3329

Pendiente:

$$m = \frac{(12)(1056.01) - (19.07)(663)}{(12)(30.3329) - (19.07)^2} = 87.026$$

Ordenada al origen:

$$b = \frac{663}{12} - 87.026\left(\frac{19.07}{12}\right) = -83.05$$

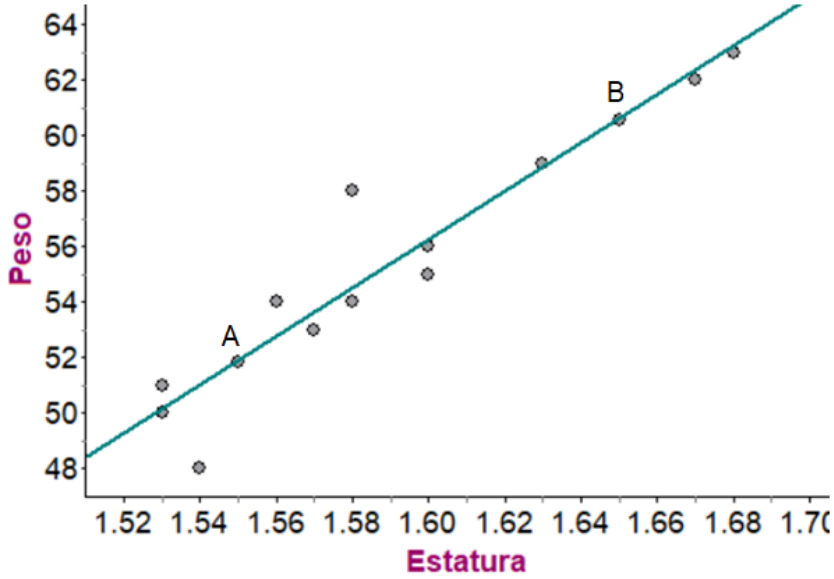
Ecuación de la recta de regresión:

$$y = 87.026x - 83.05$$

Para graficar esta recta en el mismo plano del diagrama de dispersión, conviene tomar dos valores para x en el rango donde estas los datos, y encontrar la y correspondiente de acuerdo a la ecuación anterior. Por ejemplo:

$x = 1.55$ $\quad y = 87.026(1.55) - 83.05 = 51.84$ $\quad Punto: A(1.55, 51.84)$

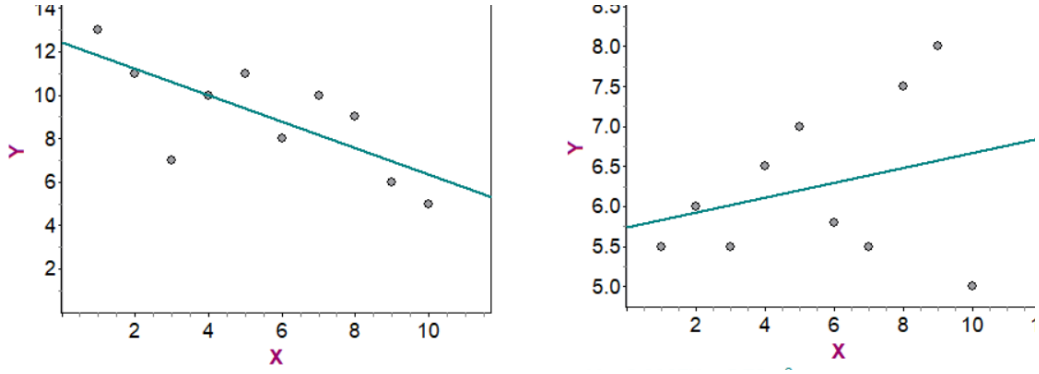
$x = 1.65$ $\quad y = 87.026(1.65) - 83.05 = 60.54$ $\quad Punto: B(1.65, 60.54)$



B. Coeficiente de correlación lineal

En cualquier colección de puntos, se puede encontrar la recta de regresión por mínimos cuadrados. Pero que exista esa recta no significa que los puntos sean muy cercanos a ella, nada más que es la recta que más se parece a la nube de puntos.

Por ejemplo, en los siguientes diagramas se ven dos nubes de puntos y las rectas de regresión correspondientes. La distancia de los puntos a la recta es muy diferente en esos dos casos.



Para tener una medida de qué tan cerca están los puntos de la recta de regresión, se calcula el Coeficiente de correlación lineal (de Pearson).

Si los *n* datos bivariados son (x_1, y_1), (x_2, y_2), (x_3, y_3), …, (x_n, y_n), el coeficiente de correlación lineal se calcula usando la fórmula

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n\sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2}\,\sqrt{n\sum_{i=1}^{n} {y_i}^2 - (\sum_{i=1}^{n} y_i)^2}}$$

El valor de r está siempre entre -1 y 1.

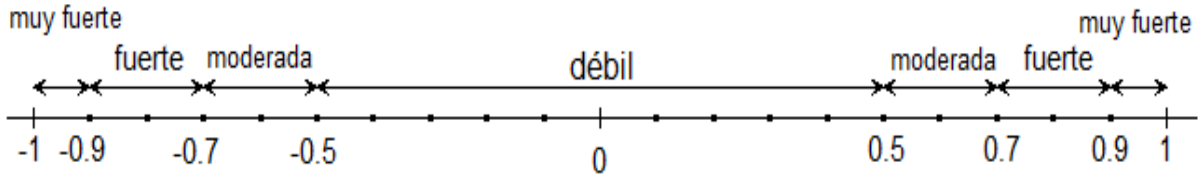
Cuando la recta de regresión tiene pendiente negativa, el valor de r está en el intervalo [-1, 0). Cuando la pendiente de la recta de regresión es positiva, r toma valores en el intervalo (0, 1].

Si r = -1 la correlación negativa es perfecta, y cuando r = 1 la correlación positiva es perfecta. Estos casos nunca se presentan en datos reales pues lo que representan es que los puntos caen exactamente sobre la recta. Pero son puntos de referencia.

Si r = 0, las variables no están correlacionadas

La interpretación de los demás valores de r es la siguiente. Si el valor de r está en

- el intervalo (-1,-0.9] o en [0.9,1) indica una correlación lineal negativa o positiva según corresponda, muy fuerte
- el intervalo (-0.9,-0.7] o en [0.7,0.9) indica una correlación lineal negativa o positiva según corresponda, fuerte.
- el intervalo (-0.7,-0.5] o en [0.5,0.7) indica una correlación lineal negativa o positiva según corresponda, moderada.
- el intervalo (-0.5,0) o en (0,0.5) indica una correlación lineal negativa o positiva según corresponda, débil.



Ejemplo 10.

De nuevo usaremos los datos del ejemplo 8 para calcular el coeficiente de correlación lineal. Los primeros cálculos que deben hacerse son:

Alumna	Estatura x_i	Peso y_i	$(x_i)(y_i)$	x_i^2	y_i^2
1	1.60	56	89.60	2.5600	3136
2	1.63	59	96.17	2.6569	3481
3	1.68	63	105.84	2.8224	3969
4	1.67	62	103.54	2.7889	3844
5	1.53	50	76.50	2.3409	2500
6	1.58	54	85.32	2.4964	2916
7	1.57	53	83.21	2.4649	2809
8	1.58	58	91.64	2.4964	3364
9	1.54	48	73.92	2.3716	2304
10	1.60	55	88.00	2.5600	3025
11	1.56	54	84.24	2.4336	2916
12	1.53	51	78.03	2.3409	2601
Total	$\sum x_i = 19.07$	$\sum y_i = 663$	$\sum x_i y_i = 1056.01$	$\sum x_i^2 = 30.3329$	$\sum y_i^2 = 36865$

Con la información de la tabla anterior, el cálculo de ***r*** *queda así:*

$$\boldsymbol{r} = \frac{(12)(1056.01) - (19.07)(663)}{\sqrt{(12)(30.3329) - (19.07)^2}\ \sqrt{(12)(36865) - (663)^2}} = 0.94$$

Por lo que, la correlación lineal entre la estatura y el peso de las adolescentes de la muestra, es muy fuerte.

Cuando el coeficiente de correlación indica que la linealidad de la relación entre las variables es cuando menos moderada, la recta de regresión se puede usar para hacer ***predicciones***, es decir, para determinar un valor aproximado de la variable dependiente que corresponde a un valor de la variable dependiente que no se encuentra en los datos.

Pero las predicciones solo se pueden hacer para valores de la variable dependiente que estén en el rango de los datos conocidos.

Ejemplo 11.

Con los datos de estatura-peso que hemos venido trabajando, la recta de regresión puede ser usada para estimar los pesos que corresponden a estaturas como las que se indican en esta tabla.

Altura (m)	Peso (kg)
1.55	51.84
1.59	55.32
1.61	57.06
1.62	57.93
1.65	60.54

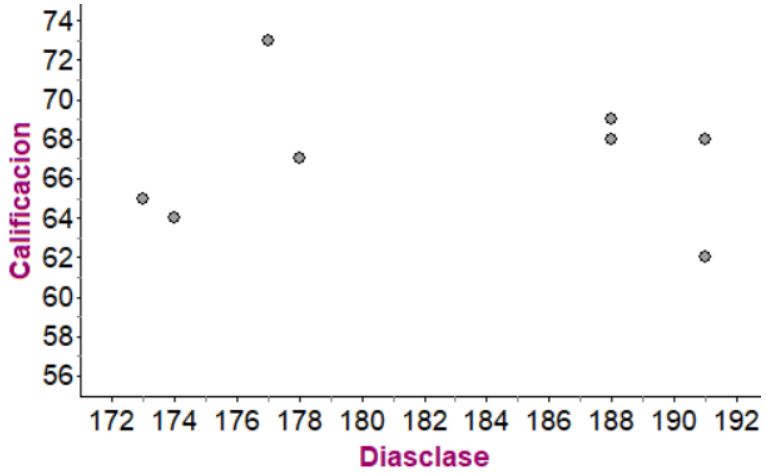
En el siguiente ejemplo se hace el análisis completo de regresión y correlación lineal.

Ejemplo 12.

En la siguiente tabla se presentan los datos de dos variables. X: es el número de días de clases en el año escolar en educación básica y Y: es la calificación promedio en un examen de Ciencias aplicado a alumnos de 13 años de varios países.

