UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

COLEGIO DE CIENCIAS Y HUMANIDADES
PLANTEL SUR

ACADEMIA DE MATEMÁTICAS

ESTADÍSTICA I
GUÍA PARA EXAMEN EXTRAORDINARIO

Candanosa Aranda, Carlos
Guillén Anguiano, Javier
Lara Álvarez, Alicia
León Cano, María Eugenia
Romero Miranda, Lourdes

Octubre de 2008

PROGRAMA DE ESTADÍSTICA Y PROBABILIDAD I

La Estadística y la Probabilidad se han vuelto requisito indispensable en la vida cotidiana para interpretar una gran variedad de información en diversos campos de estudio. En su entorno una persona encuentra reportes financieros, económicos, médicos y otros que se pueden entender y evaluar con una comprensión básica de estas disciplinas.

El curso de Estadística y Probabilidad que se imparte en quinto semestre se concibe para proporcionar a los estudiantes los elementos básicos que le permitan comprender y aplicar los procesos descriptivos para organizar, analizar e interpretar el comportamiento de datos pertenecientes a diversos campos de estudio.

PRÓPOSITOS PARTICULARES

Al finalizar el trabajo recomendado en esta guía, el alumno:

- Se apropiará de una visión de la Estadística y de su aplicación para describir el comportamiento de un conjunto de datos en una y dos variables.

- Adquirirá los elementos, métodos y técnicas para estudiar los fenómenos de naturaleza aleatoria con el fin de comprender sus características, obtener información sobre su comportamiento y evaluar sus resultados.

BIBLIOGRAFÍA RECOMENDADA

Chao, L., Introducción a la Estadística. CECSA, 1987
Christensen, H. Estadística paso a paso. Trillas, 1997
Daniel, W. Estadística Aplicada a las Ciencias Sociales y a la Educación. Mc Graw Hill, 1998
Hoel, P., Estadística Elemental. CECSA, 1979
Johnson, R. Estadística Elemental. Iberoamérica, 1990
Mendenhall, W. Estadística para Administración y Economía. Iberoamérica, 1978
Willowghby, S. Probabilidad y Estadística. PCSA, 1993
Wonnacott, T. Fundamentos de Estadística para Administración y Economía. Limusa, 1989
Spiegel, M. Probabilidad y Estadística. Mc Graw Hill, 1975

CONTENIDO

INTRODUCCIÓN

PROPÓSITO
Que el estudiante se apropie de una visión inicial de la estadística y la probabilidad a partir de los conceptos básicos y el planteamiento de ejemplos y problemas de su entorno para apreciar los alcances de la disciplina.

Noción y utilidad de 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. Los términos "estadísticas de empleo" o "estadísticas de fútbol", son muy comunes en la información escrita y hablada. 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, etc., los cuales presentan una gran variación.

La investigación científica y la toma de decisiones en la vida diaria se enfrenta a esta presencia de la variación, de modo que para realizarlas de manera óptima, la información que se colecta debe ser de tal manera que refleje la realidad; que se obtenga con objetivos definidos; que se resuma eficientemente, y se interprete adecuadamente; y esto se logra cuando se aplica la Estadística. De manera general, podemos decir que la razón principal del uso de la estadística es la existencia de la variación en estos fenómenos.

Consulta en tres fuentes distintas la definición de estadística

1.- _____

_____

2.- _____

_____

3.- _____

_____

Como puedes observar de todo lo anterior, *la Estadística es la ciencia que se encarga del desarrollo de teoría y la aplicación de métodos de recopilación, descripción y análisis de datos, para la toma de decisiones frente a la incertidumbre.*

Importancia de la estadística para los estudiantes

1. Todo ciudadano está en continuo contacto con las estadísticas en todos los medios de comunicación. Debe poder comprender la información que se le ofrece para detectar verdades y mentiras y tomar decisiones informadas.

2. Como lector de artículos de investigación debe poder comprender la información cuantitativa que se le ofrece en los artículos que lee.

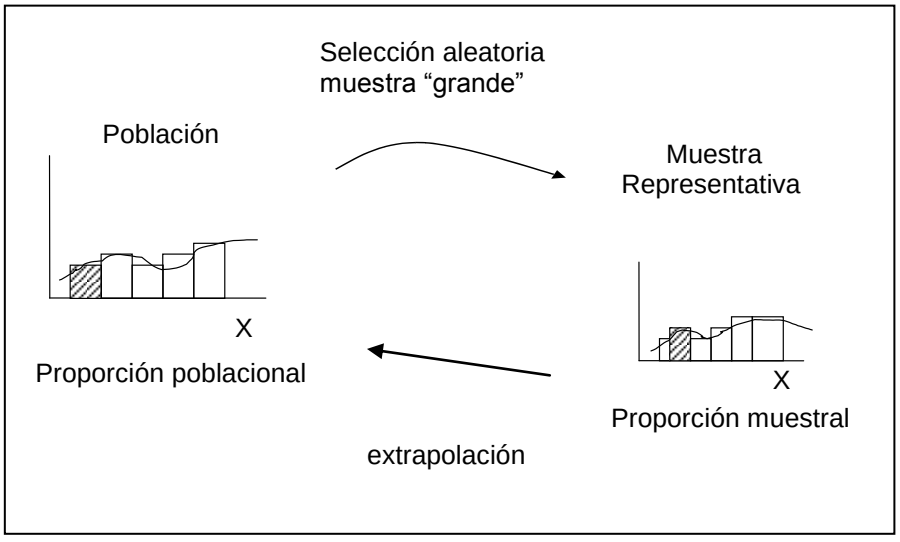
3. Como productor de investigaciones, debe poder utilizar la estadística en sus propias investigaciones, para el análisis e interpretación de resultados y la presentación de conclusiones, por ejemplo, y como justificación para la toma de decisiones.

La Estadística generalmente se divide para su estudio, en:

Estadística descriptiva
En ella se enfatizan los aspectos de presentación y descripción de los datos recogidos en la investigación. El objetivo de la estadística descriptiva es la organización de los datos para obtener información de ellos que no es obtenible a simple vista

Estadística Inferencial
Con base en la información obtenida de una pequeña parte o muestra, se hacen estimaciones y predicciones de una o varias características de la población y se realiza una toma decisiones.

Como el azar afecta tanto a la recolección de datos como a su análisis, debe ser tomado en cuenta al hacer inferencias, y es aquí donde la estadística se relaciona con la probabilidad, la cual puede definirse como el estudio matemático del azar y los fenómenos aleatorios.



Uso indebido y errores en el uso de la Estadística.

Es importante señalar que si la estadística no se utiliza adecuadamente, se puede distorsionar la información y/o tomar decisiones equivocadas.

Un error frecuente es tomar una muestra de una población bajo criterios personales del investigador o sin planificación rigurosa. También puede darse un uso indebido al manipular los resultados de algún estudio, por ejemplo para inducir respuestas a usuarios o comprometer sus decisiones.

Un poco de Historia

La palabra estadística proviene del vocablo estado, debido a que los gobiernos fueron los que comenzaron a llevar registros sobre impuestos, habitantes, nacimientos y defunciones, cosechas y datos astronómicos, etc.

La Estadística Descriptiva se origina con la recolección de datos poblacionales para censos. Estos censos ya se hacían en el imperio romano: El evangelio de Lucas dice: “Y aconteció en aquellos días que salió un edicto de parte de César Augusto, mandando que todo el mundo fuera empadronado”.

La Estadística Inferencial se origina en el Renacimiento con el desarrollo de la Teoría de la probabilidad, que a su vez se basa en el estudio de los juegos de azar. Comienza a desarrollarse plenamente con Karl Pearson y Ronald Fisher a principios del siglo XX.

Conceptos Básicos

Población
Se define como el *conjunto completo de individuos* (personas, animales o cosas) que tienen una cierta característica considerada de interés para el estudio estadístico. La mayor parte de las veces es muy grande, y algunas veces es hipotética

Muestra
La muestra es *el subconjunto de la población seleccionado* para la investigación. La selección se hace porque generalmente el costo, el tiempo y los recursos son limitados para hacer la investigación con toda la población. A partir de los resultados del estudio con la muestra (siendo ésta representativa de la población), el investigador hace inferencias sobre la población.

Al número de individuos en la muestra se le llama *Tamaño de Muestra*. Cuando el tamaño de muestra (n) es mayor de 30 se le llama muestra grande.

Parámetro
Es una medida (un número) utilizada para describir una característica de la población. (Media, mediana, varianza, etc.). Es un elemento descriptivo de la población.

Estadístico (o estadísticas)
Es una medida que se utiliza para describir una característica numérica de la muestra, no de la población. Es un elemento descriptivo de una muestra

Variables
Las características de interés en una población o una muestra se llaman variables. Como estas características no se mantienen constantes de un individuo a otro, pueden asumir más de un valor, (de ahí su nombre).

Datos
Son las observaciones, es decir, los valores que asumen las variables en cada uno de los individuos

EJERCICIOS 0.1

Selecciona la opción más apropiada, y responde la pregunta.

1.- El proceso de recoger, organizar y representar los datos demográficos de los estudiantes de un salón de clase es llamado estadística

a. Inferencial b. Descriptiva c. Paramétrica d. No paramétrica

2.- El proceso de utilizar muestras estadísticas para llegar a conclusiones sobre los parámetros de la población se llama

a. Inferencia estadística b. Muestreo
c. método científico d. estadística descriptiva

3.- El total de objetos bajo consideración o investigación del que se selecciona una muestra se llama

a. Población b. Descripción c. Parámetro d. Estadística

4.- La parte de la población escogida para hacer el análisis estadístico se llama

a. Selección b. Ejemplo c. Muestra d. Censo

5.- Una medida obtenida de una muestra se llama

a. Parámetro b. Estadístico c. Promedio d. Descripción

6.- ¿Cuándo haces uso de la estadística?

7.- En una escuela de 1,325 estudiantes el director ha decidido seleccionar un grupo de 80 estudiantes para determinar las preferencias de los estudiantes con respecto a los servicios de cafetería que ofrece la escuela. Selecciona la opción que describe más adecuadamente lo expresado en los incisos.

