Guía de Estudio
MÓDULO 18
2023

FENÓMENOS NATURALES Y PROCESOS SOCIALES

NL EDUCACIÓN
GABINETE DE IGUALDAD PARA TODAS LAS PERSONAS

EL GOBIERNO DEL NUEVO NUEVO LEÓN

PREPARATORIA ABIERTA

Coordinadora Estatal de Telebachillerato y del Subsistema de Preparatoria Abierta
Edith Alemán Ramírez

Departamento Académico de la Coordinación de Preparatoria Abierta
Elena Cisneros Rodríguez
Gretel Lizeth Marroquín Lara
Adrián Alcántara Solar
Ma. De los Ángeles Flores González
2023

¿Cómo empezar?

Estimado(a) alumno(a), la "guía de estudio" es una herramienta que te va brindará recursos de estudio, para que así tengas todo el apoyo durante el proceso autodidacta en este sistema de bachillerato no escolarizado. La guía no reemplaza al libro de texto, pero es una herramienta para facilitar el aprendizaje.

Se compone de diferentes secciones:

 Actividades: son ejercicios que podrás llevar a cabo para complementar la lectura de los conceptos clave.

 Recurso: son en su mayoría ligas que te redirigirán a una página de apoyo, puede contener información adicional o ejercicios digitales interactivos.

 Glosario: contiene la definición breve y concisa de algunas palabras que se consideran importantes en la lectura.

 Para reflexionar: este apartado plantea preguntas que desarrollarán tu pensamiento crítico, mediante lecturas, estudios de caso, etc.

Las secciones anteriores construyen tu guía de estudio y son fundamentales, pues están pensadas en función de las competencias a desarrollar de este plan modular; por lo cual te extendemos una amplia invitación a utilizar todos estos elementos para que sean de provecho en este trayecto.

Al finalizar cada unidad habrá una autoevaluación, donde podrás poner a prueba tu conocimiento. Además de servir de refuerzo práctico, te hará saber si estás listo para tu examen del módulo. ¡Mucho éxito!

Índice

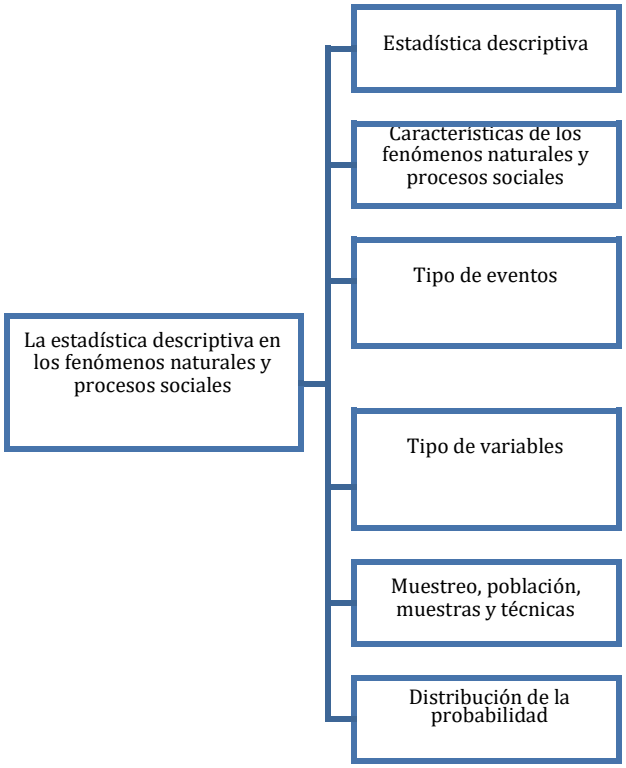
Unidad 1

1. La estadística descriptiva en los fenómenos naturales y procesos sociales

¿Qué voy a aprender y cómo?

En esta unidad aprenderás a usar los métodos estadísticos y el cálculo de las probabilidades, herramientas básicas para el manejo de la incertidumbre generada por la falta de certeza sobre el lugar o el momento de ocurrencia de un fenómeno natural y un proceso social.