País	***Días del año escolar***	***Calif. Promedio en Ciencias***
1. Irlanda	*173*	*65*
2. Francia	*174*	*64*
3. Hungría	*177*	*73*
4. Estad. Unidos	*178*	*67*
5. Canadá	*188*	*69*
6. España	*188*	*68*
7. Jordania	*191*	*62*
8. Escocia	*191*	*68*
Total	$\sum x_i$ *=1460*	$\sum y_i$ *=536*

a) Diagrama de dispersión



El diagrama muestra una muy débil relación lineal entre las variables.

b) *Recta de regresión*

País	Días x_i	Califc y_i	$(x_i)(y_i)$	x_i^2	y_i^2
1	173	65	11245	29929	4225
2	174	64	11136	30276	4096
3	177	73	12921	31329	5329
4	178	67	11926	31684	4489
5	188	69	12972	35344	4761
6	188	68	12784	35344	4624
7	191	62	11842	36481	3844
8	191	68	12988	36481	4624
Total	1460	536	97814	266868	35992

Pendiente:

$$m = \frac{(8)(97814) - (1460)(536)}{(8)(266868) - (1460)^2} = -0.01435$$

Ordenada al origen:

$$b = \frac{536}{8} - (-0.01435)\left(\frac{1460}{8}\right) = 69.62$$

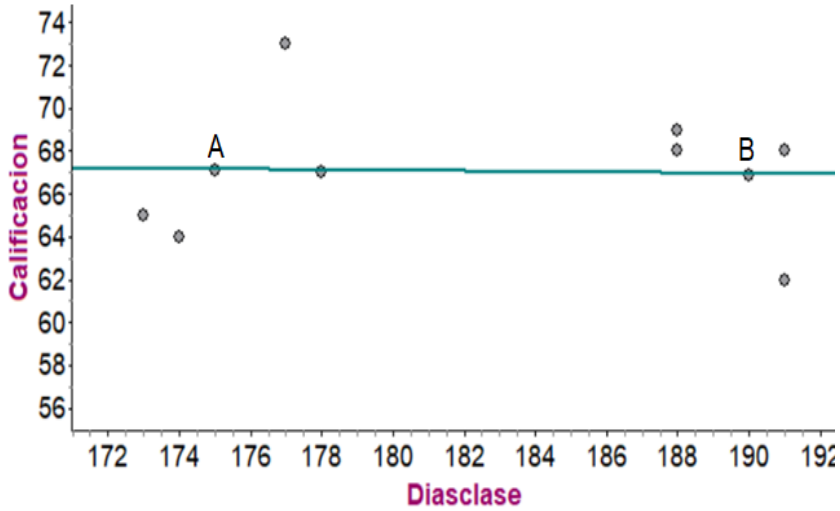
Ecuación:

$$y = -0.01435\,x + 69.62$$

Dos puntos para graficar la recta

$x = 175$ $\quad y = (-0.01435)(175) + 69.62 = 67.11$ $\quad A(175, 67.11)$

$x = 190$ $\quad y = (-0.01435)(190) + 69.62 = 66.89$ $\quad B(190, 66.89)$



c) *Coeficiente de correlación lineal*

$$\boldsymbol{r} = \frac{(8)(97814) - (1460)(536)}{\sqrt{(8)(266868) - (1460)^2}\ \sqrt{(8)(35992) - (636)^2}} = -0.0328$$

*El valor de **r** es muy cercano a 0, por lo que se puede concluir que las variables prácticamente no están correlacionadas.*

d) Predicciones

*No se pueden hacer predicciones usando la recta de regresión, por el valor de **r***

EJERCICIOS 2.3.1

1. En una tienda de descuento se tiene la siguiente situación para un determinado artículo.

Número de piezas (x_i)	1	3	5	10	12	15	24
Costo por pieza (y_i)	55	52	48	36	32	30	25

a) Encuentra la ecuación de la recta de regresión _____
b) Determina el valor del coeficiente de correlación lineal _____
c) ¿Qué nivel de correlación lineal indica el valor de ***r***? _____
d) ¿Se puede usar la recta de regresión para hacer predicciones? _____
e) Si una persona compra 20 piezas de ese artículo, ¿cuál sería el costo por pieza? _____

2. La siguiente tabla representa la densidad de un mineral (x) y su contenido de hierro (y)

x	*y*
2.8	27
3.0	30
3.2	30
3.2	34
3.4	36

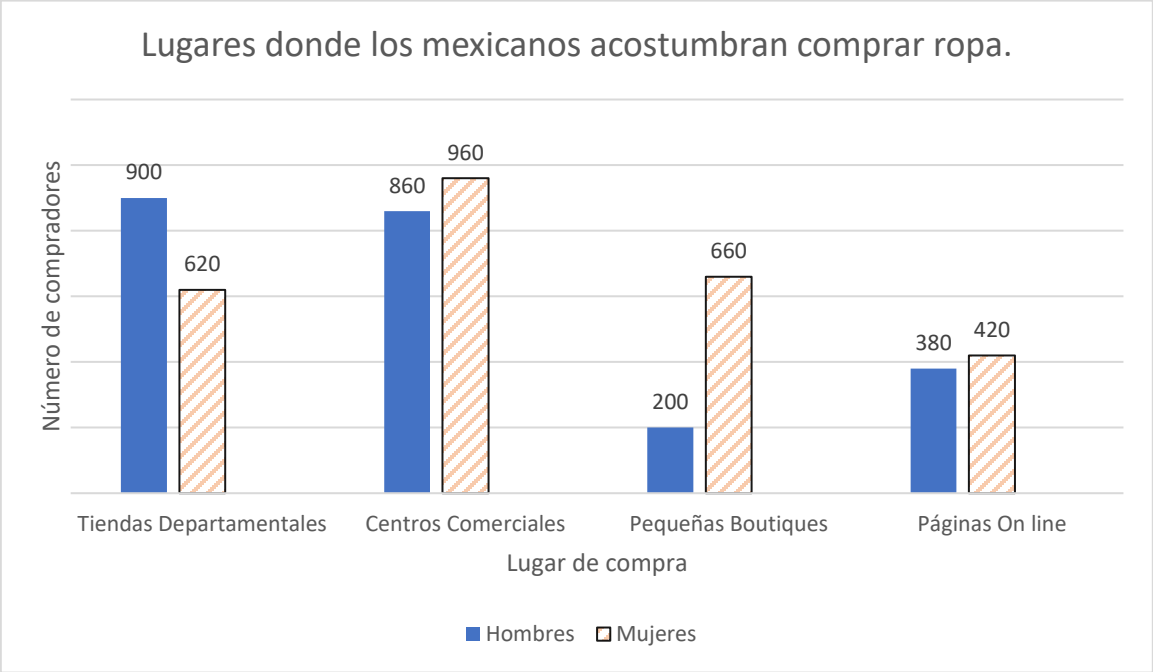
a) Construye el diagrama de dispersión.
b) Determina la ecuación de regresión lineal
c) Calcula el coeficiente de correlación ***r*** e interpreta su valor.
d) Traza la recta de regresión en el mismo plano cartesiano del diagrama de dispersión
e) ¿Se puede usar la recta de regresión para hacer predicciones? _____
f) Si la densidad del material es 2.9, determina el valor estimado del contenido de hierro.
g) Si el contenido de hierro es de 31, determina la densidad estimada del material

3. En un análisis de regresión la pendiente de la recta de mejor ajuste vale $m = 4.82$ y la ordenada al origen es $b = 5$.

a) La ecuación de esa recta de mejor ajuste es _____
b) Si el nivel de correlación lineal es fuerte, ¿qué valor aproximado para *y* arroja la recta de regresión para un valor de $x = 2$?_____

2.4 EJERCICIOS COMPLEMENTARIOS DE LA UNIDAD 2

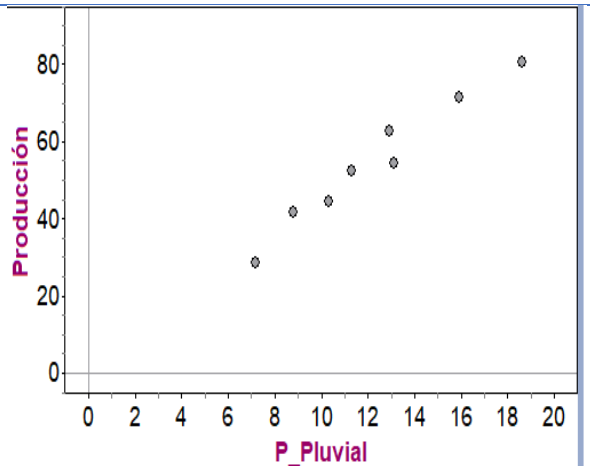
1) Se realizó una encuesta para conocer en qué lugares compran su ropa los mexicanos. La gráfica representa estos resultados, a partir de ella, realiza la tabla de contingencia correspondiente y contesta las preguntas requeridas.



a. ¿Cuáles son las dos variables representadas en la gráfica? _____ _____
b. ¿Qué tipo de variables son? _____
c. ¿Cuántas personas fueron encuestadas? _____
d. El lugar menos elegido por las mujeres para comprar ropa es_____ _____
e. El lugar preferido por los hombres para comprar ropa es_____
f. El ____ % de los encuestados prefiere comprar ropa en los centros comerciales.
g. Las pequeñas boutiques son preferidas por más _____ que _____
h. 53.2% de todos los encuestados son _____
i. ¿Quiénes compran más en línea: los hombres o las mujeres? _____

2) En cierto condado de Estados Unidos, se han registrado los niveles anuales de precipitación pluvial en pulgadas y la producción de trigo durante los últimos 8 años:

Precipitación pluvial (in)	Producción de Trigo (bushels por acre)
12.9	62.5
7.2	28.7
11.3	52.2
18.6	80.6
8.8	41.6
10.3	44.5
15.9	71.3
13.1	54.4



a. De acuerdo al diagrama de dispersión ¿Qué nivel de correlación lineal observas entre la cantidad anual de precipitación pluvial y la producción de trigo? _____

b. En tu cuaderno, encuentra los parámetros de la recta de regresión y escribe aquí la ecuación de dicha recta _____

c. Calcula el coeficiente de correlación r _____

d. ¿Cómo interpretas el valor que obtuviste para r _____

_____

e. Si la cantidad anual de lluvia se incrementara ¿qué sucedería con la producción de trigo? _____

f. Si la cantidad anual de lluvia se incrementa en una pulgada ¿De cuántos bushels sería la variación en la producción anual de trigo? _____

g. Para un año en particular se produjeron 66.53 bushels de trigo por acre, ¿Cuál fue la cantidad anual de precipitación pluvial? _____

h. A partir del modelo de regresión lineal calculado en el inciso b, se obtiene que se producen 26.75 bushels por acre cuando la cantidad anual de precipitación pluvial es de 6 pulgadas, ¿es este pronóstico confiable? _____ Argumenta tu respuesta _____

UNIDAD 3. AZAR, MODELACIÓN Y TOMA DE DECISIONES

Presentación

Ya se ha comentado que es muy costoso y muy tardado analizar a toda una población para estudiar las características que interesan, y que la Estadística ofrece una alternativa que permite obtener ciertas conclusiones acerca de la población con base en el estudio de muestras aleatorias representativas. El principal sustento teórico de estos procedimientos (que abordarás en sexto semestre) es la Probabilidad, debido a que esta rama de las matemáticas se relaciona con la toma de decisiones en condiciones de incertidumbre.

En esta unidad estudiarás los primeros conceptos relacionados con la Probabilidad, para poder contestar preguntas como:
¿Cuál es el objeto de estudio de la Probabilidad?
¿Cómo asignar una medida de probabilidad a un evento o suceso?
¿Cómo se pueden construir eventos compuestos a partir de eventos simples?
¿Cómo se modifica la probabilidad cuando se cuenta con información parcial?

Propósito

En esta unidad continuarás el desarrollo de tu 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 EXPERIMENTOS DETERMINISTAS Y ALEATORIOS

Todos los hechos que ocurren se denominan fenómenos.

Un **fenómeno determinista** es aquel cuyo resultado se predice con certeza, porque obedece a una relación causa-efecto. Por ejemplo, ¿cuándo cambiará el presidente de México? O, ¿cuándo será visto en México el siguiente eclipse total de sol?

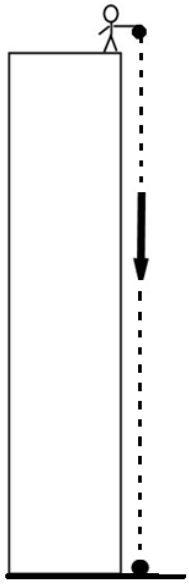
Un **fenómeno aleatorio** es aquel que tiene varios resultados posibles y estos no se pueden predecir con certeza, pues obedecen las leyes del azar. Por ejemplo, ¿cuánto crecerá la economía mexicana el próximo año? O, ¿cuántos huracanes entrarán a tierra en Norteamérica en la próxima temporada?

Para estudiar los fenómenos de nuestro entorno, lo que hacemos son experimentos. Un experimento es un proceso por medio del cual se registra una observación o una medición.

Hay dos tipos de experimentos a los que nos referiremos como experimentos deterministas y experimentos aleatorios.