A. población B. muestra C. estadístico(s) D. parámetro(s)

( ) a. Las características de los 80 estudiantes
( ) b. El grupo de 80 estudiantes
( ) c. Las medidas que el director calculará con los datos recogidos
( ) d. Los 1,325 estudiantes de la escuela
( ) e. Los valores que se obtienen con la información proveniente de la muestra
( ) f. El porcentaje de estudiantes de la escuela que no quieren cambios en los servicios de cafetería
( ) g. La frecuencia con que los 80 estudiantes han recibido malos servicios de cafetería
( ) h. El promedio del índice académico de los estudiantes de toda la escuela

Clasificación de datos y variables

Por extensión las variables reciben el mismo nombre de los datos:

*Categóricas o **Cualitativas***

Son las variables cuyos posibles valores son únicamente categorías o nombres, los cuales denotan **cualidades o atributos**, como sexo, afiliación política, color de los ojos, etc. Por lo general, estas características no se pueden describir por medio de números.

*Numéricas o **Cuantitativas***

Son aquellas variables que toman valores numéricos como resultado de un proceso de **conteo o medición**. Las preguntas que se hacen sobre estas variables se pueden responder con un número. ¿Cuánto pesas? ¿Cuánto mides? ¿Cuánto dinero ganas? ¿Cuántos hijos tienes? Además, las variables numéricas pueden ser Discretas o Continuas.

Escalas de medición

El tipo de análisis estadístico que se lleva a cabo sobre los datos depende del nivel o escala de medición de las variables de la investigación. La importancia de esta clasificación por niveles reside en el hecho de que mientras más complejo o alto es el nivel de medición, más efectivos son los métodos estadísticos que se pueden utilizar.

Medir es más que determinar las dimensiones de un objeto. **Medir** en Estadística significa **observar el valor** que toma **la variable en cada elemento** de la población o de la muestra**.**

Por ejemplo en una población de personas, se mide cuando se determina: la religión, el color de ojos, el ingreso anual, el género, el peso, la puntuación en un examen, etc. En una población de perros, se mide cuando se observa: la raza, el tamaño, el número de crías, el color de pelo, la edad, las enfermedades comunes, etc.

Escala nominal

Se utiliza cuando los datos están clasificados en **categorías** en las que **no es posible** establecer **una relación de orden**. Se refiere a atributos de los sujetos, no a cantidades. Ejemplos: tez, religión, partido político, raza, etc.

Escala ordinal

Además de agruparse en **categorías**, se muestra **un orden o secuencia** de los datos de acuerdo al grado de posesión de cierto atributo. Sin embargo, no hay un sentido numérico para este orden. La diferencia entre dos rangos no es una cantidad exacta. Ejemplo: {preescolar, primaria, secundaria, bachillerato, licenciatura, maestría, doctorado}; {soldado raso, cabo, sargento, teniente, capitán, mayor, general, coronel}.

Como puedes observar las escalas nominal y ordinal corresponden a variables de tipo Cualitativo o Categórico

Escala intervalar

Los valores de las variables son datos **numéricos**, sin embargo **no son proporcionales**. por ejemplo un temblor de 8º es veinte veces mas intenso que uno de 6º, y no dos veces además **el cero es arbitrario** y no implica ausencia del fenómeno, por ejemplo: la temperatura cero, en grados Celsius es diferente al cero en grados Fahrenheit y ninguno implica ausencia de temperatura.

Escala de razón

Los valores de la variables son datos **numéricos proporcionales** y tiene **un cero real**. Las operaciones aritméticas de producto y de cociente toman una interpretación válida. Por ejemplo: peso, altura, edad, etc. Tiene sentido hablar de que una persona de 80 años tiene el doble de años que otra de 40 años.

Las escalas intervalar y de razón corresponden a variables de tipo Cuantitativo o Numérico.

Ejercicio 0.2

1.- Selecciona la opción que representa la escala de medición para cada variable

A. nominal B. Ordinal C. Intervalar D. de razón

( ) a.- El número de cuestionarios que una persona ha llenado en el último año
( ) b.- La distancia que un carro conduce en un año
( ) c.- El tiempo que una persona ha tenido una licencia de conducir
( ) d.- La cantidad de veces que una persona fue al cine en el último semestre
( ) e.- La edad de una persona
( ) f.- Índice de criminalidad en una zona especifica del D.- F.-
( ) g.- La puntuación que obtuvo un estudiante en la Prueba de Razonamiento Matemático
( ) h.- Profesión
( ) i.- La temperatura del salón de clases
( ) j.- Nota obtenida en la clase de estadística
( ) k.- El nivel de aprobación de un programa social
( ) l.- Tiempo de trabajo con el microscopio durante el día
( ) m.- Años después de la graduación
( ) n.- partido político preferido
( ) o.- Peso
( ) p.- El tiempo usando la computadora
( ) q.- Procesador de palabras utilizado
( ) r.- El IQ de una persona
( ) s.- Altura de los árboles cercanos al salón de clase
( ) t.- Color de ojos

UNIDAD I : ESTADISTICA DESCRIPTIVA

PROPÓSITO

Que el estudiante comprenda y aplique algunas técnicas de recopilación, organización y representación de un conjunto de datos, proveniente del planteamiento, la discusión y la resolución de problemas, para interpretar y analizar el comportamiento de variables en dicho conjunto.

Distribución de Frecuencia

Como recordarás del capítulo anterior de esta guía, la Estadística Descriptiva se encarga de la organización, presentación y descripción de los datos recolectados, y de obtener información a partir de ellos.

El objetivo de la organización de datos es acomodarlos en forma útil para revelar sus características esenciales y simplificar ciertos análisis.

Cuando el tamaño de muestra es menor a 30, los datos pueden tratarse individualmente, y en este caso se les llama Datos no agrupados. Sin embargo, cuando la muestra es grande ($n \geq 30$), es laborioso hacerlo de esta forma, por lo que se lleva a cabo algún tipo de agrupación preliminar para realizar el tratamiento adecuado a los datos. En este último caso, se les llama Datos Agrupados.

Datos no agrupados

Si los datos están en una escala por lo menos ordinal, lo primero que podemos hacer es ordenarlos, en forma ascendente o descendente. Una vez ordenados los datos de la muestra se organizan en una tabla de frecuencias.

Una Tabla de Frecuencias, también llamada de *Distribución de Frecuencias*, está formada por las categorías o valores de la variable y sus correspondientes frecuencias

Utilicemos un ejemplo para identificar cada elemento de una distribución de Frecuencias.

En un grupo de Estadística I del Cch Sur, se observó la estatura de 16 alumnos y se obtuvieron los siguientes datos (en metros):

1.58	1.64	1.79	1.58	1.64	1.53	1.64	1.66
1.53	1.52	1.76	1.57	1.70	1.74	1.66	1.52

Datos ordenados

1.52	1.52	1.53	1.53	1.57	1.58	1.58	1.6	1.64	1.64	1.64	1.66	1.66	1.74	1.76	1.79

Distribución de Frecuencias

La frecuencia, también llamada frecuencia simple o absoluta, se define como el número de veces que aparece un dato x_i, y se denota por f.

Estatura x_i	Frecuencia f
1.52	2
1.53	2
1.57	1
1.58	2
1.60	1
1.64	3
1.66	2
1.74	1
1.76	1
1.79	1

La *frecuencia relativa* es el número de veces que aparece cada valor de la variable Xi, es decir cada dato, dividida entre el tamaño de la muestra. Se representa con f_r, y se tiene que: $f_r = \frac{f}{n}$

Estatura x_i	Frecuencia f	Frecuencia Relativa f_r
1.52	2	$^2/_{16} = 0.1250$
1.53	2	0.1250
1.57	1	0.0625
1.58	2	0.1250
1.60	1	0.0625
1.64	3	0.1875
1.66	2	0.1250
1.74	1	0.0625
1.76	1	0.0625
1.79	1	0.0625

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

La *frecuencia relativa acumulada* de un valor x_i es la suma de las frecuencias relativas de todos los valores menores o iguales al valor x_i, (o dividiendo las frecuencias acumuladas entre el tamaño de muestra), y se representa por F_{ra}..

Estatura x_i	Frecuencia F	Frecuencia Relativa f_r	Frecuencia Acumulada F_a	Frecuencia Acumulada Relativa F_{ar}
1.52	2	0.1250	2	$^2/_{16} = 0.1250$
1.53	2	0.1250	2+2 = 4	$^4/_{16} = 0.2500$
1.57	1	0.0625	2+2+1 = 5	$^5/_{16} = 0.3125$
1.58	2	0.1250	2+2+1+2 = 7	0.4375
1.60	1	0.0625	**8**	0.5000
1.64	**3**	0.1875	11	0.6875
1.66	2	**0.1250**	13	0.8125
1.74	1	0.0625	14	**0.8750**
1.76	1	0.0625	15	0.9375
1.79	1	0.0625	16	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 "Tres estudiantes de 16 miden 1.64 m de estatura"

fr "El 12.50% de los estudiantes miden 1.66 m de estatura"

F_a "8 de 16 estudiantes miden máximo 1.60 m de estatura"

F_{ar} "El 87.5% de los estudiantes miden hasta 1.74 m de estatura"

Ejercicios 1.1

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

+	250	560	340	780	890	960	470	340	540	440	120	340	340	550	440
	450	450	670	860	430	330	230	810	70	970	360	560	1120	370	840

Organiza los datos en una tabla de distribución de frecuencias, y
+ Escribe algunas frases de la información que proporciona la tabla de distribución de frecuencias:

a.- _____

b.- _____

c.- _____

d.- _____

Medidas de Tendencia Central

Los parámetros más útiles son las medidas de Tendencia Central, las cuales ubican el valor alrededor del cual se concentra un conjunto de datos y las Medidas de Dispersión que describen la variabilidad o dispersión de los mismos.

Las tres medidas de tendencia central o de centralización más importantes son la moda, la mediana y la media.

Consulta en dos fuentes distintas, la definición de:

Moda
1.- _____

_____

2.- _____

_____

Mediana
1.- _____

_____

2.- _____

_____

Media
1.- _____

_____

2.- _____

_____

Moda

Como pudiste observar en la bibliografía, la *moda* se define como el dato con la frecuencia más alta, es decir, el que más se repite. No siempre existe una moda y en ocasiones puede haber más de una. Además, es la única medida de tendencia central que se puede calcular para variables nominales.

Ejemplos:
En el conjunto de datos: $\{2, 3, 3, 4, 4, 4, 5, 5, 8, 8, 12, 13\}$ la moda es 4.

En la distribución $\{2, 2, 3, 3, 5, 5, 8, 8, 12, 12, 13, 13\}$ no hay moda.

Para el conjunto de datos ordinales: $\{$pequeña, pequeña, mediana, mediana, mediana, *grande*, grande, grande, extragrande, extragrande$\}$, hay dos modas: “mediana” y “grande”, porque ambos se repiten el mismo numero de veces.