Te podrás preguntar, ¿por qué centrarnos en la interpretación y el análisis de los fenómenos naturales y procesos sociales? La razón es simple, porque están presentes en el espacio donde tú y yo nos movemos, donde vivimos, trabajamos, paseamos y realizamos cualquier actividad de convivencia cotidiana; este entorno al que pueden afectar la lluvia, las inundaciones, los terremotos, las erupciones volcánicas, etcétera. En este espacio en el que pueden estar presentes la pobreza, la marginación, el desempleo, la migración, etcétera. Es en la cotidianidad de lo natural y lo social en donde nos desenvolvemos y de la cual no somos ajenos. Su entendimiento conlleva no sólo a la búsqueda de información documental sino a la medición de su impacto a partir de la aplicación de métodos estadísticos y el cálculo de probabilidades. Su conocimiento profundo además de sensibilizarnos, permite nuestra participación de manera colaborativa ante la ocurrencia y el comportamiento de fenómenos naturales y procesos sociales. Para ello, debemos asumir actitudes favorables para el trabajo y la convivencia social.



1.1La Estadística y sus principios básicos

La Estadística es la ciencia matemática que se sirve de los conjuntos de datos para obtener, a partir de ellos, conclusiones basadas en el cálculo de las probabilidades. Es el sustento de pronósticos, diseños experimentales, toma de decisiones y conclusiones basadas en el cálculo de probabilidades. El uso de la Estadística es común en diversas ciencias y situaciones cotidianas, ya que permite obtener información detallada de todo tipo de variables sociales y naturales.

1.2 Fenómenos naturales y sus procesos sociales: sus características

Los fenómenos naturales ocurren en el medio ambiente por acción de factores físicos, sin que el ser humano intervenga, pero su actividad sí los afecta. Los procesos sociales resultan de las interacciones de los seres humanos entre sí y las acciones que de ello resultan.

Fenómenos naturales y procesos sociales en tu contexto

- ¿Cuáles son las características geográficas y sociales más significativas de tu comunidad?

- ¿Qué fenómenos naturales y procesos sociales se presentan con mayor frecuencia?

- ¿Cuáles de ellos la afectan más, tanto de manera positiva como de manera negativa?

- ¿Con qué frecuencia se presentan estos fenómenos y procesos?

Escribe en la siguiente tabla ejemplos de fenómenos naturales y procesos sociales que consideres influyentes en tu contexto

Para reflexionar:

Identifica que tipo de fenómenos naturales y procesos sociales inciden directamente en tu contexto, respondiendo las siguientes preguntas relacionadas con tu entorno.

Ejemplos de fenómenos naturales que por su naturaleza puede influir en tu contexto	Ejemplos de procesos sociales que por su naturaleza pueden influir en tu contexto
1.	1.
2.	2.
3.	3.
4.	4.
5.	5.

Actividad 1.1

Antes de resolver esta actividad, lee el artículo que aparece en la sección de INICIO de la página 25. Ahora para resolver la actividad debes contar con una computadora con procesador de textos y hoja electrónica de cálculo.

1. En la tabla se muestra una lista de fenómenos que por sus características pueden ser ubicados en el entorno geográfico de una comunidad costera. En la columna de la derecha clasifica cada uno de estos como fenómeno natural o proceso social.

Fenómeno	Clasificación
Huracanes	
Inseguridad	
Lluvias	
Escolaridad en la comunidad	
Obesidad en la comunidad	
Inundaciones	
Dengue en la comunidad	
Temperatura	
Pobreza y marginación	
Sequía	
Crecimiento demográfico	
Epidemias	
Migración	
Inmigración	
Enfermedades en la comunidad	
Natalidad en la comunidad	
Servicios de salud en la comunidad	
Virus del Covid 19	

1.3 Tipos de eventos: determinísticos y aleatorios.

Las observaciones en fenómenos naturales y procesos sociales conducen a identificar eventos caracterizados por la **certidumbre** o la **incertidumbre,** ya que se pueden obtener predicciones certeras con pronósticos exactos para cualquier período determinado o eventos que por su ocurrencia incierta son más difíciles de pronosticar. A los primeros se les conoce como **determinísticos** y a los segundos como **aleatorios.**

Fenómenos naturales y procesos sociales determinísticos y aleatorios

Un fenómeno natural o proceso social es definido como **determinístico** si su observación u ocurrencia pueden ser predichas con exactitud.

Un fenómeno natural o proceso social es definido como **aleatorio** si su ocurrencia no puede ser pronosticada con exactitud hasta que ocurra.

La incertidumbre derivada de los eventos aleatorios, tanto en fenómenos naturales como en procesos sociales, es una problemática a la que se enfrentan los investigadores y los lleva a utilizar herramientas que les otorguen elementos para entender y pronosticar a partir del comportamiento de dicha incertidumbre. El conjunto de herramientas lo proporciona la Estadística.