Experimento Determinista

Es un experimento que cada vez que se repite en las mismas condiciones, produce los mismos resultados. Un ejemplo de este tipo de experimentos, es la caída libre.

	Cada vez que un objeto (libre de fricción) se deja caer desde una altura fija, se produce un movimiento que tiene siempre la misma trayectoria, recorrida en el mismo tiempo, y el objeto llega al suelo con la misma velocidad.

Experimento Aleatorio

Es un experimento con la característica de que se conocen los resultados que se pueden obtener, pero no es posible determinar cuál de esos resultados se obtendrá antes de realizar el experimento. Y esto sucede cada vez que el experimento se va a llevar a cabo.

Por ejemplo, el lanzamiento de un dado común, es un experimento aleatorio, porque se sabe que puede salir uno de los números 1, 2, 3, 4, 5 o 6, pero antes de lanzarlo no se puede saber cuál de ellos se obtendrá.	

El **objeto de estudio de la Probabilidad**, son los experimentos aleatorios.

EJERCICIOS 3.1.1

1. Se coloca una bola negra en la caja 1 y una bola roja en la caja 2.

a. Se le pide a una persona que va pasando y que no sabe cómo fueron acomodadas las bolas, que saque la que se encuentra en la caja 2 y registre su color.
¿Qué clase de experimento es este? _____ Explica tu respuesta.
_____

b. Se le pide a una persona que va pasando y que no sabe cómo fueron acomodadas las bolas, que lance una moneda al aire y si sale águila abra la caja 1 mientras que si sale sol abra la caja 2. Que observe la bola obtenida y registre su color.
¿Qué clase de experimento es este? _____ Explica tu respuesta.
_____

3.2 ESPACIO MUESTRAL Y DIFERENTES TIPOS DE EVENTO

El conjunto de todos los resultados que pueden salir al realizar un experimento aleatorio, se llama espacio muestral. Generalmente se denota por Ω (o por *S*). A cada uno de los resultados se le llama *punto muestral* o *evento elemental.*

Un *evento* es una característica que el resultado puede o no tener. Cada evento se identifica con el conjunto de resultados que tiene la característica (a los que llamaremos *resultados favorables* al evento), así que también se puede decir que un evento es un subconjunto de Ω.

Ejemplo 1.

Experimento: se lanza un dado común (es decir, un dado cúbico, marcado con los números del 1 al 6 y bien balanceado). Al caer el dado, se observa la cara superior.

El espacio muestral es

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

Algunos eventos son:

A: Sale un número impar.	A = {1, 3, 5}
B: Sale un número mayor que 4.	B = {5, 6}
C: Sale un número primo.	C = {2, 3, 5}
D: Sale un múltiplo de 3.	D = {3, 6}
E: Sale el 4.	E = {4}
F: Sale un número menor que 7.	F = {1, 2, 3, 4, 5, 6} = Ω
G: Sale un número mayor que 8.	G = ϕ (conjunto vacío)

Los eventos como F, que se identifican con todo el espacio muestral, se conocen como *eventos seguros*. Los eventos como G, que se identifican con el conjunto vacío, se conocen como *eventos imposibles*.

Ejemplo 2.

Se elige al azar un matrimonio de una cierta población y se le pregunta el número de hijos y el número de hijas que tiene.

El espacio muestral está formado por parejas:

$$\Omega = \{(a, b) \mid a \text{ es el número de hijos, } b \text{ es el número de hijas}\}$$

Algunos eventos son:

A: Tiene 2 hijos y una hija. $A = \{ (2, 1) \}$

B: No tiene hijos ni hijas. $B = \{ (0, 0) \}$

C: Sólo tiene hijas. $C = \{ (0, a) \mid a \in \{1,2,3,\ldots 12\} \}$

3.3 ENFOQUES DE LA PROBABILIDAD

La probabilidad de un evento es una medida de la facilidad con la que ocurre ese evento al realizar el experimento aleatorio.

A lo largo de la historia, se han ido descubriendo distintas formas de asignarle una medida de probabilidad a un evento. Aquí veremos tres enfoques para hacer esta asignación: el enfoque frecuencial, el clásico y el subjetivo.

A. Enfoque frecuencial

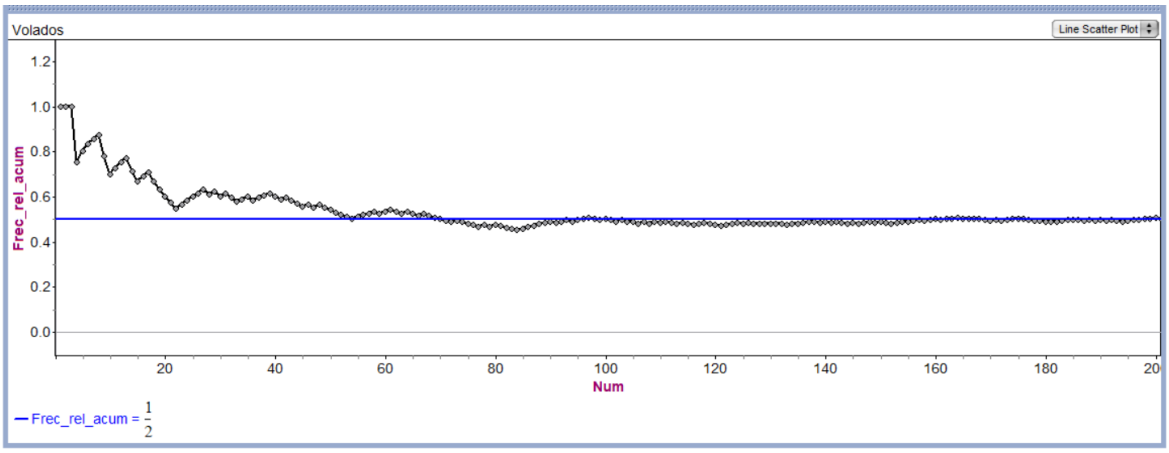
Los experimentos aleatorios sí tienen una regularidad, aunque ésta no sea del mismo tipo que la regularidad de los experimentos deterministas. Supongamos que un experimento aleatorio se repite un número grande de veces, siempre en las mismas condiciones, y se observa en cuántas de las repeticiones ocurre un evento E. La frecuencia relativa del evento E está dada por

$$fr_E = \frac{Número\ de\ veces\ que\ ocurre\ E}{Número\ de\ repeticiones\ realizadas}$$

La regularidad estadística o regularidad de frecuencias de los experimentos aleatorios, consiste en que al ir aumentando el número de repeticiones, las frecuencias relativas de cualquier evento E tienden a estabilizarse alrededor de un número *p*. Se trata de una regularidad que solo se puede observar al repetir muchas veces el experimento, y no en cada repetición particular.

Ejemplo 4.

La siguiente gráfica, muestra las frecuencias relativas del resultado *águila* al ir aumentando el número de lanzamientos de una moneda común, desde 1 hasta 200.



Simulación realizada en Fathom

Observa que en los primeros 50 lanzamientos, las frecuencias relativas tienen fuertes cambios, pero después, esas frecuencias tienden a estabilizarse alrededor de la línea azul que corresponde a la frecuencia ½ = 0.5

Esta regularidad dio lugar a definir la probabilidad de un evento como el número en torno al cual tienden a estabilizarse sus frecuencias relativas.

Sin embargo, es muy difícil determinar con precisión ese número (excepto en experimentos muy sencillos como el lanzamiento de una moneda). Por ello, el valor de la probabilidad se aproxima por una frecuencia relativa particular, usando un número grande de repeticiones. Así obtenemos:

$$P(E) \approx \frac{Número\ de\ veces\ que\ ocurre\ E}{Número\ de\ pruebas\ realizadas},$$

donde ≈ significa *aproximadamente igual.*

Nótese que usando este enfoque se obtiene solo una aproximación al valor de la probabilidad, aun cuando el número de pruebas realizadas sea muy grande, porque la variabilidad siempre estará presente por tratarse de un experimento aleatorio. Sin embargo, hay multitud de experimentos en los que no hay otra forma de asignar una medida de la facilidad con la que ocurre un evento.

Este enfoque para asignar probabilidades a eventos, permite comprender el significado de una medida de probabilidad. Cuando se afirma que la probabilidad de obtener águila al lanzar una moneda común es ½, eso **no significa** que de cada 2 lanzamientos uno será águila. **Lo que significa** es que si se hacen muchos lanzamientos de la moneda, en promedio, la mitad de los resultados obtenidos serán águilas.

Ejemplo 5.

Se elige un estudiante al azar en el plantel.
Se quiere determinar la probabilidad de los siguientes eventos:

A: nunca ha fumado

B: fuma

C: dejó de fumar

Como no se tiene ningún elemento que permita saber qué tan probable es cada evento, se aproxima esa probabilidad frecuencialmente, es decir, se repite muchas veces el experimento de seleccionar un estudiante al azar y preguntarle por su hábito en relación al tabaquismo. Supongamos que este experimento se repite 500 veces y se obtiene la siguiente información.

Evento	Frecuencia absoluta	Frecuencia relativa
Nunca ha fumado	268	0.536
Dejó de fumar	47	0.094
Fuma	185	0.370
Total	500	1.000

Esto nos permite afirmar que los valores aproximados de las probabilidades son:

$$P(A) \approx 0.536 \qquad P(B) \approx 0.094 \qquad P(C) \approx 0.370$$

EJERCICIOS 3.3.1

1) En una urna hay 7 bolas, algunas son rojas y otras son negras. Se repite 50 veces el experimento de seleccionar una bola al azar, regresando a la urna la bola elegida antes de seleccionar la siguiente (extracciones con reemplazo). En la siguiente tabla se reunió la información obtenida; R indica que fue roja y N indica que fue negra.

R	R	N	R	N	R	N	R	N	N
N	R	R	R	N	R	N	R	R	R
R	N	N	N	R	R	R	N	R	R
N	R	N	R	R	R	R	N	R	R
R	N	N	R	N	R	R	N	N	R

¿Cuántas bolas de cada color tiene la urna? Escribe tu procedimiento.

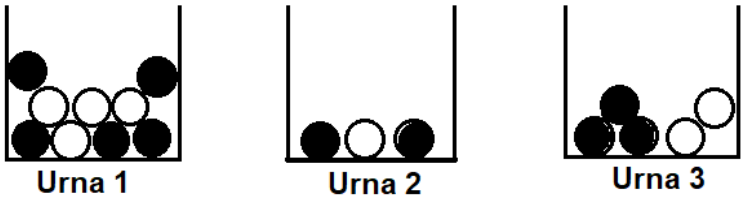
2) Se realizaron 530 encuestas a estudiantes de secundaria, y se obtuvo la siguiente información.

	Le gusta más Física	Le gusta más Historia	Le gusta más el Arte	Total
Hombre	102	88	63	
Mujer	76	104	97	
Total				

a. Completa la tabla
b. Se elige un estudiante al azar. Considera los eventos:
   A: Le gusta más Física
   B: Le gusta más Historia
   C: Le gusta más Arte
   D: Es mujer
   Usa la definición frecuencial para estimar las siguientes probabilidades
   P(A) = P(B) = P(D) =
   $P(D^c)$ = P(C) =

B. Enfoque clásico

Considera la siguiente situación. Vas a seleccionar una bola al azar de una urna y ganas si sale negra. Imagina que puedes elegir en cuál de las siguientes urnas hacer la extracción.



¿Cuál urna elegirías? _____

Si todas las bolas son iguales, excepto por el color, un razonamiento natural es analizar qué parte del total de bolas, son las bolas negras en cada caso. En la urna 1 las bolas negras son 5/9 del total, en la urna 2, son 2/3 del total y en la urna 3 son 3/5 del total.