Mediana

La *mediana* se define como el dato central de la distribución, es decir el dato que queda justo en el medio, cuando el conjunto de datos se encuentra ordenado. Se denota por $\tilde{x}$.

La mediana se puede utilizar con variables ordinales (además de la moda). Si el número de datos es impar, entonces la mediana corresponde al valor que se encuentra en el medio. Pero si el número de observaciones es par, entonces se toman los dos valores que se hallan en el medio de la distribución y se dice que la mediana se encuentra entre esos dos valores, (en el caso de variables numéricas se suman esos valores y se divide entre dos)

Ejemplos:

En el conjunto de datos: {a, b, b, c, c, *c, d*, d, g, g, k, m} la mediana esta entre c y d.

Para el conjunto de datos {2, 2, 3, 3, 5, *5*, 8, 8, 12, 12, 13} la mediana es 5

En el conjunto de datos: {2, 3, 3, 4, 4, *4, 5*, 5, 8, 8, 12, 13} la mediana es 4.5

En el siguiente conjunto de datos ordinales {pequeña, pequeña, mediana, mediana, mediana, *grande*, grande, grande, grande, grande, grande, extragrande, extragrande}, la mediana es "grande"

La mediana divide al conjunto de datos justo a la mitad por lo que nos proporciona información del estilo: "El 50% de los datos esta por debajo de la mediana y el otro 50% por arriba de ella"

Media

Si los datos son numéricos (en escala intervalar o de razón), entonces es posible calcular una tercera medida de tendencia central: la *media aritmética,* la cual consiste en la suma de todos los valores dividida por el número de ellos.

Se denota con $\bar{x}$ y queda expresada como: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

La media aritmética es lo que usualmente conocemos como "promedio", y se interpreta como tal. Una característica de la media es que resulta sensible a datos extremos, lo que no sucede con la mediana ni con la moda.

Ejemplos

En el conjunto de datos: {2, 3, 3, 4, 4, 4, 5, 5, 8, 8, 12, 13}, la moda es 4, la mediana es 4.5 y la media es 6.45.

Para el conjunto de datos {2, 3, 3, 4, 4, 4, 5, 5, 8, 8, 12, 93}, la moda es 4, la mediana es 4.5 y la media resulta 13.72.

Un ejemplo más:

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

Calculemos las Medidas de Tendencia Central

moda = 1.64

mediana = $\tilde{X} = \frac{1.60+1.64}{2} = 1.62$

media = $\bar{x} = \frac{\sum_{i=1}^{16} x_i}{16} = \frac{25.96}{16} = 1.6225$

Información proporcionada:

moda: "La estatura más frecuente entre los estudiantes es de 1.64 m"

mediana: "El 50% de los estudiantes miden menos de 1.62 m y el otro 50% mide más de 1.62m"

moda: "Los estudiantes tienen una estatura promedio de 1.6225 m "

Ejercicios 1.2

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

250	560	340	780	890	960	470	340	540	440	120	340	340	550	440
450	450	670	860	430	330	230	810	70	970	360	560	1120	370	840

Calcula las tres medidas de tendencia central y escribe la información que proporcionan

a.- _____

b.- _____

c.- _____

Medidas de Dispersión

A las Medidas de Dispersión también se les llama Medidas de Variación. La variación es la cantidad de dispersión, o "separación", que presentan los datos.

Rango

El rango de un conjunto de números es la diferencia entre el mayor y el menor de todos ellos. Se denota por *R* y se tiene que $R = x_n - x_1$

Varianza

La varianza es la suma de los cuadrados de las diferencias de los datos con relación a su media aritmética, dividida entre el tamaño de la muestra menos 1.

Se denota por S^2, y se tiene $S^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Si se dispone de una tabla de distribución de frecuencias el calculo varía, utilizando la expresión :

$S^2 = \dfrac{\sum_{i=n}^{k}(x_i - \bar{x})^2 * f_i}{n-1}$ en la cual, k es el número de datos distintos en la muestra.

Desviación Estándar

Un inconveniente de la varianza es que sus unidades de medición se encuentran al cuadrado, por lo que no se puede comparar con la media aritmética. Debido a esto, se define la Desviación Estándar como la raíz cuadrada de la varianza.

Se denota por S, y se tiene $S = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

De igual manera, existe una expresión equivalente: $S = \sqrt{\dfrac{\sum_{i=n}^{k}(x_i - \bar{x})^2 * f_i}{n-1}}$

Coeficiente de Variación

El coeficiente de variación es una medida relativa de la variación. Mide la dispersión de los datos con respecto de su media.

Se denota por CV y se expresa en porcentaje: $CV = \left(\frac{S}{\bar{x}} \right) \cdot 100\%$

El coeficiente de variación se utiliza principalmente cuando se desea comparar dos distribuciones de frecuencia que tienen diferente unidad de medida.

Ejemplo:

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

Calculemos las Medidas de Dispersión

Rango R = 1.79 – 1.52 = 0.27

Para realizar los cálculos de la varianza “a mano”, resulta conveniente construir una tabla como la siguiente

Estatura x_i	Frecuencia f	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 * f_i$
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

$\bar{x}$ = 1.6225 Σ = 0.1095

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

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

Coeficiente de Variación CV = $\frac{0.08544}{1.6225}$ 100 % = 5.266%

Démosle sentido a estos números:

R "La máxima diferencia de estaturas entre los estudiantes es de 27 cm."

S "Las estaturas de los estudiantes se desvían en promedio 8.54 cm. de su media." (equivalente a 0.08544 m.)

CV "Las estaturas varían 5.266% con respecto a su media"

Medidas de Posición

Los cuantiles son medidas de posición "no central" que se utilizan para resumir o describir las propiedades de conjuntos grandes de datos numéricos. Los cuantiles que se calculan más a menudo son: cuartiles, deciles y percentiles.

Cuartiles

Son tres valores numéricos que dividen a la muestra ordenada en cuatro partes iguales. Se denotan por Q_1, Q_2, Q_3.

Primer cuartil, es un valor tal que 25% de las observaciones son menores y 75% son mayores. $Q_1 = x_{\frac{n+1}{4}}$. Recuerda que el subíndice indica la posición del dato en el conjunto.

Segundo cuartil, es un valor tal que 50% de las observaciones son menores y 50% son mayores. Coincide con el valor de la mediana. $Q_2 = x_{\frac{2(n+1)}{4}}$

Tercer cuartil, es un valor tal que 75% de las observaciones son menores y 25% son mayores. $Q_3 = x_{\frac{3(n+1)}{4}}$

Busca en la bibliografía recomendada, qué son y cómo se calculan los deciles y los percentiles

A continuación te mostramos un ejemplo sobre el cálculo de los cuartiles.

Ejemplo:
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

Q_1 Q_2 Q_3

Calculemos algunas Medidas de Posición

$Q_1 = x_{\frac{16+1}{4}}$ = 1.55

"El 25% de los estudiantes miden menos de 1.55 m y el otro 75% mide más"

$Q_2 = x_{\frac{2(16+1)}{4}}$ = 1.62

"El 50% de los estudiantes miden menos de 1.62 y el otro 50% mide más"

$Q_3 = x_{\frac{3(16+1)}{4}}$ = 1.66

"El 75% de los estudiantes miden menos de 1.66 y el otro 25% mide más"

Ejercicios 1.3

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

250	560	340	780	890	960	470	340	540	440	120	340	340	550	440
450	450	670	860	430	330	230	810	70	970	360	560	1120	370	840

Calcula las medidas dispersión y las de posición y escribe la información que proporciona cada una

a.- _____

b.- _____

c.- _____

d.- _____

e.- _____

f.- _____

Datos Agrupados

Distribución de frecuencia

Cuando la muestra es grande (n mayor que 30) resulta conveniente organizar los datos en intervalos de clase para construir su distribución de frecuencias.

Para ejemplificar esta situación, analicemos los datos siguientes correspondientes 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

El número de intervalos de clase depende del número de observaciones. Una mayor cantidad de datos requiere un mayor número de clases. Por lo general la distribución de frecuencias debe tener como mínimo 5 intervalos, pero no más de 15.

Aunque, no existe una regla formal para determinar el número de intervalos y el tamaño de los mismos, existen algunas reglas empíricas que resultan útiles en esta decisión

Denotemos con K al número de intervalos de clase y con C su tamaño; utilizaremos la Regla de Sturges:

$$K = \frac{Rango}{1 + 3.322\ Log(n)}; \qquad C = \frac{Rango}{K}$$

Para nuestro ejemplo, $K = \frac{52 - 16}{1 + 3.322\ Log(55)} = 5.30$

Como K debe ser un número entero, se redondea y se tienen K = 5 intervalos.

Los intervalos serán de tamaño, $C = \frac{52 - 16}{5} = 7.2$, el cuál se redondea hasta la precisión de nuestros datos, es decir a enteros, por lo que C = 7.

Tomemos el dato menor como el límite inferior del primer intervalo, (aunque existen otros criterios, este es el más sencillo), y construyamos los intervalos de modo que cada uno sea de tamaño 7, es decir, de manera en cada uno se cuenten 7 enteros.

Intervalo de Clase
16 – 22
23 – 29
30 – 36
37 – 43
44 – 50
51 – 57

Por ejemplo, en el intervalo 16 – 22 hay 7 enteros:

$$\{16,17,18,19,10,21,22\}$$

Observa que, como se llevan a cabo redondeos, resultaron 6 intervalos en lugar de 5, pero recuerda que la Regla no es una Ley, sólo es un guía para el cálculo. Lo importante es que el último intervalo de clase cubra al dato mayor de la muestra.

Frecuencia Simple o Absoluta de los Intervalos de clase.

En la sección anterior se definió la frecuencia como el número de veces que aparece un dato, en el caso de datos agrupados, la definición varia ligeramente:

La Frecuencia (simple o absoluta) de un intervalo es el número de datos que caen en el mismo.

Intervalo de Clase	**Frecuencia**
16 – 22	**9**
23 – 29	**26**
30 – 36	**11**
37 – 43	**3**
44 – 50	**4**
51 – 57	**2**

¿Qué información proporciona esta primera tabla?

"De 55 personas 4 tienen entre 44 y 50 años"

"9 de cada 55 personas tienen 22 años o menos"

"Sólo 2 de 55 personas tienen 51 años o más"

Frecuencia Relativa de los Intervalos de clase.

Se define, igual que en la sección anterior, como la Frecuencia Simple dividida por el tamaño de muestra.