Estadística

Ciencia referente a la recolección, análisis e interpretación de datos, que busca explicar condiciones regulares en fenómenos de tipo aleatorio.

Concepto de probabilidad

La **probabilidad** es la medida que permite cuantificar la oportunidad o posibilidad de que ocurra un evento aleatorio.

Se asigna un valor numérico de 0 a 1, donde cero es la certeza de que no ocurra y 1 es la certeza de que si ocurra. También se maneja en porcentaje donde 100% es la certeza de que el evento ocurra. Por ejemplo, la probabilidad de un evento es de 1 entre 5 entonces,

$$P\% = \frac{1}{5}x100 = 20\%$$

Asignación de la probabilidad de un evento aleatorio

Dividiendo el número de casos favorables (k) entre el número total de casos (n) se obtiene la probabilidad (P) del evento aleatorio(A). $P(A) = \frac{k}{n}\ para\ 0{,}1\ o\ P(A) = \frac{K}{n}x100\ para\ 0{,}100$ (técnica conocida como *enfoque clásico*) esta fórmula aplica para muestras donde todos los eventos tienen la misma probabilidad de ocurrir y se les denomina **eventos equiparables**.

Ejemplo: Se quiere determinar la probabilidad de que llueva una sola vez durante 3 días consecutivos

Evento	Día 1	Día 2	Día 3
1	Llueve	Llueve	Llueve

Partiendo del enfoque clásico, el evento de que llueva una sola vez durante 3 días

2	Llueve	Llueve	No llueve
3	Llueve	No llueve	Llueve
4	No llueve	Llueve	Llueve
5	Llueve	No llueve	No llueve
6	No llueve	Llueve	No llueve
7	No llueve	No llueve	Llueve
8	No llueve	No llueve	No llueve

consecutivos es de 3 y el número total de caos son 8, entonces: $P(A) = \frac{k}{n}x100$

$$P(A) = \frac{3}{8}x\ 100 = 37.5\%$$

Desde el punto de vista intuitivo es lógico pensar que para medir el impacto de un fenómeno natural o un proceso social sobre el entorno solo es necesario estudiar la frecuencia con la que se presenta; se manejan la frecuencia absoluta y la frecuencia relativa.

Frecuencia absoluta

Número de veces que un evento es observado dentro de un período específico o en un espacio geográfico determinado.

Ejemplo 1: en cierta región solo llovió 10 de 30 días, entonces la frecuencia absoluta es igual a 10.

Frecuencia relativa

Representa el cociente entre la frecuencia absoluta de un evento y el número total de eventos observados dentro de un período de tiempo o un espacio geográfico determinado.

Siguiendo con el ejemplo anterior, la frecuencia relativa (Fr) resulta de dividir la frecuencia absoluta (Fa) que en este caso es 10 entre el número de días del fenómeno observado que es 30.

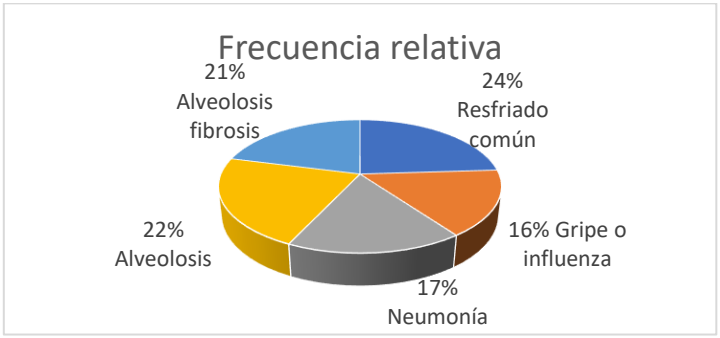
$$Fr = \frac{Fa}{n}x100 \qquad Fr = \frac{10}{30} = 0.333 \qquad Fr = \frac{10}{30}x100 = 33.3\%$$

Ejemplo 2: Para prevenir contagios y transmisión de enfermedades en el período invernal, se estudia un registro de una comunidad rural de personas de más de 70 años.

El registro del hospital comunitario se presenta en la primera tabla, así como las frecuencias del evento en la segunda tabla.