¿Cuál de estas fracciones es mayor?

$$\frac{5}{9} = \frac{25}{45} \qquad \frac{2}{3} = \frac{30}{45} \qquad \frac{3}{5} = \frac{27}{45}$$

Por tanto, conviene más la urna 2.

De manera que lo que nos sirve es la razón entre el número de resultados favorables y el total de resultados que pueden salir. Generalizando esta idea, para medir la facilidad con la que ocurre un evento podemos calcular la fracción:

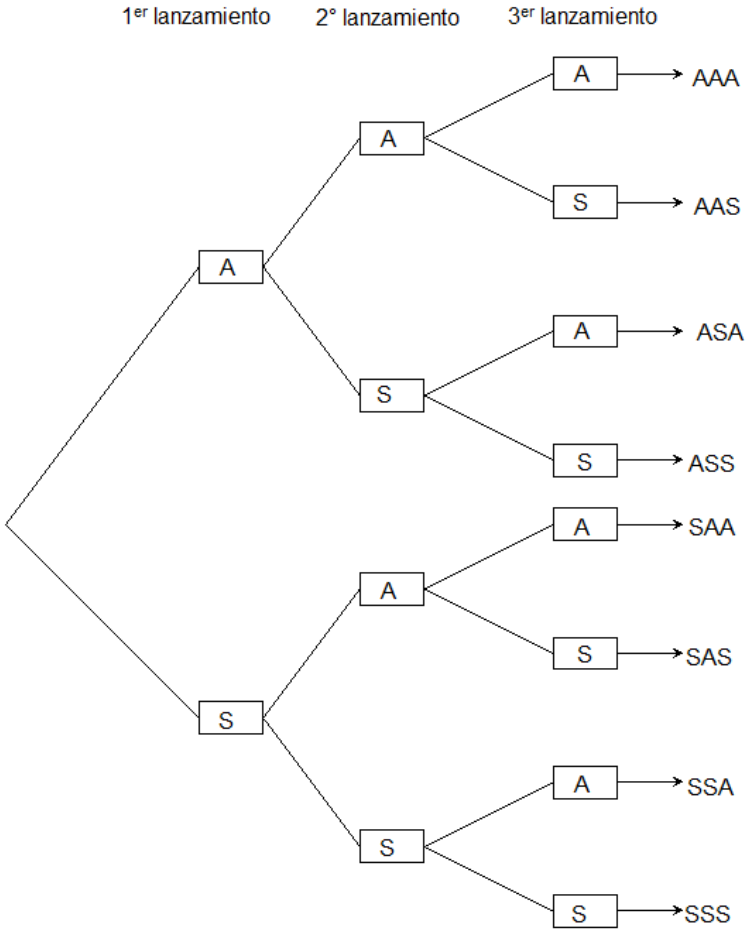
$$P(E) = \frac{Número\ de\ resultados\ favorables}{Total\ de\ resultados\ posibles\ equiprobables}$$

Esta forma de asignar probabilidades solo puede aplicarse cuando todos los elementos del espacio muestral tienen la misma posibilidad de ocurrir.

Ejemplo 6.

Se lanza una moneda común 3 veces consecutivas.

a) Para determinar el espacio muestral vamos a usar un diagrama de árbol en el que *A* representa águila y *S* sol.



Así que el espacio muestral es

Ω = {AAA, AAS, ASA, ASS, SAA, SAS, SSA, SSS}

El conjunto tiene 8 elementos. El número de elementos de un conjunto, se llama, cardinalidad del conjunto lo denotaremos por N(Ω) = 8

b) Vamos a considerar los siguientes eventos.
A: Se obtienen exactamente dos soles
A = {ASS, SAS, SSA}

B: Salen tres resultados iguales
B = {AAA, SSS}

C: Salen dos resultados iguales consecutivos
C = {AAA, AAS, ASS, SAA, SSA, SSS}

D: Salen al menos 2 águilas
D = {AAA, AAS, ASA,S AA}

E: No salen águilas. E = {SSS}

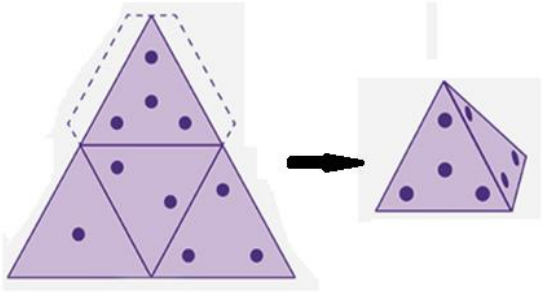
Contando los casos favorables en cada caso, tenemos:

P(A) = $\frac{3}{8}$ P(B) = $\frac{2}{8} = \frac{1}{4}$ P(C) = $\frac{6}{8} = \frac{3}{4}$

P(D) = $\frac{4}{8} = \frac{1}{2}$ P(E) = $\frac{1}{8}$

EJERCICIOS 3.3.2

1) Se construye un tetraedro con las caras numeradas como se muestra en la imagen.



Al lanzar este dado de 4 caras, se considera que el número obtenido es el que queda en la base. Se lanza dos veces el tetraedro.

a. Traza en tu cuaderno un diagrama de árbol para los dos lanzamientos y escribe el espacio muestral.
b. Calcula la probabilidad de los siguientes eventos:
A: Salen dos números iguales.
B: Salen solo números mayores que 1
C: Salen dos números que suman 6.

C. Enfoque subjetivo

Cuando se plantean preguntas como: ¿Cuál es la probabilidad de que en la segunda mitad de este siglo se encuentre una vacuna que prevenga el VIH?, los investigadores se encuentran ante situaciones en las que no es posible aplicar ninguno de los enfoques descritos anteriormente.

No se puede aplicar el enfoque frecuencial porque no hay un experimento aleatorio que repetir. No es posible aplicar el enfoque clásico porque no hay elementos que permitan asegurar que es igualmente probable que sí se encuentre a que no se encuentre.

En casos como este, solo queda acudir a los expertos. En el ejemplo, los expertos son científicos que han estudiado el VIH en todo el mundo y han desarrollado distintas formas de tratamiento. La probabilidad que ellos den del evento mencionado, se considera una aproximación de la probabilidad real. A este enfoque para asignar probabilidades se le conoce como enfoque subjetivo, y solo se aplica cuando no es posible aplicar ninguno de los enfoques anteriores.

3.4 CÁLCULO DE PROBABILIDADES DE EVENTOS SIMPLES Y COMPUESTOS

Hay distintas formas en que se puede extraer una muestra aleatoria de una población. Dos aspectos importantes a tomar en cuenta son: el orden y el reemplazo.

El orden en la extracción importa cuando se consideran muestras distintas si primero sale A y luego B que si primero sale B y luego A. No importa el orden si solo se toma en cuenta que han sido seleccionados los elementos A y B sin importar cuál se extrajo primero y cuál después.

El reemplazo ocurre cuando se selecciona un elemento, se toma nota de su valor o su característica de interés, y se regresa a la población antes de seleccionar el siguiente. En cambio, si cada objeto que se selecciona no se regresa a la población antes de seleccionar el siguiente, estamos ante una extracción sin reemplazo.

El número de posibles resultados y el número de resultados favorables, que se requieren para aplicar el enfoque clásico, cambian dependiendo de cómo sea la extracción.

A. Extracciones con orden y con reemplazo

Cuando se hacen extracciones con reemplazo, cada vez que se va a elegir un elemento, se puede elegir cualquiera de la población, así que es posible que un mismo elemento sea seleccionado dos o más veces.

Ejemplo 7.

Se van a formar números de 3 cifras eligiendo dígitos al azar y con reemplazo

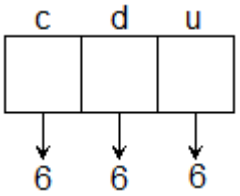
de la colección de dígitos

1, 3, 4, 6, 7, 9

Algunos ejemplos de los números de tres cifras que se pueden formar son:

169 691 961 311 444 767

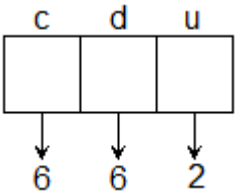
Para elegir la cifra de las unidades, hay 6 opciones. Por cada dígito elegido para esa posición, hay nuevamente 6 opciones para la cifra de las decenas. Por cada pareja escogida en decenas y unidades, hay de nuevo 6 opciones para la cifra de las centenas. De manera que el total de números de tres cifras que se puede formar es $6^3 = 216$



Vamos a calcular la probabilidad de los siguientes eventos:

E: El número formado es par.

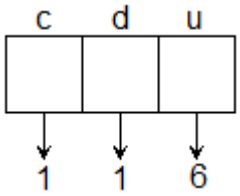
Los casos favorables son los números de 3 cifras que tiene uno de los dígitos 4 o 6 en el lugar de las unidades, es decir, para esa posición hay 2 posibilidades. La cantidad de números pares que se pueden formar es $6^2(2) = 72$



Por tanto, $P(E) = \frac{6(6)(2)}{6(6)(6)} = \frac{1}{3} = 0.333$

F: Salen tres dígitos iguales.

En los casos favorables el dígito de las unidades puede ser cualquiera de los 6 dígitos posibles, pero una vez seleccionado ese, solo ese mismo número puede salir en las decenas y en las centenas. Así que hay $1^2(6) = 6$ números formados por 3 cifras iguales.



Por tanto, $P(F) = \frac{1(1)(6)}{6(6)(6)} = \frac{1}{36} = 0.0278$

B. Extracciones con orden y sin reemplazo

Cuando se hacen extracciones sin reemplazo, los elementos que ya han sido elegidos, no pueden volver a seleccionarse porque no se regresan a la población.

Ejemplo 8.

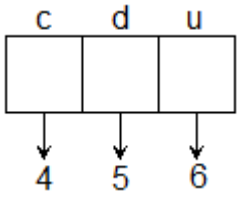
Se van a formar números de 3 cifras eligiendo tres números con orden y sin reemplazo de la colección de dígitos

1, 3, 4, 6, 7, 9

Algunos ejemplos de los números de tres cifras que se pueden formar son:

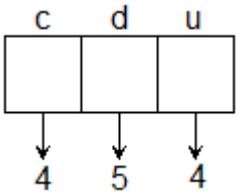
169 691 961 314 413 746

Para elegir la cifra de las unidades, hay 6 opciones. Para el dígito de las decenas, ya solo quedan 5 opciones y para el de las centenas quedan 4 opciones. De manera que el total de números de tres cifras que se puede formar es 4(5)(6) = 120



G: El número formado es impar.

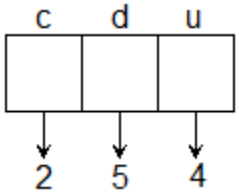
Los casos favorables son los números de 3 cifras que tiene uno de los dígitos 1, 3, 7 o 9 en el lugar de las unidades, es decir, para esa posición hay 4 posibilidades. Como el dígito que se use ahí ya no se regresa, para las decenas hay 5 posibilidades y para las centenas hay 4. La cantidad de números pares que se pueden formar es 4(5)(4) = 80



Por tanto, $P(G) = \frac{4(5)(4)}{4(5)(6)} = \frac{2}{3} = 0.667$

H: Sale un número menor que 400

En los casos favorables el dígito de las centenas puede ser 1 o 3, el de las decenas puede ser cualquiera de los 5 dígitos que no han sido usados, y el de las unidades cualquiera de los 4 dígitos no seleccionados aún. Así que hay 2(5)(4) = 40 casos favorables.