Intervalo de Clase	Frecuencia	**Frecuencia Relativa**
16 – 22	9	$^{9}/_{55}$ = **0.1636**
23 – 29	26	**0.4727**
30 – 36	11	**0.2000**
37 – 43	3	**0.0545**
44 – 50	4	**0.0727**
51 – 57	2	**0.0364**

¿Qué nueva información proporciona esta segunda tabla?

La frecuencia relativa es una medida proporcional de la frecuencia para cada intervalo:

"El 20.00% de las personas tienen entre 30 y 36 años"

"Sólo el 3.64% de las personas tienen 51 años o más"

Frecuencia Acumulada de los Intervalos de clase.

Se construye sumando la frecuencia simple de cada intervalo con las frecuencias de los intervalos que le preceden.

Intervalo de Clase	Frecuencia	Frecuencia Relativa	**Frecuencia Acumulada**
16 – 22	9	0.1636	**9**
23 – 29	26	0.4727	9 + 26 = **35**
30 – 36	11	0.2000	9+26+11 = **46**
37 – 43	3	0.0545	**49**
44 – 50	4	0.0727	**53**
51 – 57	2	0.0364	**55**

Observa que la frecuencia acumulada del último intervalo es igual al tamaño de la muestra, ¿porqué debe suceder esto? _____

¿Qué tipo de información proporciona esta tercera tabla?

"De 55 personas 35 tienen menos de 30 años"

"9 de cada 55 personas tienen máximo de 22 años"

"53 de 55 personas tienen de hasta 50 años"

Frecuencia Acumulada Relativa de los Intervalos de clase.

La frecuencia acumulada relativa se construye, sumando la frecuencia relativa de cada intervalo con las frecuencias relativas de los intervalos que le preceden, o dividiendo la frecuencia acumulada entre el tamaño de muestra.

Intervalo de Clase	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	**Frecuencia Acumulada Relativa**
16 – 22	9	0.1636	9	**0.1636**
23 – 29	26	0.4727	35	0.1636 + 0.4727 = **0.6364**
30 – 36	11	0.2000	46	**0.8364**
37 – 43	3	0.0545	49	**0.8909**
44 – 50	4	0.0727	53	**0.9636**
51 – 57	2	0.0364	55	**0.9999**

Observa que la frecuencia acumulada relativa del último intervalo es aproximadamente igual a 1, ¿porqué sucede esto? _____

¿Cómo obtener información de esta cuarta tabla?

La frecuencia acumulada relativa es una medida proporcional de la frecuencia acumulada hasta el limite superior de cada intervalo:

"Sólo el 16.36% de las personas tienen de hasta 22 años"

"El 63.64% de las personas tienen máximo de 29 años"

"El 89.09% de las personas tienen menos de 44 años"

Ejercicios 1.4

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 una tabla de 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 de Clase	Frecuencia Simple	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Acumulada. Relativa

2.- Escribe algunos ejemplos de la información que se obtiene a partir de cada tipo de Frecuencia del ejercicio 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, en el que aparecen las primeras reacciones a la picadura. Construye una tabla de distribución de frecuencias. (Observa que la precisión de estos datos es de décimas)

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

Intervalo de Clase	Frecuencia Simple	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Acumulada Relativa

4.- Escribe algunos ejemplos de la información que se obtiene a partir de cada columna 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 ----------	Frecuencia Relativa	---------- ----------	---------- ----------
300 - 499	105			
500 - 599	90			
600 - 699	45			
700 - 799	60			1

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

a.- La frecuencia simple 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 _____.

Medidas de Tendencia Central para datos agrupados

Cuando la muestra es grande y los datos se agrupan en intervalos de clase, el cálculo de las medidas de tendencia central varía significativamente. Se hace necesario, además, definir algunos conceptos nuevos, identifica cuáles.

Moda

La moda se definió como el dato con la mayor frecuencia, de manera similar definimos ahora la *Clase Modal*, como aquel intervalo de clase con la mayor frecuencia.

Una vez que identificamos la clase modal, se utiliza la siguiente fórmula para calcular la moda:

$$LR_{\text{inf}} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \cdot C$$

A continuación describimos cada elemento utilizado en esta fórmula:

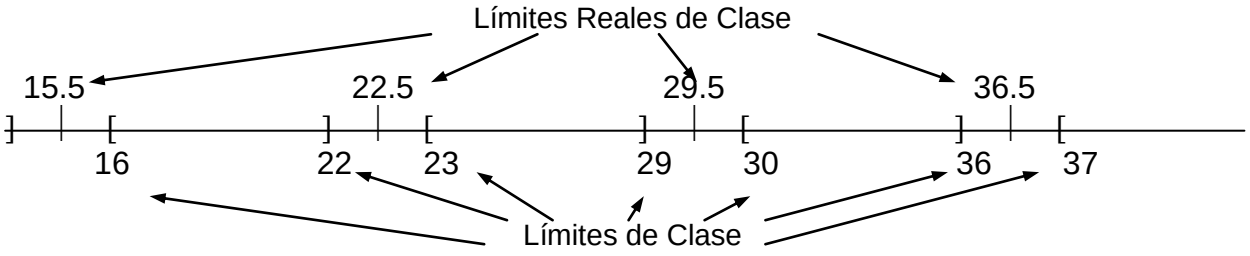
LR_{inf} = límite real inferior de la clase modal.

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

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

C = Tamaño de clase de la clase modal.

Para aclarar lo que son los límites reales observa y analiza el siguiente esquema



Veamos el cálculo de la moda con el ejemplo de la edad de 55 personas:

moda = $22.5 + \left(\frac{17}{17+15} \right) \cdot 7 \approx 26.2$

"La edad más frecuente es de 26.2 años"

Intervalo de Clase	Frecuencia
16 – 22	9
23 – 29	***26***
30 – 36	11
37 – 43	3
44 – 50	4
51 – 57	2

Mediana

La mediana se definió como el dato central cuando el conjunto se encuentra ordenado, ahora definimos la *Clase Mediana*, como aquel intervalo de clase que cubre el 50% de los datos. Para identificarla busquemos el intervalo cuya frecuencia acumulada relativa sea igual o mayor a 0.5

Una vez que identificamos la clase mediana, se utiliza la siguiente fórmula para calcular la mediana:

$$\tilde{x} = LR_{\text{inf}} + \left(\frac{\frac{n}{2} - Fa_1}{f_{med}} \right) \cdot C$$

Cada elemento utilizado en esta fórmula se describe a continuación:

LR_{inf} = límite real inferior de la clase mediana.

Fa_1 = frecuencia acumulada de la clase que precede a la clase mediana.

f_{med} = frecuencia simple de la clase mediana.

C = tamaño de clase de la clase modal.

n = tamaño de muestra

Veamos el cálculo de la mediana con el ejemplo de la edad de 55 personas:

Intervalo de Clase	Frecuencia	Frecuencia Acumulada	Frecuencia Acumulada Relativa
16 – 22	9	***9***	0.1636
23 – 29	***26***	35	***0.6364***
30 – 36	11	46	0.8364
37 – 43	3	49	0.8909
44 – 50	4	53	0.9636
51 – 57	2	55	0.9999

$$\text{mediana} = 22.5 + \left(\frac{\frac{55}{2} - 9}{26} \right) \cdot 7 \approx 27.5$$

"El 50% tales personas tienen una edad menor o igual a 27.5 años y el otro 50% tiene una edad mayor a 27.5 años"

Media

La media igual que antes, se define como el promedio de los datos. Vamos a necesitar el concepto de marca de clase, el cuál es el punto medio de cada intervalo.

No es necesario identificar ninguna clase en particular, y la fórmula para calcular la media es: $\bar{x} = \frac{\sum_{i=1}^{n}(x_i^*)(f_i)}{n}$

Los elementos en esta fórmula son:

x_i^* = marca de clase de cada clase

f_i = frecuencia simple de cada clase.

Veamos el cálculo de la media con nuestro conocido ejemplo de la edad de 55 personas:

Como en otros cálculos, resulta conveniente utilizar una tabla como la siguiente:

Intervalo de Clase	Marca de clase x_i^*	Frecuencia f_i	$(x_i^*)(f)$
16 – 22	19	9	19 * 9 = 171
23 – 29	26	26	676
30 – 36	33	11	363
37 – 43	40	3	120
44 – 50	47	4	188
51 – 57	54	2	108
			∑ = 1626

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i^*)(f_i)}{n} = \frac{1626}{55} \approx 29.6$$

“La edad promedio de tales personas es de 29.6 años”

Medidas de Dispersión para datos agrupados

Rango

Si sólo disponemos de una tabla de frecuencias, el Rango se define como la diferencia entre el límite real superior de la última clase y el límite real inferior de la primera.

Varianza

La varianza para datos agrupados se calcula de manera similar, con algunas modificaciones: las marcas de clase de cada intervalo toman el lugar de los datos y es necesario multiplicar por cada frecuencia simple.

$$S^2 = \frac{\sum_{i=1}^{n} (x_i^* - \bar{x})\, f_i}{n-1}$$

Desviación estándar

Sigue siendo la raíz cuadrada de la varianza: $S = \sqrt{\frac{\sum_{i=1}^{n} (x_i^* - \bar{x})\, f_i}{n-1}}$

Coeficiente de Variación

Se define de la misma forma, como : $CV = \left(\frac{S}{\bar{x}} \right) \cdot 100\%$

Utilicemos nuestro conocido ejemplo de la edad de 55 personas y calculemos las medidas de dispersión para tales datos, es útil una tabla como la siguiente.

Intervalo de Clase	Marca de clase x_i^*	Frecuencia f_i	$x_i^* - \bar{x}$	$(x_i^* - \bar{x})^2$	$(x_i^* - \bar{x})^2 * f_i$
16 – 22	19	9	-10.6	112.3600	1011.2400
23 – 29	26	26	-3.6	12.9600	336.9600
30 – 36	33	11	3.4	11.5600	127.1600
37 – 43	40	3	10.4	108.1600	324.4800
44 – 50	47	4	17.4	302.7600	1211.0400
51 – 57	54	2	24.4	595.3600	1190.7200

$\bar{x}$ = 29.6 $\quad\quad$ Σ = 4201.60

Rango 57.5 – 15.5 = 42

Varianza $S^2 = \frac{4201.60}{54}$ = 77.8074

Desviación Estándar $S = \sqrt{0.0073}$ = 8.8208

Coeficiente de Variación CV = $\frac{8.8208}{29.6}$ 100 % = 29.80%

¿Qué dicen estos números?