Enfermedad	Número de casos registrados (n)
Resfriado común	36
Gripe o influenza	24
Neumonía	26
Alveolosis	33
Alveolosis fibrosa	31
Número total de casos "N"	150

Enfermedad	Frecuencia absoluta	Frecuencia relativa
Resfriado común	36	0.24
Gripe o influenza	24	0.16
Neumonía	26	0.17
Alveolosis	33	0.22
Alveolosis fibrosa	31	0.21
Número total de casos "N"	150	1



Ejemplo: Prevalencia del resfriado común

$N=150$

$n= 36$

$$Fr = \frac{36}{150} x100 = 24\%$$

Contar con registros de la ocurrencia de eventos permite medir su impacto.

Probabilidad Condicional

Es aquella que depende de que se haya cumplido otro hecho relacionado. Si tenemos un evento, que denominamos A, condicionado a otro evento, al cual denominamos B, la notación sería P(A | B) y la fórmula sería la siguiente:

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

Ejemplo:

Calcular la probabilidad de obtener un 6 al tirar un dado sabiendo que ha salido par.

Solución:

$$P(6|par) = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Las propiedades de la probabilidad condicional son las siguientes:

P (A | B) + P($\underline{A}$ | B) = 1

Esto significa que la probabilidad de A dado B, más la probabilidad del complemento de A (los elementos del universo que no pertenece a A) dado B, es igual a 1.

B ⊆A → P (A | B) = 1

Esta propiedad implica que si A es un subconjunto de B (o son dos conjuntos iguales), la probabilidad de que ocurra A dado B es 1.

$$P(A) = P(A \mid B) \cdot P(B) + P(A \mid \underline{B}) \cdot P(\underline{B})$$

Lo anterior quiere decir que la probabilidad de A es igual a la probabilidad de A dado B por la probabilidad de B más la probabilidad de A, dado el complemento de B por el complemento de B.

Ejemplos:

Una clase consta de seis niñas y 10 niños. Si se escoge un comité de tres al azar, hallar la probabilidad de:

a) Seleccionar tres niños

-Primero se sabe que en total son 16 niños

-Los casos favorables en niños serían 10.

$$P(3\ niños) = \frac{10}{16} * \frac{9}{15} * \frac{8}{14} = 0.214$$

Se resta un número dado que sale un niño en cada ronda imaginando que salió niño.

b) Seleccionar exactamente dos niños y una niña

$$P(2\ niños\ y\ 1\ niña) = \frac{10}{16} * \frac{9}{15} * \frac{6}{14} + \frac{10}{16} * \frac{6}{15} * \frac{9}{14} + \frac{6}{16} * \frac{10}{15} * \frac{9}{14} = 0.214$$

¡Ahora te toca a ti!

Actividad 1.2

Retoma la lista de fenómenos naturales y procesos sociales de la actividad 1 y resuelve la siguiente actividad.

1. Clasifica la lista de fenómenos naturales y procesos sociales en determinísticos y aleatorios, (4 en cada caso) argumentado las razones que te llevaron a realizar dicha clasificación.

Fenómeno	Clasificación

negro	azul	amarillo	rojo	azul
azul	rojo	negro	amarillo	rojo
rojo	amarillo	amarillo	azul	rojo
negro	azul	rojo	negro	amarillo

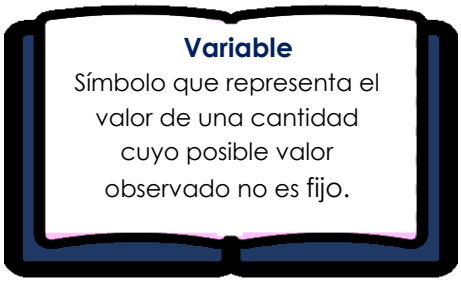
2. Se le pidió a un grupo de personas que indicarán su color favorito, y se obtuvieron los siguientes resultados:

Con los resultados obtenidos, elabora una tabla de frecuencias.

Color	Frecuencia absoluta	Frecuencia relativa
Negro		
Azul		
Amarillo		
Rojo		
Total		

1.4 Las variables como instrumento de medición de la incertidumbre y la variabilidad de un fenómeno aleatorio.

A las características de los fenómenos naturales y los procesos sociales, también se les llama variables y éstas son objeto de estudio y medición.



Se observa que los fenómenos de naturaleza aleatoria presentan características que varían en los distintos momentos de espacio o tiempo en el que son medidas, debido a la incertidumbre relacionada con el fenómeno observado. Por ejemplo, los niveles de precipitación pluvial o el analfabetismo varían en distintos momentos de tiempo y espacio. La incertidumbre derivada de la aleatoriedad de estos fenómenos dificulta la medición por lo que los **Los métodos estadísticos** se vuelven imprescindibles. Dichos métodos dependen de las

características métricas que son los elementos que forman parte de los objetos y sujetos que pueden ser sujetos de medición.