Por tanto, $P(H) = \frac{2(5)(4)}{6(5)(4)} = \frac{1}{3} = 0.333$

EJERCICIOS 3.4.1

1) Se forman placas con tres letras y 3 números. Las letras se eligen de las 26 letras del abecedario distintas de Ñ, y los números de los 10 dígitos. La elección se hace con orden y con reemplazo.

    a. Escribe 5 placas distintas que se pueden formar con las condiciones indicadas. Por ejemplo, AXN349

    b. Determina cuántas placas distintas se pueden formar

    c. Se forma una placa al azar. Calcula la probabilidad de los siguientes eventos

        A: Salen 3 letras distintas

        B: Salen 3 letras iguales

        C: La primera letra es Z

        D: Los números de la placa no empiezan con 0.

E: Salen tres números distintos.

F: El número de 3 cifras es menor que 700

C. Probabilidad de eventos compuestos

Se forman eventos compuestos cuando se relacionan dos o más eventos mediante operaciones como uniones, intersecciones, diferencias y complementos.

Ejemplo 9.

Se lanza un dado común dos veces consecutivas. El espacio muestral está formado por todas las parejas de números enteros del 1 al 6, formadas con orden y con reemplazo.

$$\Omega = \begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{Bmatrix}$$

El número de elementos del espacio muestral es N(Ω) = 36.

Consideremos los eventos:

A: Salen dos números que suman 8.

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

B: Salen dos números cuya diferencia es 4.

$$B = \{(1,5)\ \ (2,6)\ \ (5,1)\ \ (6,2)\}$$

C: Salen dos números iguales.

$$C = \{(1,1)\ \ (2,2)\ \ (3,3)\ \ (4,4)\ \ (5,5)\ \ (6,6)\}$$

Sus probabilidades son:

$$P(A) = \frac{5}{36} \qquad P(B) = \frac{4}{36} = \frac{1}{9} \qquad P(C) = \frac{6}{36} = \frac{1}{6}$$

Ahora consideremos los siguientes eventos compuestos:

A ∩ B: Salen dos números que suman 8 y su diferencia es 4.

$$A \cap B = \{(2,6)\ \ (6,2)\}$$

A $\cup$ C: Salen dos números que suman 8 o son iguales.

$$A \cup C = \begin{Bmatrix} (1,1)\ \ (2,2)\ \ (3,3)\ \ (4,4)\ \ (5,5)\ \ (6,6) \\ (2,6)\ \ (3,5)\ \ (5,3)\ \ (6,2) \end{Bmatrix}$$

C^c: Salen dos números distintos.

$$C^c = \begin{Bmatrix} (1,2)\ \ (1,3)\ \ (1,4)\ \ (1,5)\ \ (1,6) \\ \vdots \qquad\qquad \vdots \\ (6,1)\ \ (6,2)\ \ (6,3)\ \ (6,4)\ \ (6,5) \end{Bmatrix}$$

Sus probabilidades son:

$$P(A \cap B) = \frac{2}{36} = \frac{1}{18} \qquad P(A \cup C) = \frac{10}{36} = \frac{5}{18} \qquad P(C^c) = \frac{30}{36} = \frac{5}{6}$$

Ejemplo 10.

Se realiza una encuesta a 500 estudiantes del plantel para investigar sobre el género y el hábito de fumar. La información obtenida es la siguiente.

		Hábito de tabaquismo			
		Nunca ha fumado	Dejó de fumar	Fuma	Total
Género	Hombre	124	18	103	245
	Mujer	144	29	82	255
	Total	268	47	185	500

Se selecciona un estudiante del plantel al azar. Consideremos los siguientes eventos:

A: Es hombre A^c: es mujer B: Nunca ha fumado

C: Dejó de fumar D: Fuma.

Aproximemos frecuencialmente sus probabilidades. Observa que para ello, se requieren los números de los totales.

$$P(A) = \frac{245}{500} = 0.49 \qquad P(A^c) = \frac{255}{500} = 0.51 \qquad P(B) = \frac{268}{500} = 0.536$$

$$P(C) = \frac{47}{500} = 0.094 \qquad P(D) = \frac{185}{500} = 0.370$$

Con la misma información, es posible aproximar la probabilidad de los siguientes eventos compuestos.

$$P(A \cap B) = \frac{124}{500} = 0.248 \qquad P(A^c \cap D) = \frac{82}{500} = 0.164 \qquad P(A^c \cap C) = \frac{29}{500} = 0.058$$

$$P(A \cup B) = \frac{124 + 144 + 18 + 103}{500} = \frac{389}{500} = 0.778$$

$$P(A^c \cup D) = \frac{144 + 29 + 82 + 103}{500} = \frac{358}{500} = 0.716$$

$$P(A \cup C) = \frac{18 + 124 + 103 + 29}{500} = \frac{274}{500} = 0.548$$

EJERCICIOS 3.4.2

1. En el experimento de lanzar dos dados comunes, calcula la probabilidad de los siguientes eventos.

   X: Salen dos números cuya diferencia es 3 Y: Sale al menos un 2

   Z: Salen dos números menores que 4 Z^c $X \cap Y$ $X \cup Y$

2. Se lanza al aire una moneda común tres veces consecutivas. Calcula la probabilidad de los siguientes eventos.

   M: Salen exactamente 2 águilas. N: Salen al menos 2 soles.

   O: Salen a lo más 2 soles. Q: No salen soles

   O – Q $N \cap O$ Q^c $M \cup Q$

3. En una alcaldía de la CdMx, se aplicaron encuestas sobre el desempeño del alcalde en funciones, separando las respuestas por rango de edad de los encuestados. La siguiente tabla de contingencia muestra los resultados.

	Muy malo	Malo	Regular	Bueno	Muy bueno	Total
De 18 a 24 años	55	60	45	30	10	200
De 25 a 44 años	35	40	55	25	5	160
De 45 a 60 años	25	28	32	35	20	140
Total	115	128	132	90	35	500

Supongamos que se selecciona una persona al azar en esa alcaldía. Aproxima frecuencialmente las probabilidades de los siguientes eventos.

A: Tiene entre 18 y 24 años

B: Tiene entre 25 y 44 años

C: Tiene más de 44 años

D: Considera que el desempeño del alcalde ha sido muy malo

E: Considera que el desempeño del alcalde ha sido malo

F: Considera que el desempeño del alcalde ha sido regular

G: Considera que el desempeño del alcalde ha sido bueno

H: Considera que el desempeño del alcalde ha sido muy bueno

$C \cap H$ $A \cup D$ $B - F$ $D \cup E$ F^c $(G \cup H)^c$

3.5 PROBABILIDAD CONDICIONAL Y EVENTOS INDEPENDIENTES

A. Probabilidad condicional

Imagina que lanzas dos veces un dado común. Los casos posibles son las 36 parejas de números que escribimos en la sección anterior. Supongamos que estamos interesados en calcular la probabilidad del evento

A: la suma de los números obtenidos es 8.

Los casos favorables a este evento son las parejas:

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

Así que su probabilidad es $P(A) = \frac{5}{36}$.

Si se sabe que en el primer lanzamiento salió 6, entonces ya solo son posibles las parejas

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

De estas 6 parejas, la única cuyos números suman 8 es

$$\{(6,2)\}$$

Así que conocer la información de que ha ocurrido B, reduce los casos favorables a 1 y el total de casos posibles a 6. La nueva probabilidad es

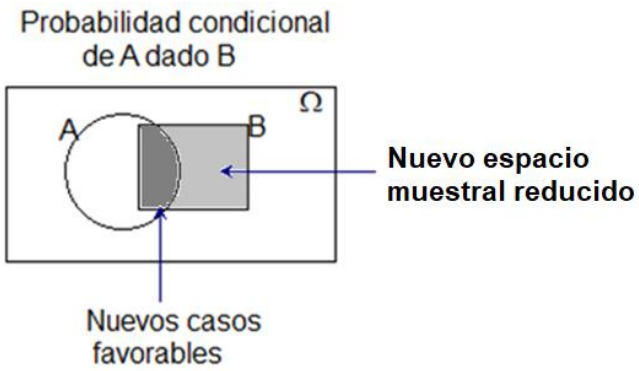
$$\mathrm{P}(A \mid B) = \frac{1}{6},$$

donde $P(A \mid B)$ se lee: "Probabilidad de A dado B" y se conoce como una probabilidad condicional.

Gráficamente, podemos representar la situación de la probabilidad condicional así:

Recuerda que el número de elementos de un conjunto B lo denotamos por N(B). La representación anterior muestra que la probabilidad condicional se puede calcular mediante la fracción $\frac{N(A \cap B)}{N(B)}$. Si dividimos tanto el numerador como el denominador de esta fracción entre el número de elementos del espacio muestral Ω, obtenemos:

$$P(A \mid B) = \frac{N(A \cap B)}{N(B)} = \frac{N(A \cap B)/N(\Omega)}{N(B)/N(\Omega)} = \frac{P(A \cap B)}{P(B)}$$



Lo anterior tiene sentido siempre y cuando $P(B) \neq 0$, ya que no está definida la división entre 0.

En resumen, para cualquier par de eventos A y B, con $P(B) \neq 0$, se define la probabilidad condicional de A dado B como:

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

Ejemplo 11.

En una secundaria, se sabe que el 32% de los alumnos reprueban Matemáticas, el 18% reprueba Español y el 10% reprueban ambas asignaturas.

Se selecciona un alumno de esa secundaria al azar. Consideremos los eventos

M: el alumno reprueba Matemáticas.

E: el alumno reprueba Español.

Entonces, la información del enunciado se resume en

$$P(M) = 0.32 \qquad P(E) = 0.18 \qquad P(M \cap E) = 0.1$$

Vamos a calcular las siguientes probabilidades condicionales.

a) Que el alumno repruebe Matemáticas dado que reprobó Español.

$$P(M \mid E) = \frac{P(M \cap E)}{P(E)} = \frac{0.1}{0.18} = 0.5556$$

b) Que el alumno repruebe Español dado que reprobó Matemáticas.

$$P(E \mid M) = \frac{P(E \cap M)}{P(M)} = \frac{0.1}{0.32} = 0.3125$$

Ejemplo 12.

En cierto poblado, las mujeres representan el 52% de la población y los hombres el otro 48%. Se sabe que el 20% de las mujeres y el 5% de hombres están sin trabajo. Un economista que estudia la situación de empleo, elige al azar una persona.

Consideremos los siguientes eventos sobre la persona seleccionada:

M: Es mujer H: Es hombre E: Está empleada

D: Está desempleada

Supongamos que la población total es de 8000 personas. Queremos determinar las siguientes probabilidades condicionales sobre la persona seleccionada.

- F: Que sea mujer dado que está desempleada
- G: Que sea hombre dado que está empleado
- H: Que está desempleado dado que es hombre
- I: Que esté empleada dado que es mujer

Es útil construir una tabla de contingencia con la información que se tiene sobre el espacio muestral. Para ello, calculamos:

52% de 8000 = 4160 48% de 8000 = 3840

20% de 4160 = 832 5% de 3840 = 192

	Desempleados	Empleados	Total
Mujeres	832	3 328	4 160
Hombres	192	3 648	3 840
Total	1 024	6 976	8 000

Cada una de las entradas de la tabla representan la cantidad de elementos de los siguientes eventos

	Desempleados	Empleados	Total
Mujeres	N(M ∩ D)	N(M ∩ E)	N(M)
Hombres	N(H ∩ D)	N(H ∩ E)	N(H)
Total	N(D)	N(E)	N(Ω)

Así que

$$P(F) = P(M \mid D) = \frac{832}{1024} = 0.8125$$

$$P(G) = P(H \mid E) = \frac{3648}{6976} = 0.523$$

$$P(H) = P(D \mid H) = \frac{192}{3840} = 0.05$$

$$P(I) = P(E \mid M) = \frac{3328}{4160} = 0.80$$

B. Probabilidad de la intersección de dos eventos

En algunos casos, es fácil calcular directamente la probabilidad condicional y la fórmula que hemos visto se usa más bien para determinar la probabilidad de una intersección. Para ello, simplemente debemos despejar la probabilidad de la intersección.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \rightarrow \qquad P(A \cap B) = P(B)P(A \mid B)$$

Ejemplo 13.