R “La máxima diferencia de edades entre estas personas es de 42 años”

S “La edades de tales personas se desvían en promedio 8.82 años de su media.”

CV “Las estaturas varían 29.80% con respecto a su media”

Consulta la bibliografía recomendada para saber cómo calcular las medidas de posición para datos agrupados.

Ejercicios 1.5

1.- Calcula e interpreta las medidas de tendencia central y las medidas de dispersión para los datos agrupados, correspondientes a

a) el número de vuelos internacionales recibidos en el aeropuerto de la ciudad de México durante los dos meses anteriores (del ejercicio 1.4 - 1)

b) un estudio realizado con 40 personas para conocer la reacción sistémica a la picadura de abeja (del ejercicio1.4 - 3)

c) los resultados obtenidos al entrevistar a 300 estudiantes de bachillerato que trabajan mientras estudian (del ejercicio1.4 - 5)

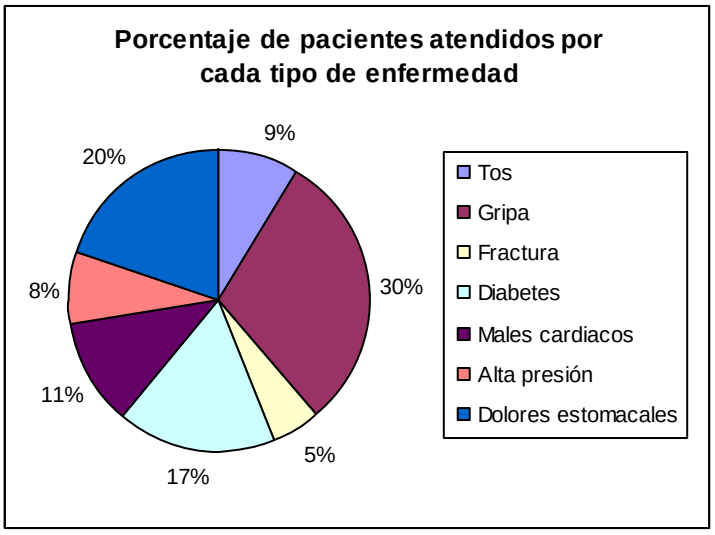
Representación Gráfica

Además de la distribución de frecuencias y de las medidas de tendencia central y de dispersión, resulta conveniente construir alguna representación gráfica de los datos. De esta manera, se tiene una imagen que describe visualmente el comportamiento de los datos.
Cuando los datos son de tipo cualitativo es adecuado utilizar gráficas de barras o circulares. Si los datos son de tipo cuantitativo, el polígono de frecuencias o los histogramas de frecuencias, son los más útiles.

Toda gráfica debe tener: Un título descriptivo, el nombre de la variable que representa, las unidades de la variable, y en su caso la escala utilizada.

Grafica Circular

Se conoce también como Diagrama de pastel, de sectores y otros. Se divide un círculo de manera proporcional a la distribución de los valores de la variable. Ayuda a percibir la importancia relativa de cada categoría respecto al total. Se utiliza también para representar datos discretos.

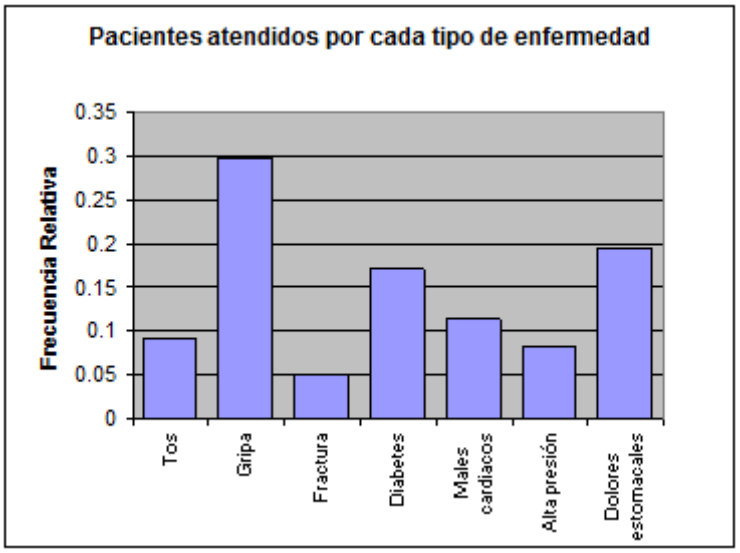


Gráfica de barras

En este tipo de gráfica se muestran en un sistema de ejes cartesianos los valores de la variable, y los valores de la frecuencias, absolutas o relativas.

Los valores de la variable se localizan sobre un eje horizontal y las frecuencias sobre uno vertical. Las barras son rectángulos cuyo ancho es arbitrario, pero debe ser el mismo para todas las barras, y cuya longitud es la frecuencia o el porcentaje de observaciones dentro de la categoría.

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



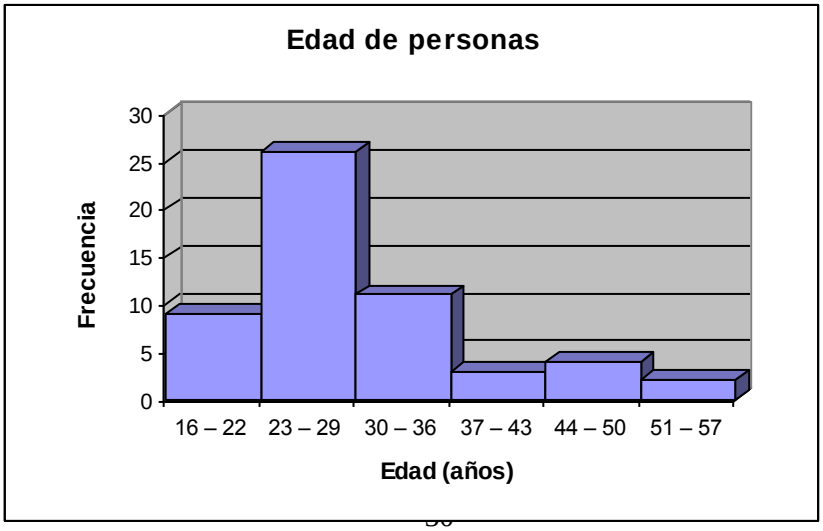
Para una distribución de frecuencias se tienen diferentes representaciones gráficas, tales como:

Histograma

Consiste en un gráfico de barras o rectángulos cuya altura corresponde a la frecuencia de cada valor o de cada intervalo localizada sobre el eje vertical.

Para datos no agrupados, cada frecuencia se representa por una barra cuya área sea proporcional a ella. Típicamente, el ancho de cada barra se escoge como 1 y así, la altura y el área de la barra son iguales a la frecuencia del valor.

Para datos agrupados, el ancho de los rectángulos corresponde al tamaño de los intervalos de clase. Las barras, por lo tanto, son contiguas, y se encuentran centrados en las marcas de clase.

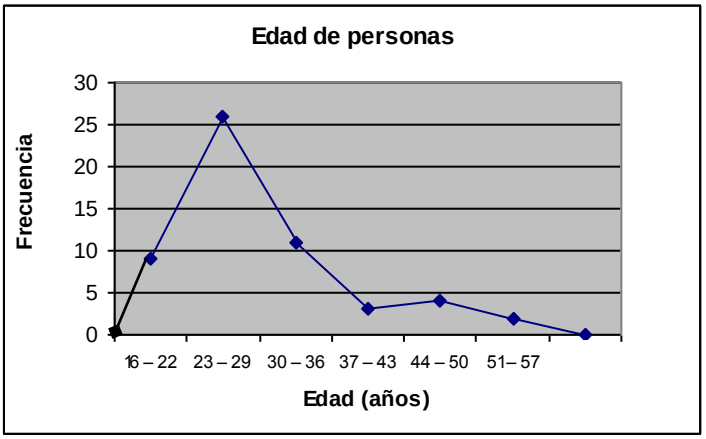


Polígono de Frecuencias

Consiste en una gráfico de líneas trazado sobre un sistema de ejes cartesianos.

Para datos no agrupados, se trazan los puntos que corresponden a los valores de la variable cuantitativa y la frecuencia (absoluta o relativa), a continuación se unen los puntos mediante segmentos de recta, los extremos se unen con el eje horizontal con el primer valor menos una unidad y el extremo derecho mas una unidad.

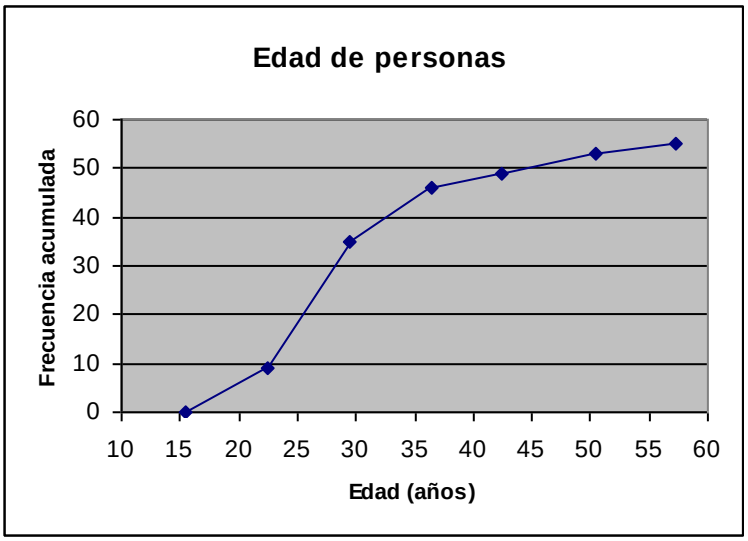
Para datos agrupados los vértices tienen como coordenadas las marcas de clase y las frecuencias correspondientes. Se debe cerrar sobre el eje horizontal en dos puntos que corresponden a las marcas de clase de dos intervalos, uno anterior y el otro posterior al primero y al último intervalo, cuya frecuencia es cero.



Ojiva

Consiste en un poligono de frecuencias acumuladas, por lo tanto es una gráfica de líneas generalmente ascendente.
Para datos no agrupados se trazan los puntos que corresponden a los valores de la variable cuantitativa y la frecuencia (acumulada o relativa acumulada), a continuación se unen los puntos mediante segmentos de recta, el extremo derecho no se une con el eje horizontal.

Para datos agrupados los vértices tienen como abscisa los valores de la variable representados por los límites reales superiores y como ordenada la frecuencia acumulada o frecuencia relativa acumulada (ojiva porcentual).