Variables cuantitativas

En el estudio de los fenómenos naturales y procesos sociales es común encontrarse con **variables continuas**. Las variables continuas cumplen con la propiedad de que entre dos valores observados hay una infinidad de posibles valores observables. Siempre es posible identificar el rango de variación de todos sus valores posibles donde este puede llegar a ser un rango continuo infinito entre dos datos y a dicho rango se le conoce como intervalo de variación, de ahí que a las variables continuas también se les conoce como **variables intervalares**. Como ejemplos de variables continuas se tienen el peso y la estatura de una persona, porcentaje de personas desempleadas en una comunidad, cantidad de energía eléctrica consumida por una comunidad en un intervalo de tiempo. En otro tipo de eventos se identifican las **variables discretas** que son aquellas que solo pueden adoptar valores numéricos enteros, de tal modo que entre dos valores consecutivos hay por lo menos un valor no observable, que carecen de sentido y toman solo un número finito de valores. Algunos ejemplos son el número de personas obesas en una comunidad, número de personas que padecen de cierta enfermedad, número de personas morosas.

Variables cualitativas

El estudio del nivel socioeconómico es claramente una variable cualitativa ya que se clasifica como bajo, medio y alto por lo que su naturaleza no es numérica, entonces se habla de una **variable categórica**, aquella cuyos niveles de respuesta son de tipo cualitativo como: color, raza, credo, filiación política, nivel educativo, entre otros.

Variables categóricas nominales y ordinales

Aún y cuando la naturaleza de las variables categóricas no es numérica, para fines de estudio estadístico es necesario asignar un valor numérico a cada una de sus niveles de respuesta. Por ejemplo, para el nivel socioeconómico jerárquico bajo, medio y alto, se pueden asignar a criterio, los valores 1,2 y 3. Siendo así el caso, a la variable categórica se le denomina ordinal.

Cuando la variable categórica no es de tipo jerárquico, pero igual se les asigna un número a sus categorías, entonces se dice que la variable es **nominal**, por ejemplo, cuando se quiere determinar el índice de deserción escolar considerando la variable asociada con la escuela de procedencia que pudiera explicar el fracaso escolar (bachillerato tecnológico y comercial, público o privado). Las escuelas no tienen una jerarquía.

Clasificación de las variables por su grado de asociación.

Hipótesis: Supuesto al que se le otorga un cierto grado de posibilidad para extraer de ello un efecto o consecuencia.

Los vínculos entre los fenómenos naturales y los procesos sociales permiten el planteamiento de **hipótesis** sobre las posibles relaciones de causa y efecto que pudieran existir entre las variables asociadas, dando lugar a la identificación de las **variables independientes** y las **variables dependientes**. La variable dependiente se trata del factor que se ve modificado o influenciado por una variable independiente. Mientras que la variable dependiente son los factores que el investigador quiere poner a prueba para demostrar una hipótesis.

Un ejemplo sería el incremento de las enfermedades respiratorias (efecto) debido a la disminución de la temperatura (causa). En dicho ejemplo, la variable dependiente es la prevalencia de las enfermedades respiratorias y la variable independiente es la temperatura mensual/anual.

Actividad 1.3

Identifica variables cualitativas y cuantitativas de la siguiente tabla.

Variables identificadas con eventos aleatorios relacionados con el entorno de la comunidad de Santiago
Sexo de los habitantes de Santiago.
Escolaridad de los habitantes mayores de 18 años.
Temperatura media anual en la comunidad.
Tipo de enfermedades más frecuentes.
Porcentaje anual de habitantes que se contagiaron de dengue durante el verano.
Porcentaje de habitantes que se contagiaron de alguna enfermedad respiratoria durante el invierno.
Número de altas semanales en la cínica de la comunidad.
Número diario de camas disponibles en la clínica de la comunidad.
Porcentaje de familias cuyo sustento económico es obtenido en actividades derivadas de la comercialización de productos del mar.
Número anual de huracanes y tormentas tropicales que han impactado a la comunidad.
Porcentaje anual de la población que ha emigrado hacia la comunidad.
Porcentaje anual de la población que ha emigrado fuera de la comunidad.
Índice de analfabetismo en la comunidad.
Número diario de personas detectadas con diabetes en el hospital de la comunidad.