De una urna que contiene 6 bolas negras y 4 blancas, se hacen dos extracciones **sin reemplazo**. Vamos a calcular la probabilidad de que las dos bolas extraídas sean blancas.

Usaremos los siguientes eventos.

B_1: La primera bola seleccionada es blanca

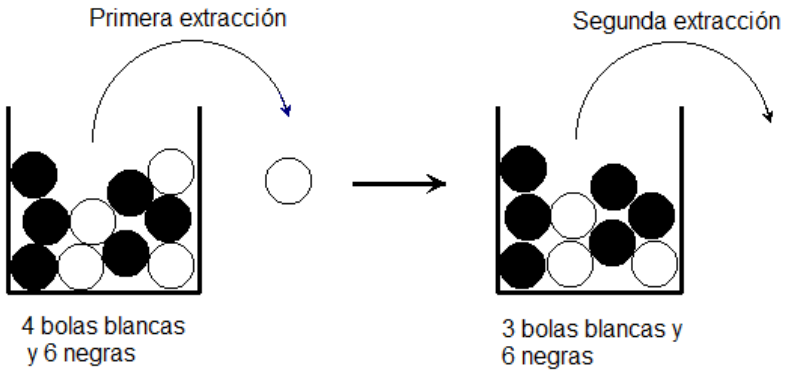
B_2: La segunda bola seleccionada es blanca

Así que buscamos $P(B_1 \cap B_2)$. En la primera extracción se tiene que

$$P(B_1) = \frac{4}{10}$$

Como la extracción es sin reemplazo, la composición de la urna cambia para la segunda extracción, conteniendo ahora 6 bolas negras y 3 blancas. Por esta razón, en la segunda extracción se tiene que

$$P(B_2 \mid B_1) = \frac{3}{9}$$



Por lo tanto

$$P(B_1 \cap B_2) = P(B_1)P(B_2 \mid B_1) = \frac{4}{10}\left(\frac{3}{9}\right) = \frac{12}{90} = 0.133$$

C. Independencia de dos eventos

Dos eventos A y B son independientes si la ocurrencia o no ocurrencia de uno de ellos, no altera la probabilidad del otro, es decir, si P(A | B) = P(A).

Como $P(A \cap B) = P(B)P(A \mid B)$, lo anterior conduce a la afirmación:

$$A \text{ y } B \text{ son independientes si y solo si } P(A \cap B) = P(A)P(B).$$

Ejemplo 14.

De una urna que contiene 6 bolas negras y 4 blancas, se hacen dos extracciones **con reemplazo**. Vamos a analizar los eventos:

B_1: La primera bola seleccionada es blanca

N_2: La segunda bola seleccionada es negra

Como la extracción es con reemplazo, la composición de la urna no cambia después para la segunda extracción. De manera que la probabilidad $P(N_2)$ no se ve afectada

por el color que haya tenido la primera bola seleccionada, es decir, se trata de eventos independientes. De ahí que:

$$P(B_1 \cap N_2) = P(B_1)P(N_2) = \frac{4}{10}\left(\frac{6}{10}\right) = \frac{24}{100} = 0.24$$

Ejemplo 15.

Para estudiar la relación entre ser zurdo o diestro con el género de una persona, se entrevistaron 350 personas elegidas al azar y se obtuvo la siguiente información.

	Diestro	Zurdo	Total
Hombre	156	26	182
Mujer	144	24	168
Total	300	50	350

a) Supongamos que se elige una persona al azar de la población entrevistada. Vamos a aproximar frecuencialmente las probabilidades de los siguientes eventos.

D: Es diestro $P(D) = \frac{300}{350}$

Z: Es zurdo $P(Z) = \frac{50}{350}$

H: Es hombre $P(H) = \frac{182}{350}$

M: Es mujer $P(M) = \frac{168}{350}$

b) Verifiquemos si son independientes o no las siguientes parejas de eventos:

- D y H

$P(D \cap H) = \frac{156}{350} = 0.4457 \quad y \quad P(D)P(H) = \frac{300(182)}{350^2} = 0.4457$

Así que D y H sí son independientes.

- Z y M

$$P(Z \cap M) = \frac{24}{350} = 0.06857 \quad y \quad P(Z)P(M) = \frac{50(168)}{350^2} = 0.06857$$

De donde se concluye que Z y M también son independientes.

De hecho, si A y B son independientes, entonces también lo son las siguientes parejas de eventos:

A y B^c A^c y B A^c y B^c

EJERCICIOS 3.5.1

1. Se lanza dos veces un dado común. Calcula la probabilidad de que

   a. Haya salido un 3 si se sabe que la suma de los dos lanzamientos fue 6.

   b. La suma de los dos lanzamientos sea 6, si se sabe que uno de los dados cayó en 3.

2. Una urna contiene 8 bolas rojas y 4 azules. Se extraen dos bolas **sin reemplazo**. Calcula la probabilidad de que

   a. Ambas sean rojas.

   b. Ambas sean azules

3. Se tienen dos urnas: la urna 1 tiene 7 bolas verdes y 5 amarillas, la urna 2 contiene 4 bolas verdes y 6 amarillas. Se extrae una bola de cada urna al azar. Considera los eventos:

   A_1: La bola extraída de la urna 1 es amarilla

   A_2: La bola extraída de la urna 2 es amarilla

   a. ¿Son independientes los eventos anteriores? _____ Explica tu respuesta. _____

   b. Calcula la probabilidad de que las dos bolas extraídas de las urnas sean amarillas.

ANEXO DE RESPUESTAS

Ejercicios 1.1.1

1. a. Alumnos de escuelas primarias y secundarias del estado de Chihuahua
   b. Alumnos de 300 escuelas primarias y secundarias seleccionadas al azar
   c. Una encuesta (levantamiento).
2. 2.1 – b 2.2 – a 2.3 – b 2.4 – a 2.5 – d
3. a. Todas las mujeres de 15 años y más que asisten a la clínica 22 del IMSS.
   b. El conjunto de estudiantes matriculados este semestre en el CCH Sur.
   c. Mexicanos con credencial del INE vigente con teléfono fijo.

Ejercicios 1.2.1

Variable	Valores	Tipo de variables
a) Opinión sobre un maestro del CCH.	Bueno, malo, regular	Cualitativa ordinal
b) Cantidad de café que sirve una máquina automática en una descarga.	De 250 a 300 mililitros	Cuantitativa continua
c) Cantidad de libros que un estudiante consulta en la biblioteca en un semestre.	0, 1, 2, … 10	Cuantitativa discreta
d) Carreras que eligen estudiantes de 6° semestre.	Administración, Actuaría, Contaduría, etcétera	Cualitativa nominal
e) Peso del contenido de las cajas de cereal que indican 800 gr.	De 750 a 850 gr	Cuantitativa continua
f) Tipo de medalla obtenida por los tres mejores deportistas de una prueba.	Oro, plata y bronce	Cualitativa ordinal

Ejercicios 1.3.1

Cantidad a pagar x_i	Frecuencia f_i	Frecuencia Relativa f_r	Frecuencia Acumulada (F_a)	Frecuencia relativa acumulada (F_{ra})
70	**1**	0.033	1	**0.033**
230	3	0.100	**4**	0.133
340	5	0.167	9	**0.300**
370	2	0.067	**11**	**0.367**
450	5	0.167	**16**	**0.533**
485	**3**	0.100	**19**	**0.633**
560	**5**	**0.167**	24	**0.800**
870	4	0.133	**28**	**0.933**
970	1	**0.033**	**29**	**0.967**
1120	**1**	0.033	**30**	1.000
Total	30	**1**		

b. Las respuestas pueden variar, por ejemplo:

1. Los montos más frecuentes de pago son: \$340, \$450 y \$560. Estos 3 montos abarcan al 17 × 3 = 51% de las familias.
2. De las 30 familias, 16 pagan a lo más \$450
3. El 94% de las familias pagan cuando mucho \$870
4. Los montos menos frecuentes están en los extremos, es decir, muy poco (\$70) o mucho (\$970 y \$1120)

Ejercicios 1.3.2

1. Fórmula de Sturges para el número de intervalos

$$k = 1 + 3.322Log(n) = 1 + 3.322Log(60) = 1 + 3.322(1.78) = 1 + 5.91 = 6.91$$

Tomamos $k = 7$ redondeado. Entonces, la amplitud sugerida es

$$c = \frac{Rango}{K} = \frac{73 - 29}{7} = \frac{44}{7} = 6.29$$

Tomamos *c = 6.* Como no se cubre el rango con 7 intervalos de longitud 6, usaremos 8 intervalos. Distribuyendo los valores que sobran, se obtienen los intervalos:

$[27 - 33)$
$[33 - 39)$
$[39 - 45)$
$[45 - 51)$
$[51 - 57)$
$[57 - 63)$
$[63 - 69)$
$[69 - 75]$

Contando cuantos datos caen en cada intervalo y calculando las demás frecuencias, se obtiene

Intervalo o clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
$[27 - 33)$	2	0.033	2	0.033
$[33 - 39)$	5	0.083	7	0.117
$[39 - 45)$	6	0.100	13	0.217
$[45 - 51)$	12	0.200	25	0.417
$[51 - 57)$	8	0.133	33	0.5500
$[57 - 63)$	12	0.200	45	0.750
$[63 - 69)$	8	0.133	53	0.883
$[69 - 75]$	7	0.117	60	1
Total	60	1.000		

2. Las respuestas pueden variar, por ejemplo:
    a. De los 60 días, en 25 hubo menos de 51 vuelos.
    b. Las cantidades más frecuentes de vuelos, están entre 45 y 51, y entre 57 y 63. Los dos intervalos anteriores abarcan el 40% de los días.
    c. Alrededor de 20% de los días hubo menos de 45 vuelos y alrededor de 88% de los días hubo menos de 69 vuelos.
    d. Solo el 25% de los días hubo de 63 a 75 vuelos.

3. Fórmula de Sturges para el número de intervalos

$$k = 1 + 3.322Log(n) = 1 + 3.322Log(40) = 1 + 3.322(1.60) = 1 + 5.32 = 6.32$$

Tomamos $k = 7$. Entonces, la amplitud sugerida es

$$c = \frac{Rango}{K} = \frac{16.5 - 3.8}{7} = \frac{12.7}{7} = 1.8143$$

Tomaremos intervalos de longitud 1.9 Como 7 intervalos de longitud 1.9 abarcan 13.3 segundos y el rango es de 12.7 segundos, distribuimos el sobrante y los intervalos quedan como en la siguiente tabla:

Y calculando las frecuencias se obtiene:

Intervalo o clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
$[3.5 - 5.4)$	1	0.025	1	0.025
$[5.4 - 7.3)$	2	0.050	3	0.075
$[7.3 - 9.2)$	9	0.225	12	0.300
$[9.2 - 11.1)$	9	0.225	21	0.525
$[11.1 - 13.0)$	14	0.350	35	0.875
$[13.0 - 14.9)$	3	0.075	38	0.950
$[14.9 - 16.8]$	2	0.050	40	1
Total	40	1.000		

4. Las respuestas pueden variar, por ejemplo:
    a. Las mayores frecuencias están entre los 7.3 y los 13 segundos. En este rango se ubica el 80% de los casos registrados en los datos.
    b. Más de 87% de los casos registrado, presentaron una reacción antes de los 13 segundos.
    c. Sólo el 30% de las personas, presentaron reacción antes de los 7.3 segundos.
    d. Solo el 5% presentó una reacción de los 14.9 segundos en adelante.