Ejercicios 1.6

Construye una representación gráfica para:

a) la cuenta de la luz (en pesos) del mes de marzo de 30 familias escogidas aleatoriamente (del ejercicio 1.1 - 1)

b) el número de vuelos internacionales recibidos en el aeropuerto de la ciudad de México durante los dos meses anteriores (del ejercicio 1.4 - 1)

c) un estudio realizado con 40 personas para conocer la reacción sistémica a la picadura de abeja (del ejercicio 1.4 - 3)

d) los resultados obtenidos al entrevistar a 300 estudiantes de bachillerato que trabajan mientras estudian (del ejercicio 1.4 - 5)

Ejercicios adicionales

La siguiente tabla contiene los datos obtenidos al entrevistar a estudiantes, elegidos al azar, de 5º. semestre de CCH.

Nombre	Género (M o F)	Edad (años cumplidos)	Tipo sanguíneo	Color favorito	Número de Hermanos **	Peso (kg)
Verónica	F	17	O^+	Azul	2	63
Guillermo	M	16	O^+	Morado	1	67
Viviana	F	17	O^+	Azul	3	60
Nuria	F	17	A^+	Azul	2	62
Alfredo	M	17	O^+	Rojo	3	75
Gerson	M	17	O^+	Negro	6	74
Nohemí	F	18	A^+	Azul	3	54
Alejandra	F	16	O^+	Blanco	2	61
Viridiana	F	16	O^+	Violeta	2	50
Elizabeth	F	16	O^+	Blanco	3	45
Rogelio	M	17	O^+	Azul	3	74
Amaranta	F	17	A^+	Blanco	1	54
Fabiola	F	16	O^+	Morado	2	54
Zicarú	F	18	O^+	Rosa	3	51
Karla	F	18	A^+	Turquesa	2	55
Andrea	F	17	O^+	Negro	3	60
Alfonso	M	17	O^+	Azul	3	64
Rubí	F	15	B^+	Morado	2	62
Claudia	F	17	O^+	Violeta	3	60
Wendi	F	17	O^+	Negro	3	58

**incluyéndose a sí mismo(a)

1.- Identifica el tipo de variable representada en cada columna

2.- Realiza un análisis descriptivo (distribución de frecuencias, medidas de tendencia central y de dispersión, representación gráfica, etc.) de cada variable (por separado).

UNIDAD II : DATOS BIVARIADOS

PROPÓSITO
Que el estudiante comprenda la forma en que se establece una relación entre dos variables, a partir de tablas, diagramas, regresiones y correlaciones, y describa la naturaleza e intensidad de dicha relación.

Datos bivariados

Se llaman datos bivariados a aquellos que provienen de dos variable medidas al mismo tiempo sobre cada individuo.
Por ejemplo: Edad y Género, Escolaridad e Ingreso, Peso y Estatura, etc.

Dependiendo de la naturaleza de cada variable se da el tratamiento a los datos.

Caso 1: Dos variables Cualitativas

Cuando los datos bivariados provienen de dos variables cualitativas, resulta conveniente organizarlos en una Tabla de Contingencia. Las columnas de esta tabla representan a las categorías de la variable 1 y los renglones representan a las categorías de la variable 2; la frecuencia aparecerá en las celdas centrales de la tabla.

Analicemos este caso con un ejemplo.

La siguiente tabla muestra el número de pacientes hospitalizados por la misma enfermedad en los últimos 6 meses

	Hospital			
Género	Los Ángeles	Médica Sur	20 de Noviembre	López Mateos
Hombres	36	**44**	43	28
Mujeres	34	50	**52**	53

Identifica las dos variables: _____ y _____ .

El número 44 del primer renglón y la segunda columna significa que:

“44 pacientes eran hombres **y** estuvieron hospitalizados en el hospital Médica Sur”

El número 52 del tercer renglón y la tercera columna significa que:

“_____

_____”

Al sumar las frecuencias absolutas de cada fila y de cada columna, se obtiene la frecuencia absoluta marginal.

	Hospital				
Género	Los Ángeles	Médica Sur	20 de Noviembre	López Mateos	Total
Hombres	36	44	43	28	
Mujeres	34	50	52	53	**189**
Total	**70**		95		

¿Que información obtenemos de estos valores?

"70 pacientes (en total) estuvieron hospitalizados en el hospital Los Ángeles"

"189 pacientes (en total) eran mujeres"

"_____ pacientes (en total) estuvieron hospitalizados en el 20 de Noviembre"

"_____ pacientes (en total) eran mujeres"

"_____ pacientes (en total) estuvieron hospitalizados en _____" etc.

Ahora, ¿podríamos saber sobre cuántos pacientes se hizo el estudio?

Claro!, tendríamos que sumar todas las celdas, lo que es equivalente a sumar la última columna o el último renglón que agregamos, y concluimos que: "Se hizo el estudio con n = _____ pacientes"

Frecuencias relativas

Si dividimos todas las celdas de la tabla sobre el tamaño de muestra (total de pacientes), obtenemos una nueva tabla, la cual nos proporciona la Frecuencia Relativa respecto al total.

	Hospital				
Género	Los Ángeles	Médica Sur	20 de Noviembre	López Mateos	Total
Hombres	0.1058				0.4441
Mujeres		0.1470			
Total			0.2794		

¿Qué porcentaje de pacientes eran hombres y estuvieron hospitalizadas en Los Ángeles?

Podemos responder la pregunta anterior utilizando la primera celda de la tabla:

"El 10.58% de los pacientes eran hombres y estuvieron hospitalizados en Los Ángeles"

¿Cómo interpretamos el resultado de la celda en el segundo renglón-segunda columna?

"_____"

"El porcentaje de pacientes que estuvieron en el hospital 20 de Noviembre es _____%"

Por otro lado, si dividimos los valores de cada renglón por el total del mismo, obtenemos la Frecuencia Relativa respecto al Genero.

	Hospital			
Género	Los Ángeles	Médica Sur	20 de Noviembre	López Mateos
Hombres	$^{36}/_{151} = 0.2384$			
Mujeres		$^{50}/_{189} = 0.2645$		

De aquí, obtenemos que:

"El 23.84% de los **pacientes hombres** estuvieron en el hospital Los Ángeles"

"El 26.45% de los **pacientes mujeres** estuvieron en el hospital Médica Sur"

De los pacientes mujeres, el _____% estuvo en el hospital López Mateos"

Ahora, si dividimos los valores de cada columna sobre el total de la misma, obtenemos la Frecuencia Relativa respecto al Hospital.

	Hospital			
Género	Los Ángeles	Médica Sur	20 de Noviembre	López Mateos
Hombres	$^{36}/_{70} = 0.5142$			
Mujeres			$^{52}/_{95} = 0.5473$	

De la tabla anterior, obtenemos que:

"De los pacientes que estuvieron en Los Ángeles, el 51.42 % eran mujeres "

"El 54.73% de los pacientes que estuvieron en el hospital 20 de Noviembre eran _____"

Ejercicios 2.1

1.- La tabla de contingencia siguiente representa el Estado Civil y la preferencia por ciertos periódicos de distintas personas.

	Periódico preferido			
Estado Civil	El Universal	Excélsior	Reforma	La Jornada
Soltero	11	6	7	14
Casado	6	10	10	8
Viudo	5	6	6	9
Separado	7	8	5	12

Con base en la tabla, responde las preguntas y completa la información

a) El periódico Excélsior lo prefieren _____ personas

b) Se entrevistó a _____ personas Viudas.

c) ¿Cuántas personas son solteras y prefieren el periódico la Jornada? _____

d) ¿Qué porcentaje de personas son casadas y prefieren el periódico Reforma? _____

c) De las personas que prefieren el Excélsior, el _____ % son separadas

e) De las personas que prefieren el Universal, ¿qué porcentaje son solteros? _____

f) De las personas separadas, el _____ % prefiere leer la Jornada

g) De las personas viudas, ¿qué porcentaje prefiere leer el Reforma? _____

2.- La siguiente tabla 1 muestra los datos obtenidos al observar el tipo sanguíneo y el género de 20 personas.

Genero	F	M	F	F	M	M	F	F	F	F	M	F	F	F	F	F	M	F	F	F
Tipo Sang.	O^+	O^+	O^+	A^+	O^+	O^+	A^+	O^+	O^+	O^+	O^+	A^+	O^+	O^+	A^+	O^+	O^+	B^+	O^+	O^+

a) Organiza estos datos en una tabla de contingencia

b) Escribe algunos enunciados sobre la información que se obtiene de ella

c) Representa gráficamente

Caso 1: Dos variables Cuantitativas

Cuando los datos bivariados provienen de dos variables cuantitativas resulta de interés estudiar la relación que guarda una con la otra. La relación puede ser de muy distinta naturaleza: lineal, cuadrática, exponencial, logarítmica, trigonométrica, etc. En estadística la relación que nos interesa es la Relación Lineal, por lo que se llevan a cabo Análisis de Correlación Lineal y de Regresión Lineal

El análisis de correlación, se usa para medir la fuerza de asociación entre las variables. El objetivo medir la covarianza que existe entre esas dos variables numéricas.

El análisis de regresión se usa con propósitos de predicción. Se busca desarrollar un modelo estadístico útil para predecir los valores de una variable dependiente o de respuesta basados en los valores de al menos una variable independiente o explicativa.

Ejemplo
Se decidió examinar la relación entre la estatura, (en metros), y el peso, (en kilogramos), a partir de una muestra de 12 alumnas de cierta escuela. 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

Diagrama de dispersión

Es una grafica donde aparecen los valores muestrales considerados como parejas ordenadas (x_1, y_1), (x_2, y_2), …, (x_n, y_n).

Si los valores muestrales dan una configuración de puntos como el del diagrama de dispersión, el modelo se llama de regresión lineal simple.