Presión sanguínea sistólica registrada en las personas que acuden a la clínica de la comunidad.
Tipo de accidente por el que ingresan los pacientes a la sala de emergencia en la clínica de la comunidad.

1. Con base en las características del entorno en el cual se encuentra ubicada la comunidad de Santiago, en el siguiente cuadro se presenta un conjunto de variables relacionadas con la naturaleza del entorno de esta comunidad.

Variables cuantitativas		Variables cualitativas	
Continuas	Discretas	Ordinales	Nominales

A partir de las variables registradas, clasifícalas en cuantitativas (continuas o discretas) y cualitativas (ordinales y nominales).

Utiliza la tabla.

Variables independientes	Variables dependientes

2. De las variables que clasificaste en la tabla anterior identifica aquellas que puedan tener una relación de causa-efecto. Utiliza la siguiente tabla para registrar cada variable independiente identificada con su correspondiente variable dependiente.

1.5 Muestreo: población, muestra y técnica.

Técnicas de muestreo y su importancia en el análisis de la información.

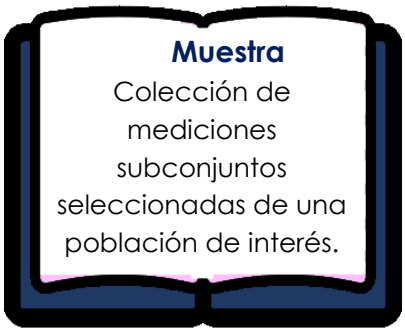
Las **fuentes de información** son el insumo sin el cual el estudio de los fenómenos naturales y sociales no podría llegar a buen término.

Ya que se identificaron las fuentes de información, el paso natural es la **recolección de mediciones o datos** que permitan estudiar el comportamiento del fenómeno.

Censos y muestras.

Una de las fuentes más confiables para obtener información es el **censo**, que se define como un método de recolección de datos que implica la medición o recopilación de información de todos los miembros de una población; es decir el conjunto de todos los elementos o individuos que son objeto del estudio que verifican una característica o varias.

En algunas ocasiones los recursos económicos y el tiempo son limitados por lo que se toma una muestra del total de la población, a esta alternativa se le conoce como **método de muestreo.**



La información permite al investigador obtener datos confiables analizando la **muestra** de una población.

Concepto de muestra

Una muestra es un subconjunto extraído de la población mediante la aplicación de algún método de muestreo que permite obtener conclusiones válidas para la población donde sean confiables. Al momento de tomar una muestra de la población, es importante evitar el sesgo.

El **sesgo** indica una diferencia entre el valor estimado y el valor real. Es un error que puede llevar a conclusiones incorrectas, se da como resultado de una tendencia o inclinación que puede llevar a tomar solo los datos que comprueban la hipótesis, dejando de lado los datos en contra de ella.

Métodos de muestreo

El método más sencillo es el **muestreo aleatorio simple**.

Muestreo aleatorio simple

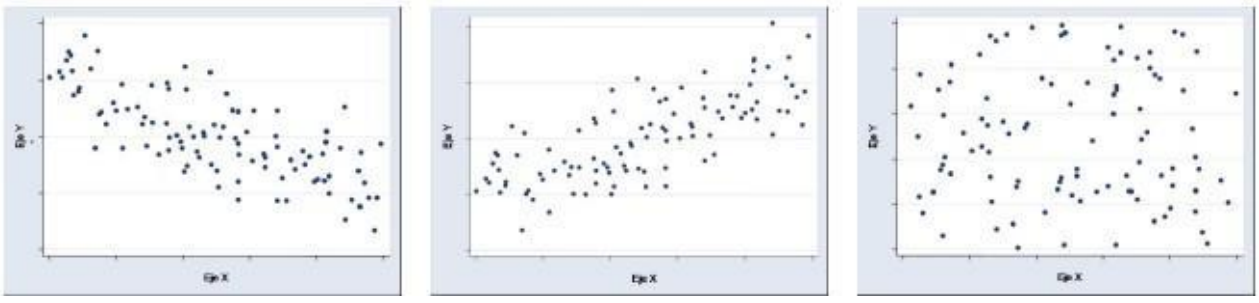
Es en el que todas las muestras tienen la misma probabilidad de ser seleccionadas.

Muestra aleatoria simple

Si "n" es el número de objetos o individuos, entonces en una muestra aleatoria simple se asegura que todas las muestras posibles diferentes del tamaño deseado, tienen la misma probabilidad de ser seleccionados.