5. Tabla completa

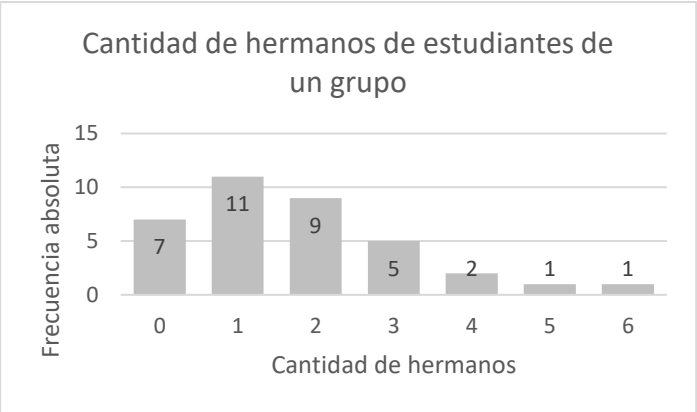
Intervalo de Clase (Ganancia Semanal)	Frecuencia Simple	Frecuencia Relativa	Frecuencia Relativa	Frecuencia Acumulada Relativa
$[300-400)$	105	0.350	105	0.350
$[400-500)$	42	0.140	147	0.490
$[500-600)$	90	0.300	237	0.790
$[600-700)$	45	0.150	282	0.940
$[700-800]$	18	0.060	300	1

a. La frecuencia simple del primer intervalo nos dice que: 105 estudiantes ganan menos de 400, pero al menos 300.
b. El 30% de los estudiantes ganan entre 500 y 600.

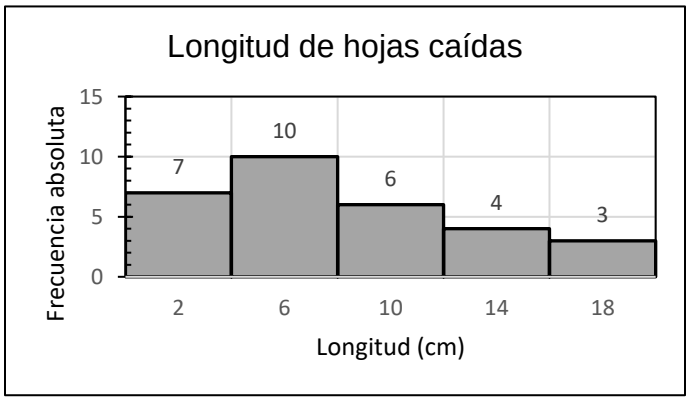
c.- La frecuencia acumulada de la cuarta clase quiere decir que: 282 estudiantes ganan menos de 700, pero al menos .
d.- El porcentaje de estudiantes que ganan máximo $599.99 es de 79%.

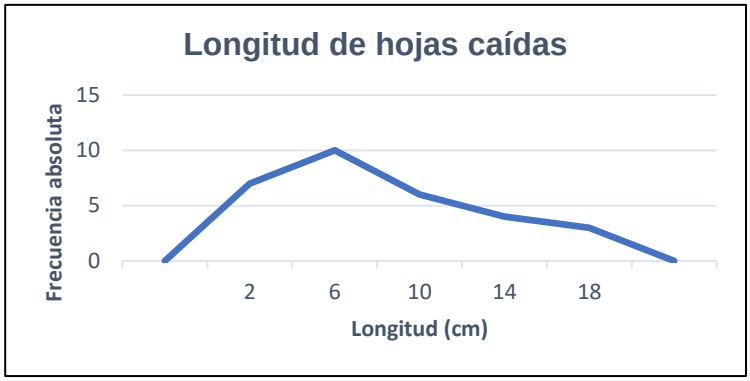
Ejercicios 1.4.1

1. a.

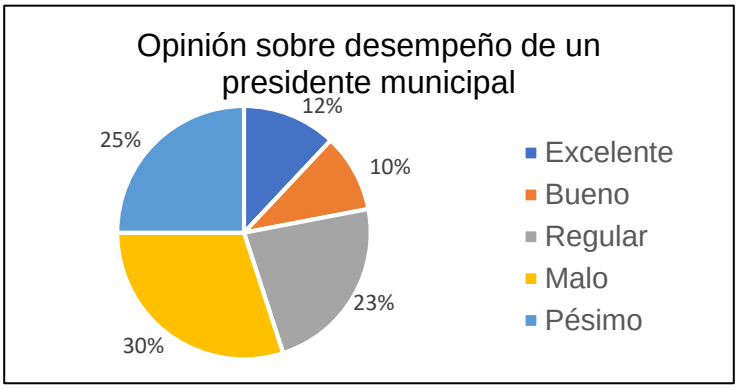


b.

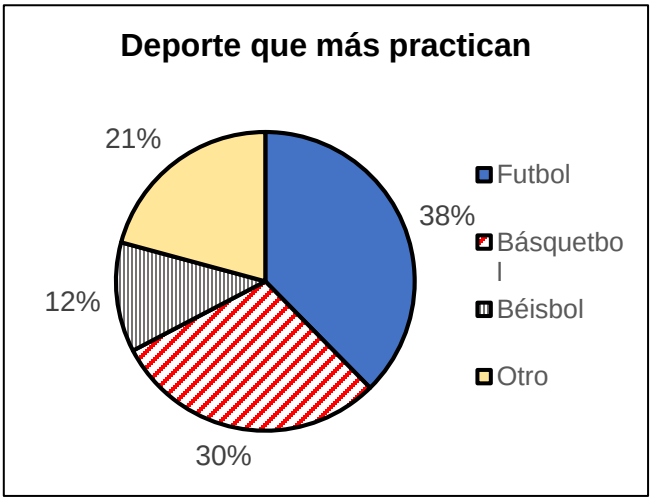


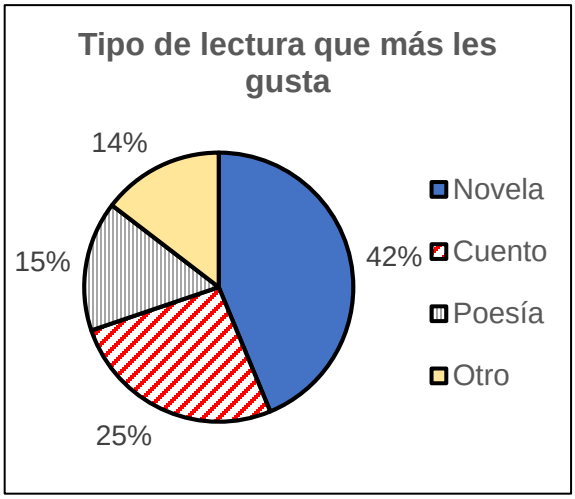

Longitud de hojas caídas
Frecuencia absoluta
15
10
5
0
2
6
10
14
18
Longitud (cm)

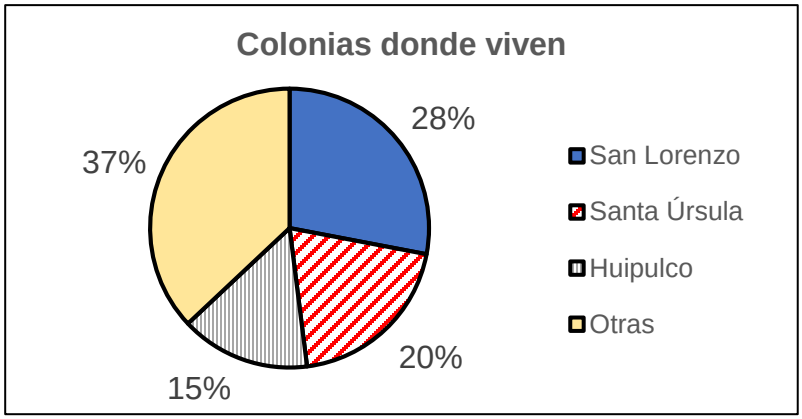
c.


Opinión sobre desempeño de un presidente municipal
12%
10%
23%
30%
25%
Excelente
Bueno
Regular
Malo
Pésimo

2.


Deporte que más practican
21%
38%
12%
30%
Futbol
Básquetbol
Béisbol
Otro


Tipo de lectura que más les gusta
14%
15%
42%
25%
Novela
Cuento
Poesía
Otro



Ejercicios 1.5.1

1. a) Mediana: Mdn = 27.5 mm. b) Moda: Mo = 18 años. c) Media: $\overline{X} = 28.6$ min.
2. a) Media: $\overline{X} = 0.1172$ km, Mediana Mdn = 0.115 km, Modas: 0.08 y 0.12 km.
   a) Media: $\overline{X} = 1.91$ hijos, Mediana Mdn = 2 hijos, Moda: 0 hijos.
3. a) Entre los 16 y los 30 años. b) Entre los 30 y los 65 años. c) 28 años. d) 36,4 años.
4. Hay muchas formas de responder. Un ejemplo:
   f. 196, 197, 198, 199, 200, 200, 201, 202, 203, 204

Ejercicios 1.5.2

1. Colección A: rango = 9, desviación típica S = 3.5777
   Colección B: rango = 9, Desviación típica S = 4.5935
   Más dispersa la colección B. Aunque tienen rango igual, la desviación típica es mayor en la colección B porque los datos están más alejados de su media.
2. Desviación típica entre 2 y 4 unidades: colección 3. Desviación típica menor a 2 unidades: colección 1. Desviación típica mayor a 4 unidades: Colección 2.
3. En la empresa B, pues la mayoría de ellos estarán entre $38 000 y $42 000, y tienen un poder de compra similar.
   CV_A = 44.44%, CV_B = 5%. Interpretación: la dispersión es mucho mayor en la empresa A porque la desviación típica es el 44.44% de la media, mientras que en la empresa B la desviación típica es el 5% de la media.

Ejercicios 1.5.3

1. Q_1 = 68.5, Q_2 = Mdn = 71, Q_3 = 75
2. a. La empresa A porque ¾ partes de los salarios están entre $150 y $300 diarios. b. Empresa A: entre $150 y $225, Empresa B: entre $150 y $300.
   c. Empresa A: entre $300 y $650, Empresa B: entre $400 y $600.
   d. Empresa A: la segunda, de $200 a $225, Empresa B: la primera y la segunda están igualmente concentradas.

Ejercicios 2.1.1

1. a) Tabla de contingencia

	Periódico preferido				
Estado Civil	El Universal	Excélsior	Reforma	La Jornada	Total
Soltero	11	9	10	14	44
Casado	10	6	10	8	34
Viudo	6	4	5	5	20
Separado	10	8	5	9	32
Total	37	27	30	36	130

Tabla de frecuencias relativas

	Periódico preferido				
Estado Civil	El Universal	Excélsior	Reforma	La Jornada	Total
Soltero	8.46%	6.92%	7.69%	10.77%	33.85%
Casado	7.69%	4.62%	7.69%	6.15%	26.15%
Viudo	4.62%	3.08%	3.85%	3.85%	15.38%
Separado	7.69%	6.15%	3.85%	6.92%	24.62%
Total	28.46%	20.77%	23.08%	27.69%	100%

Tabla de porcentajes por renglón

	Periódico preferido				
Estado Civil	El Universal	Excélsior	Reforma	La Jornada	Total
Soltero	25.00%	20.45%	22.73%	31.82%	100%
Casado	29.41%	17.65%	29.41%	23.53%	100%
Viudo	30.00%	20.00%	25.00%	25.00%	100%
Separado	31.25%	25.00%	15.63%	28.13%	100%
Total	28.46%	20.77%	23.08%	27.69%	100%

Tabla de porcentajes por columna

	Periódico preferido				
Estado Civil	El Universal	Excélsior	Reforma	La Jornada	Total
Soltero	29.73%	33.33%	33.33%	38.89%	33.85%
Casado	27.03%	22.22%	33.33%	22.22%	26.15%
Viudo	16.22%	14.81%	16.67%	13.89%	15.38%
Separado	27.03%	29.63%	16.67%	25.00%	24.62%
Total	100%	100%	100%	100%	100%

b.