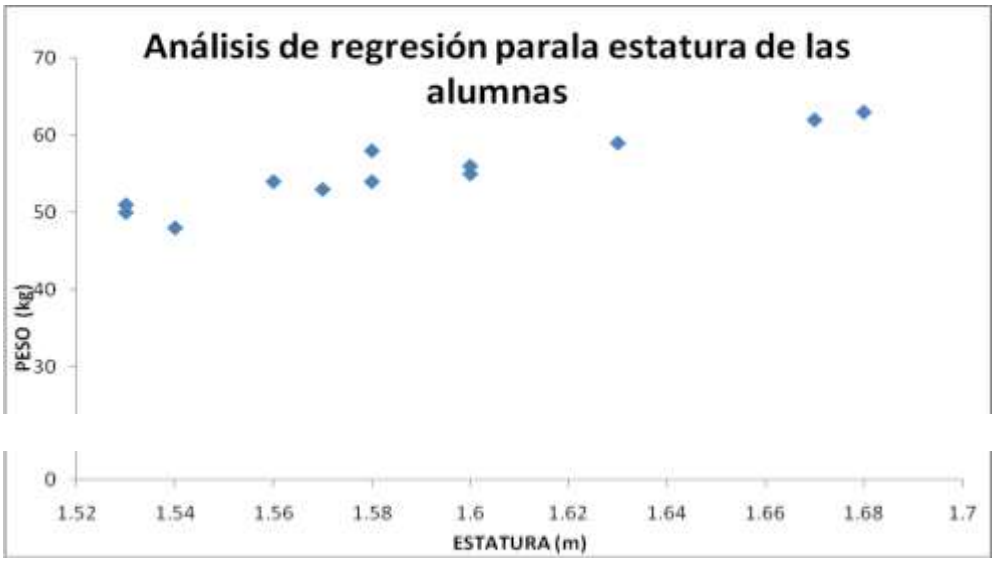


Diagrama de Dispersión

Análisis de Correlación Lineal

El objetivo es ver si existe o no una relación de carácter lineal entre las dos variables, y si existe, entonces medir el grado de intensidad de la linealidad. Esto comúnmente se realiza calculando el coeficiente de correlación lineal de Pearson:

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

El coeficiente toma valores en el intervalo [-1, 1].

Un valor negativo de r significa que la relación entre las variables es inversamente proporcional, (a mayor X menor Y)

Un valor positivo de r significa que la relación entre las variables es directamente proporcional, (a mayor X mayor Y)

Un valor cercano a 0, indica que la relación entre las variables es casi nula, es decir, no hay relación entre ellas.

Un valor cercano a 1 significa que la relación entre las variables es fuertemente lineal.

Análisis de Regresión Lineal

Si se cumplen ciertas suposiciones, la ordenada b de la muestra y la pendiente m de la muestra se pueden usar como estimaciones de los parámetros respectivos de la población m* y b*. Así, la ecuación de regresión muestral que representa el modelo de regresión en línea recta es:

$$Y^*_i = mX_i + b$$

donde

Y^* = valor pronosticado de Y para cada observación
X_i = valor de X para cada observación

Método de Mínimos Cuadrados: se refiere a encontrar la línea recta que mejor se ajuste a los datos, de manera que las diferencias entre los valores reales Y_i y los valores pronosticados a partir de la recta ajustada de regresión Y^*_i sean tan pequeñas como sea posible.

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

Regresando a nuestro ejemplo de estatura y peso de alumnas, para realizar los cálculos es útil construir una tabla como la siguiente:

Alumna	Estatura (m) X	Peso (kg) Y	XY	X^2	Y^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
	19.07	663	1056.01	30.3329	36865

Al sustituir los valores correspondientes para r, m y b se obtiene:

$r = 0.94$, $m = 87.03$, $b = -83.06$

Por tanto, la relación lineal es fuerte y es positiva; y, la ecuación de regresión lineal es

$$Y^* = 87.03\,X - 83.06$$

Utilizando dicha ecuación podemos predecir, por ejemplo, el peso de una alumna cuya estatura es de 1.55 m

$$Y^* = 87.03(1.55) - 83.06 = 51.83$$

De acuerdo a este modelo, una alumna cuya estatura fuera de 1.55 m., tendría un peso de 51.8 kg.

Ejercicios 2.2

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

No. de piezas (x)	1	3	5	10	12	15	24
Costo por pieza (Y)	55	52	48	36	32	30	25

a) El coeficiente de correlación lineal vale _____

b) La recta de regresión lineal por mínimos cuadrados es _____

c) 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) Calcula el coeficiente de correlación r

c) Determina la ecuación de regresión lineal

d) Traza la recta de regresión sobre el diagrama de dispersión

e) Si la densidad del material es 2.9, determina el valor estimado del contenido de hierro.

f) Si el contenido de hierro es de 31, determina la densidad estimada del material

3.- En un análisis de regresión la pendientes de la recta de mejor ajuste vale $\hat{\beta}_1 = 4.86$ y la ordenada al origen es $\hat{\beta}_0 = 5$.

a) La ecuación de esa recta de mejor ajuste es _____

b) Considerando la recta de regresión de la pregunta anterior, ¿qué efecto causa un valor de $x = 2$? _____

UNIDAD IV : PROBABILIDAD

PROPÓSITO

Que el estudiante estudie los fenómenos aleatorios, resolviendo problemas utilizando los tres enfoques, subjetivo, frecuentista y clásico, para comprender conceptos fundamentales que le permiten interpretar a la probabilidad y a sus reglas relacionadas directamente con la Inferencia Estadística.

PROBABILIDAD

La probabilidad tiene un papel crucial en la aplicación de la inferencia estadística y la toma de decisiones bajo incertidumbre. Sin una adecuada comprensión de las leyes básicas de la probabilidad, una inferencia (o una decisión), cuyo fundamento es la información proporcionada por una muestra aleatoria, puede estar equivocada.

Fenómenos Aleatorios y Fenómenos Determinísticos.

Todos los hechos o sucesos que ocurren se denominan fenómenos.

Fenómeno Determinista.- Es el fenómeno cuyo resultado se predice con certeza, porque obedece a una relación causa-efecto y al variar poco las causas varía poco el efecto.
Ejemplo: cuánto costarán 35 litros de gasolina si un litro cuesta $6.10, cuándo será visto en México el siguiente eclipse total de sol; al disparar un proyectil con el mismo ángulo de elevación y las mismas condiciones describe la misma parábola, etc.

Fenómeno Aleatorio.- Es un fenómeno que tiene varios resultados y estos no se pueden predecir con certeza, pues obedecen las leyes del azar.

Ejemplo: el resultado probable de una rifa; cuál será el equipo ganador de fútbol en el próximo campeonato; qué cara quedará arriba al lanzar un dado; si llueve o no llueve mañana; el tiempo que tardará un árbol en alcanzar 3m de altura etc.

Un *Experimento aleatorio* es una acción que se considera con propósito de análisis y que tiene como fin determinar la probabilidad de uno o de varios resultados. En la práctica, un experimento es el proceso por medio del cual una observación o medición es registrada.

Un experimento aleatorio se caracteriza por:
a) El experimento se puede repetir indefinidamente bajo las mismas condiciones
b) Cualquier mínima modificación en las condiciones iniciales pueden modificar el resultado final
c) Se puede determinar el conjunto de los posibles resultados del experimento, pero no se puede predecir previamente un resultado

Espacio Muestral es el conjunto de (todos) los **posibles resultados** en un experimento aleatorio. Generalmente se denota con Ω (o con *S*). A cada uno de estos resultados, también se les llama *puntos muestrales*.

Ejemplos:
1.- Experimento: Se lanza una moneda y se observa la cara superior (es decir, lo que "cae").

$$\Omega = \{ s, a \}$$

2.- Experimento: Se lanza un dado común y se observa la cara superior

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

Cualquier subconjunto de Ω es denominado *Evento aleatorio*, y se denota normalmente con las letras mayúsculas *A*, *B*, C, ...

Si un espacio muestral contiene *n* elementos, hay un total de 2^n subconjuntos o eventos (y a esto se le conoce como conjunto potencia).

Ejemplo
Experimento: Se lanza un dado común y se observa la cara superior.

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

Evento A: el número que "cae" es par. $A = \{ 2, 4, 6 \}$
Evento B: el número que "cae" es primo. $B = \{ 1, 2, 3, 5 \}$

A un evento que contiene un solo elemento, se le llama *evento simple o elemental.*

A un evento que contiene más de un elemento, se le llama *evento compuesto.*

A un evento que contiene el mismo número de elementos que Ω, se le llama *evento seguro.*

Un evento que no tiene elementos es llamado *evento imposible.*

Ejemplo:
Experimento: Se lanza una moneda tres veces.

$$\Omega = \{ (S,S,S), (S,S,A), (S,A,S), (A,S,S), (A,A,S),(A,S,A),(S,A,A), (A,A,A) \}$$

Evento elemental: C: Que salgan tres soles; $C = \{ (S,S,S) \}$

Evento compuesto: D: Que salgan dos soles; $D = \{ (S,S,S), (S,S,A), (S,A,S), (A,S,S) \}$,

Evento imposible: E: que salgan cuatro soles $E = \phi$

Evento seguro: F: Que salgan entre 0 y 3 soles $F = \Omega$

Enfoques de Probabilidad

La probabilidad clásica se refiere a situaciones ideales, donde todos los casos o resultados posibles tienen la misma probabilidad de ocurrencia (son equiprobables). La probabilidad frecuencial proporciona estimaciones de la probabilidad que pueden variar, dependiendo del número de observaciones realizadas. La frecuencia subjetiva de un evento es asignada por el investigador con base en su experiencia.

Probabilidad Clásica

Supongamos un espacio muestral $\Omega = \{a_1,\ldots a_N\}$ de manera que los a_i son sucesos elementales igualmente probables y sea un suceso $E = \{a_1,\ldots a_k\}$ ($k \leq N$). Se define la probabilidad P del evento E, como

$$P(E) = \frac{N(E)}{N(\Omega)}$$

Ejemplo:
Experimento: Se lanza una moneda tres veces.

$$\Omega = \{ (S,S,S), (S,S,A), (S,A,S), (A,S,S), (A,A,S),(A,S,A),(S,A,A), (A,A,A) \}$$

Evento C: Que salgan tres soles; $P(C) = \frac{1}{8}$

Evento D: Que salgan dos soles; $P(D) = \frac{4}{8}$

Evento E que salgan cuatro soles; $P(E) = P(\phi) = \frac{0}{8} = 0$

Evento F: Que salgan entre 0 y 3 soles; $P(F) = \frac{8}{8} = 1$

Cómo puedes observar, una función de probabilidad tiene las siguientes verdades básicas o axiomas.

1. Si E es un evento cualquiera, entonces $0 \leq P(E) \leq 1$
2. Si Ω o S, es el evento seguro, entonces $P(\Omega) = 1$ o $P(S) = 1$
3. Si $E_1, E_2, \ldots E_k$ son eventos *mutuamente excluyentes,* entonces

$$P(E_1 \text{ o } E_2 \text{ o } \ldots E_k) = P(E_1) + P(E_2) + \ldots + P(E_k)$$

Operaciones Básicas con Eventos

Ya que los eventos aleatorios son subconjuntos del conjunto Ω, espacio muestral, se pueden aplicar las conocidas operaciones con conjuntos, a los eventos, como son la unión, la intersección y la diferencia de eventos.