Selección de una muestra aleatoria simple

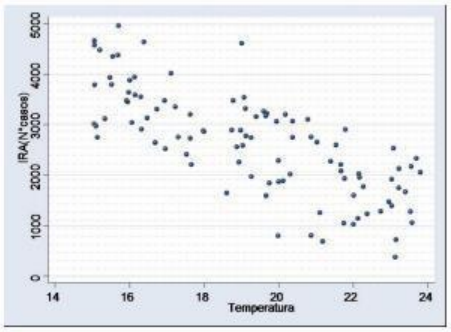
Un método muy utilizado para la obtención de una muestra aleatoria simple es primero hacer una lista, donde cada elemento es identificado con un número elegido de manera aleatoria. Después se utiliza una técnica exploratoria que grafica las posibles. Relaciones de dependencia entre dos variables. Conocida como **diagrama o gráfico de dispersión**, esta técnica utiliza un conjunto de registros apareados de la forma (X, Y), donde X es la variable independiente y la variable dependiente, graficadas en el plano cartesiano. La forma que tome la nube de puntos resultante, indicará el grado de asociación entre las variables. Las gráficas muestran las 3 posibilidades de asociación entre variables que pueden ser identificadas a partir de este método.



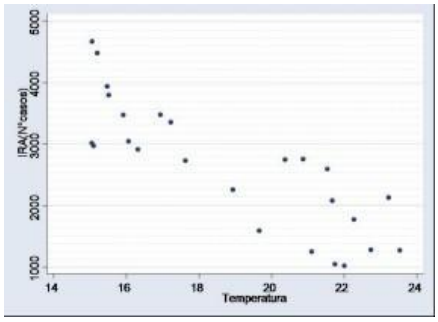
Gráfica 1: el valor de la variable Y disminuye conforme aumenta X.
Gráfica 2: el valor de la variable Y aumenta conforme aumenta X.
Gráfica 3: no hay relación entre las variables.

Ejemplo: Un estudiante cuenta con 100 registros de temperatura y número asociado de casos de infecciones respiratorias agudas (IRA), recolectadas en
100 semanas consecutivas. Elabora un diagrama de dispersión y obtiene la siguiente gráfica.

Se observa que, a mayor temperatura, menor la incidencia del IRA.



Por último, el estudiante aplica un muestreo aleatorio simple, donde elige una muestra aleatoria simple de tamaño n= 25 a partir de la población de 100 registros apareados de temperatura y número de casos de infecciones respiratorias agudas (IRA), obteniendo la siguiente gráfica:



Se observa que, a pesar de contar con un menor número de observaciones, los datos muestreados representan adecuadamente la tendencia observada en la gráfica de los datos totales. El muestreo aleatorio simple es muy útil porque en muchos de los casos no se cuenta con la información completa de la población. Dicho método nos da la confianza, pero no la certeza de haber producido una muestra representativa de la población.

Muestreo aleatorio estratificado

Cuando toda la población puede ser dividida en subgrupos, el método a utilizar es el **muestreo aleatorio estratificado**, en el que se seleccionan diferentes muestras aleatorias de cada subgrupo, de manera independiente. A cada subgrupo se le llama **estrato**. En el muestreo se selecciona una muestra aleatoria de cada estrato.

Un muestreo estratificado permite hacer **inferencias**, es decir, ligar dos o más proposiciones del estudio de una población.

Muestreo de conveniencia.

Es el muestreo que se presenta cuando se solicita una muestra de respuesta voluntaria, cuya información generalmente no es representativa de la población.

¡Ahora te toca a ti!

Comprende la utilidad del diagrama de dispersión para apoyar la verificación de hipótesis planteadas con respecto a la posible relación de causa-efecto entre 2 variables.

Actividad 1.4:

Elabora una gráfica de dispersión e identifica los tipos de muestreo en los ejemplos que se presentan a continuación.

Cantidad de tinta (Litros)	Número de errores
0,47	16
0,48	14
0,69	30
0,7	31
0,59	15
0,59	17
0,37	10
0,62	21
0,39	11
0,35	13
0,68	28
0,52	17
0,42	11
0,51	18
0,5	19
0,34	10
0,41	14
0,3	7
0,53	20
0,33	7
0,36	11
0,4	16
0,4	13
0,69	24
0,61	24
0,32	9
0,66	28
0,64	23
0,45	17
0,59	20
0,6	21
0,56	19
0,6	20
0,55	18
0,44	15
0,49	16
0,63	25
0,65	26
0,38	10
0,67	24

Una litográfica se encuentra haciendo todos los ensayos y pruebas para determinar la cantidad de tinta de cada color que deberían tener las máquinas para la impresión de posters.