- 27 personas prefieren leer el Excélsior.
- Se entrevistó a 20 personas viudas.
- ¿Cuántas personas son solteras y prefieren el periódico la Jornada? 14

- ¿Qué porcentaje de personas son casadas y prefieren el periódico Reforma? 7.69%
- ¿Qué porcentaje de personas no son casadas? 73.85%
- ¿Qué porcentaje de personas no leen Excelsior? 79.23%
- De las personas separadas, el 28.13% prefiere leer la Jornada
- De las personas viudas, ¿qué porcentaje prefiere leer el Reforma? 16.67%
- ¿Qué estado civil tiene el mayor porcentaje de lectores de La Jornada? Personas solteras
- De las personas que prefieren el Reforma, el 16.67% son separadas
- De las personas que prefieren el Universal, ¿qué porcentaje son solteros? 29.73%
- ¿Cuál de los periódicos tiene entre sus lectores el mayor porcentaje de casados? El Reforma

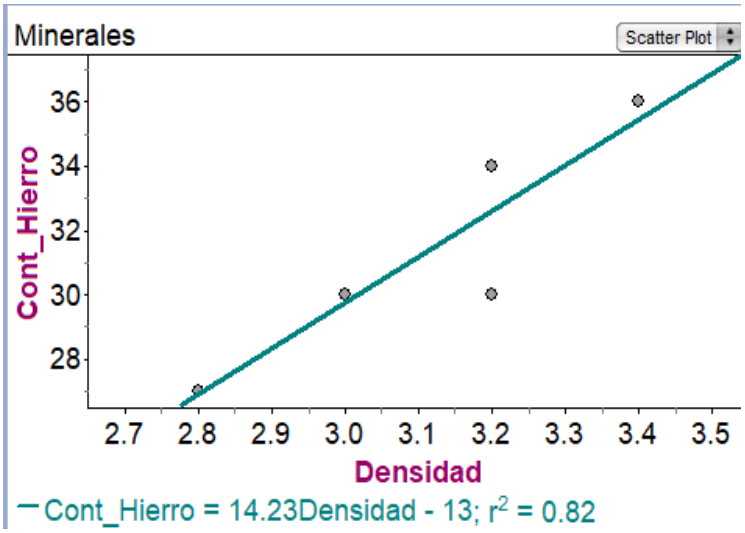
c. Las variables no están relacionadas.

Ejercicios 2.3.1

1. a) $y = -1.41x + 53.8$
   b) *r = -0.95*
   c) Muy fuerte correlación lineal negativa, es decir, mientras mayor sea el número de piezas, menor es el costo de cada una.
   d) Sí
   e) $25.6 por pieza

2.

a)



b) $y = 14.23x - 13$

c) $\boldsymbol{r}$ = 0.9055 Hay una muy fuerte correlación lineal positiva.

d) Sí.

e) El contenido de hierro estimado para una densidad de 2.9 es 28.267

f) Si el contenido de hierro fuera 31, la densidad del mineral sería de 3.09

3. a) $y = 5x + 4.82$

b) $y = 4.82(2) + 5 = 14.64$

Ejercicio 3.1.1

a. Determinista. Cada vez que se repite se obtiene color rojo.
b. Aleatorio. Antes de realizarlo no se sabe si saldrá rojo o negro.

Ejercicios 3.3.1

1. De las 50 repeticiones, en 30 salió roja y en 20 negra. Por tanto la frecuencia relativa de cada color es: roja: 0.6, negra: 0.4.
   Asumiendo que esta es una buena aproximación de la probabilidad, es de esperarse que el número de bolas rojas sea aproximadamente 0.6(7) = 4.2 y las bolas negras sean aproximadamente 0.4(7) = 2.8. Conclusión: 4 rojas y 3 negras.
2. a.

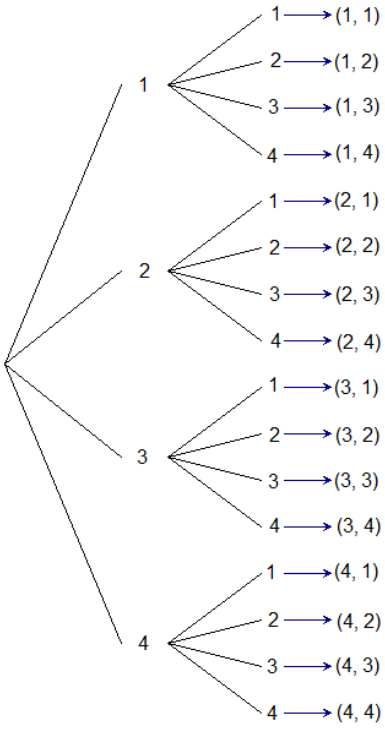
	Le gusta más Física	Le gusta más Historia	Le gusta más el Arte	Total
Hombre	102	88	63	253
Mujer	76	104	97	277
Total	178	192	160	530

b. $P(A) \approx 0.3358$ $P(B) \approx 0.3623$ $P(D) \approx 0.5226$

$P(D^c) \approx 0.4773$ $P(C) \approx 0.3019$

Ejercicios 3.3.2

a.



Espacio muestral

$$\Omega = \begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) \\ (2,1) & (2,2) & (2,3) & (2,4) \\ (3,1) & (3,2) & (3,3) & (3,4) \\ (4,1) & (4,2) & (4,3) & (4,4) \end{Bmatrix}$$

b.

$$P(A) = \frac{4}{16} = \frac{1}{4}$$

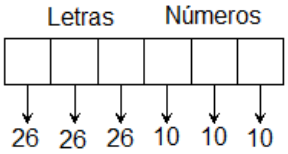
$$P(B) = \frac{9}{16}$$

$$P(A) = \frac{3}{16}$$

Ejercicios 3.4.1

a. La respuesta puede variar, ejemplo: AYY755

b.



Total de placas distintas:

$$26^3(10^3) = 17576000$$

c. $P(A) = \frac{26(25)(24)(10^3)}{26^3(10^3)} = 0.8876$ $P(B) = \frac{26(1)(1)(10^3)}{26^3(10^3)} = 0.0015$

. $P(C) = \frac{1(26)(26)(10^3)}{26^3(10^3)} = 0.0385$ $P(D) = \frac{26^3(9)(10^2)}{26^3(10^3)} = 0.9$

. $P(E) = \frac{26^3(10)(9)(8)}{26^3(10^3)} = 0.72$ $P(F) = \frac{26^3(7)(10^2)}{26^3(10^3)} = 0.7$

Ejercicios 3.4.2

1. $P(X) = \frac{6}{36} = \frac{1}{6}$ $P(Y) = \frac{11}{36}$ $P(Z) = \frac{9}{36} = \frac{1}{4}$ $P(Z^c) = \frac{25}{36}$

. $P(X \cap Y) = \frac{2}{36} = \frac{1}{18}$ $P(X \cup Y) = \frac{15}{36}$ $P(Z - Y) = \frac{4}{36} = \frac{1}{9}$

2. $P(M) = \frac{3}{8}$ $P(N) = \frac{4}{8} = \frac{1}{2}$ $P(O) = \frac{7}{8}$ $P(Q) = \frac{1}{8}$

. $P(O - Q) = \frac{6}{8} = \frac{3}{4}$ $P(N \cap O) = \frac{3}{8}$ $P(Q^c) = \frac{7}{8}$ $P(M \cup Q) = \frac{4}{8} = \frac{1}{2}$

3. $P(A) = \frac{200}{500} = 0.4$ $P(B) = \frac{160}{500} = 0.32$ $P(C) = \frac{140}{500} = 0.28$

. $P(D) = \frac{115}{500} = 0.23$ $P(E) = \frac{128}{500} = 0.256$ $P(F) = \frac{132}{500} = 0.264$

. $P(G) = \frac{90}{500} = 0.18$ $P(H) = \frac{35}{500} = 0.07$ $P(C \cap H) = \frac{20}{500} = 0.04$

. $P(A \cup D) = \frac{260}{500} = 0.52$ $P(B - F) = \frac{105}{500} = 0.21$ $P(D \cup E) = \frac{360}{500} = 0.72$

. $P(F^c) = \frac{368}{500} = 0.736$ $P((G \cup H)^c) = \frac{375}{500} = 0.75$

Ejercicios 3.5.1

1. Sean los eventos S: suma 6, y T: en uno de los dados salió 3.

$$a. \quad P(T \mid S) = \frac{1}{5} \qquad b. \quad P(S \mid T) = \frac{1}{11}$$

2. Eventos:

A_1: La primera bola extraída es azul
A_2: La segunda bola extraída es azul
R_1: La primera bola extraída es roja
R_2: La segunda bola extraída es roja

a. $P(R_1 \cap R_2) = \frac{8}{12}\left(\frac{7}{11}\right) = \frac{14}{33}$

b. $P(A_1 \cap A_2) = \frac{4}{12}\left(\frac{3}{11}\right) = \frac{1}{11}$

3. a. Sí son independientes porque la extracción de la urna 1 no afecta la probabilidad de que se extraiga una amarilla de la urna 2.

   $b.$ $P(A_1 \cap A_2) = P(A_1)P(A_2) = \frac{5}{12}\left(\frac{6}{10}\right) = \frac{1}{4}$

BIBLIOGRAFÍA

A. Bibliografía básica

1. Behar G., Roberto et al. 2004. 55 Respuestas a dudas típicas de Estadística. España. Ediciones Díaz de Santos S.A.
2. Chao, L. (1989). *Introducción a la estadística* (2ª edición). D.F., México. CECSA.
3. Domínguez y Domínguez, J. 2009. Estadística y probabilidad. El mundo de los datos y el azar. México. Oxford University Press.
4. Haber, A. et al. 1986. Estadística General. México. Addison-Wesley Iberoamericana.
5. Infante, S. 2012. Métodos Estadísticos, un Enfoque Interdisciplinario. España, Ed. Mundi-Prensa.
6. Sánchez O. 2010. Probabilidad y Estadística. México. Ed. McGraw-Hill.
7. Willoughby, S. (2000). *Probabilidad y estadística*. D.F., México. Publicaciones Cultural.

B. Bibliografía complementaria

1. Bonet, J. (2003). *Lecciones de estadística. Estadística descriptiva y probabilidad.* Alicante, España. Editorial Club Universitario.
2. Christensen, H. (1997). *Estadística paso a paso* (3ª edición). D.F., México. Trillas.
3. Cristófoli, M. 2010. Manual de Estadística con Excel. Argentina. Omicron Editorial.
4. Chung, K. (1974). *Elementary Probability Theory with Stochastic Processes.* New York, USA. Springer-Verlag,
5. Daniel, W. (2007). *Bioestadística. Base para el análisis de las ciencias de la salud*. DF, México. Limusa.
6. Guisande, G., Cástor et al. 2012. Tratamiento de datos con R, Statistica y SPSS. España. Ediciones Díaz de Santos S.A.
7. Montgomery, D. et al. 2004. Probabilidad y estadística aplicadas a la ingeniería. México. Ed. McGraw-Hill
8. Prieto, L., Herranz, I. (2005) ¿Qué significa "*Estadísticamente significativo"? Madrid, España.* Díaz de Santos.
9. Triola, M. (2009). *Estadística* (10ª edición). D.F., México. Pearson Addison Wesley.
10. Willoughby, S. (2000). *Probabilidad y estadística*. D.F., México. Publicaciones Cultural.