UNION $A \cup B$ Unión de eventos originales: es el evento que sucede si y solo si A sucede o B sucede o ambos suceden

INTERSECCION $A \cap B$ Intersección de los eventos originales, es el evento que sucede si y sólo si A y B suceden simultáneamente.

DIFERENCIA A - B La diferencia de los eventos originales A y B, es el evento que sucedo solo en A pero no en B.

Gráficamente estas operaciones se pueden representar a través de los diagramas de Venn.

Sea Ω el espacio muestral y A y B eventos tal que $A, B \subset S$ gráficamente, en la figura 1 se presenta el caso donde los eventos A y B no tienen elementos del espacio muestral en común y en la figura 2 se presenta el caso donde los eventos A y B tienen elementos del espacio muestral en común..

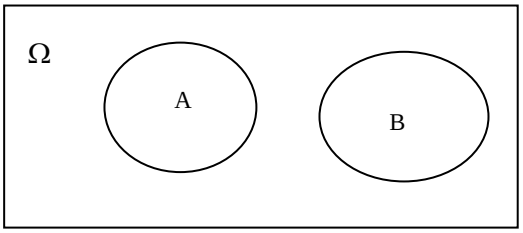


Fig. 1

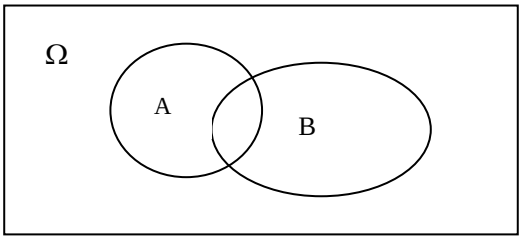


Fig. 2

Dos eventos A y B son mutuamente exclusivos, cuando no pueden ocurrir simultáneamente, es decir, $A \cap B = \varnothing$, lo que ocurre en la fig. 1.

Ejemplo: Experimento: Se lanza un dado.
Espacio muestral = total de caras en que puede caer el dado, o sea seis formas de interés:
$\Omega = \{ 1,2,3,4,5,6 \}$, $N(\Omega) = 6$
Sean A, B, C los eventos: A: Que caiga un número impar $= \{ 1, 3, 5 \}$, $N(A) = 3$
B: Que caiga un número mayor de 2 y menor que 5 $= \{ 3, 4 \}$, $N(B) = 2$
C: Que caiga un número par $= \{ 2, 4, 6 \}$, $N(C) = 3$

a).- Unión:

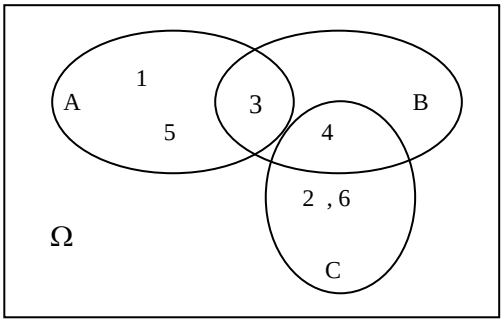
$A \cup B = \{ 1, 3, 5 \} \cup \{ 3, 4 \} = \{1,3,4,5\}$, $N(A \cup B) = 4$
$A \cup C = \{ 1, 3, 5 \} \cup \{ 2,4,6 \} = \{1,2,3,4,5,6\} = S$, $N(A \cup C) = N(S) = 6$

b).- Intersección:

$A \cap B = \{ 1, 3, 5 \} \cap \{ 3, 4 \} = \{3\}$, $N(A \cap B) = 1$
$A \cap C = \{ 1, 3, 5 \} \cap \{ 2,4,6 \} = \{\Phi\}$, $N(A \cap C) = N\{\Phi) = 0$

c).- Diferencia:
$A - B = = \{ 1, 3, 5 \} - \{ 3, 4 \} = \{ 1, 5 \}$, $N(A - B) = 2$

d).- Complemento:
$A^c = \{ 2, 4, 6\} = C$ $N(A^c) = N(C) = 3$



Probabilidad frecuencial y regularidad estadística

Las frecuencias relativas de un evento tienden a estabilizarse cuando el número de observaciones se hace cada vez mayor.

Ejemplo:
La regularidad estadística en el experimento del lanzamiento de monedas, indica que las frecuencias relativas del evento: que salga sol {s }, se tiende a estabilizar aproximadamente en 0.5= 1/2.

Si un experimento se repite N veces bajo las mismas condiciones, la probabilidad de un evento A, denotada por P(A), es el valor en el que se estabilizan las frecuencias relativas del evento A, cuando el número de observaciones del experimento se hace cada vez mayor.

Ejemplo:

En los últimos certámenes de belleza ha habido: 7 reinas Europeas, 1 Africana, 5 Latinoamericanas, 3 norteamericanas y 2 Asiáticas.

Calcula la probabilidad de que la reina de belleza de este año sea:
A) Latinoamericana
B) Africana o Asiática
C) Europea
D) No norteamericana

Probabilidad Condicional

Una situación de interés consiste en determinar la probabilidad de un evento si ha ocurrido otro. Por ejemplo, si lanzamos un dado, ¿cuál es la probabilidad de obtener un 3 si se sabe que cayo un número impar?

La información "*se sabe que es impar"* condiciona la probabilidad de ocurrencia del evento *"cae 3"*, es decir, de las 3 posibles resultados impares solamente nos interesan aquel que es 3 ; así, la probabilidad (*llamada probabilidad condicional)*, es $\frac{1}{3} = 0.3333$

Observe que si se calcula solamente P("cae 3"), se obtiene $\frac{1}{6} = 0.1666$, pero la influencia del evento impar modifica su probabilidad a 0.3333

Definición

Sean A y E dos eventos de un espacio muestral Ω, con P(E) > 0. La probabilidad de que ocurra el evento A dado que ha ocurrido E, es decir, la **probabilidad condicional** de A dado E, se define como: $P(A|E) = \frac{P(A \cap E)}{P(E)}$

Además, despejando a P(A∩E), y haciendo (A∩E) = (E∩A), se tiene:

$$P(A \cap E) = P(E \cap A) = P(E)\, P(A/E)$$

Ejemplo
En cierta ciudad, las mujeres representan el 50% de la población y los hombres el otro 50%. Se sabe que el 20% de las mujeres y el 5% de hombres están sin trabajo. Un economista estudia la situación de empleo, elige al azar una persona desempleada. Si la población total es de 8000 personas, ¿Cuál es la probabilidad de que la persona escogida sea?:

a) Mujer b) Hombre c) Mujer sabiendo que está empleada
d) sin empleo dado que es hombre e) Empleada si se sabe que es mujer

Es útil construir una tabla de contingencia para el espacio muestral

	Desempleados	Empleados	Total
Mujeres	800	3200	4000
Hombres	200	3800	4000
Total	1000	7000	8000

Sea los eventos:
E : que la persona seleccionada esté empleada
D : que la persona seleccionada esté desempleada
M : que la persona seleccionada sea mujer
H : que la persona seleccionada sea hombre

Cada una de las entradas de la tabla representan:

	Desempleados	Empleados	Total
Mujeres	M∩D	M∩E	M
Hombres	H∩D	H∩E	H
Total	D	E	

	Desempleados	Empleados	Total
Mujeres	800/8000=.1	3200/8000=.4	4000/8000=.5
Hombres	200/8000=.025	3800/8000=.475	4000/8000=.5
Total	1000/8000=.125	7000/8000=.875	8000/8000=1

P(M) = 0.50 P(H) = 0.50 P(E) = 0.875 P(D) = 0.125

P(M/E) = P(M∩E)/P(E) = 0.40/0.875 = 0.4571

P(D/H) = P(D∩H)/P(H) = 0.025/0.5 = 0.05

P(E/M) = P(M∩E)/P(M) = 0.40/0.5 = 0.08

P(M/D) = P(M∩D)/P(D) = 0.10/0.125 = 0.8

P(H/D) = P(H∩D)/P(D) = 0.025/0.125 = 0.2

Regresando al contexto del problema, estos números significan que:

“La probabilidad de que la persona escogida sea Mujer es del 50%”

“La probabilidad de que la persona escogida sea Hombre es del 50%”

“La probabilidad de que la persona escogida sea Mujer sabiendo que está empleada es del 45.74 %”

“La probabilidad de que la persona escogida este sin empleo dado que es hombre es del 5%”

“La probabilidad de que la persona escogida este Empleada si se sabe que es mujer es del 8%”

Ejercicios 3.1

1.- Se ha recibido un cargamento de toronjas con las siguientes características: 10% son rosadas sin semilla, 20% son blancas sin semilla, 30% son rosadas con semilla y 40% son blancas con semilla. Se selecciona aleatoriamente una toronja del cargamento. Calcula la probabilidad de que:

Sea sin semilla
Sea blanca
Sea rosada o sin semilla
Sea rosada dado que es sin semilla
Sea sin semilla dado que es rosada.

2.- En una ciudad hay una alta incidencia de cirrosis entre la población. Se sospecha que se debe al alto índice de consumo de alcohol. Se hacen estudios estadísticos que asocian “presencia de la enfermedad” con “consumo de alcohol”. Se encuentra que el 40% de la población consume alcohol, el 20% padece la enfermedad y el 5% consume alcohol y padece la enfermedad. ¿Se verifica la creencia?

3.- Relaciona ambas columnas, colocando en los paréntesis de la derecha la letra que corresponda a la aseveración correcta.

A	Lanzamiento de una moneda para observar sus resultados	( )	Distribución de frecuencias
B	Tipo de sangre de las personas	( )	Muestra
C	Número de veces que se repite un dado	( )	Variable
D	Característica que interesa estudiar en una muestra o en una población	( )	Fenómeno aleatorio
E	Subconjunto representativo de un universo	( )	Frecuencia
F	Arreglo de los datos observados	( )	Variable numérica continua
G	Lanzar un objeto hacia arriba y observar que baja	( )	Población
H	Altura de los árboles del CCH Sur	( )	Frecuencia relativa
I	Cociente del número de veces que se repite un dato entre el número total de datos	( )	Fenómeno determinista
J	Universo donde interesa estudiar una característica	( )	Variable categórica nominal.