Como prueba inicial, han decidido establecer la relación de errores de impresión según el grado de llenado de los recipientes de tinta de la máquina.

Las variables a estudiar para este ejemplo de gráfico de dispersión en calidad son:

- Cantidad de tinta en litros
- Número de errores de impresión.

Al estar el número de errores influenciado por la cantidad de tinta, lo ubicamos como el eje y. Por consiguiente, el eje x es la cantidad de tinta. Ahora sí, hacemos el gráfico de dispersión.

Puedes utilizar papel milimétrico para obtener mejor resultado en la elaboración de la gráfica.
En caso de que te sea posible utilizar internet puedes utilizar un simulador o un programa de Excel.

Un gobierno quiere estudiar el ingreso promedio por núcleo familiar.
Se selecciona a todas las familias que habitan en Ciudad de México.
Se determina que se necesitarán 1.000 elementos para hacer la muestra.
Se elabora una lista de las
2.756.319 familias que habitan en Ciudad de México.
A cada familia se le asigna un número comenzando por el 1 hasta llegar al 2.756.319. Se escogen 1.000 números al azar utilizando un programa

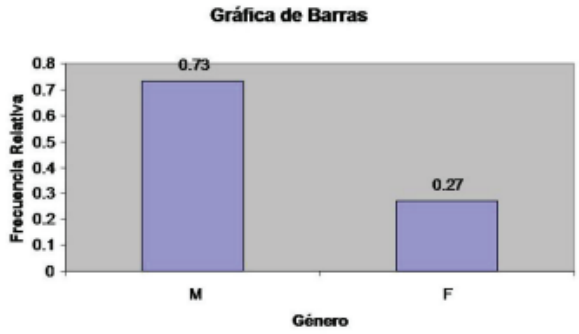
Un equipo de investigadores está analizando las opiniones de la población sobre la reforma de una ley. Para seleccionar la muestra representativa, se divide la población en cuatro grupos etarios (de 18 a 30 años, de 30 a 45 años, de 46 a 60 años y mayores de 60 años) y después se escoge a los individuos de manera aleatoria.

1.6 Las distribuciones de probabilidad y su relación con las variables relacionadas con fenómenos naturales y procesos sociales

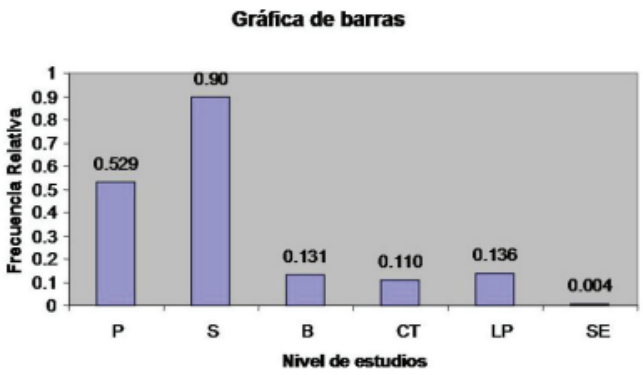
Distribuciones de frecuencias para variables categóricas

Al realizarse un estudio demográfico sobre equidad y género en la economía familiar, por ejemplo, los investigadores aplican la **distribución de frecuencias**, entre la población de los hogares, de tal manera que el impacto sea medido a partir de la frecuencia relativa que ambas respuestas tienen entre el número total de hogares.

La distribución de frecuencias se resume en una gráfica de barras, cuya frecuencia relativa (Frecuencia absoluta/total de observaciones) o proporción queda presentada por la altura de cada rectángulo.
Suponiendo que el investigador encontró que el 0.73 o 73% del género masculino y el 0.27 o 27% del género femenino, sostienen su hogar, la gráfica resultaría así:



Ejemplo 2: Ahora se pretende conocer el nivel de escolaridad del jefe de familia (sin importar su género), bajo las siguientes categorías; primaria (P), secundaria (S9, bachillerato (B) carrera técnica (CT), licenciatura y posgrado (LP) y sin estudios (SE)
De un hogar seleccionado al azar, la probabilidad de frecuencia de estudios de bachillerato es P(x=bachillerato) =0.110 o 11.0%



1.7 Histograma con distintas variables

Histograma de frecuencias

Se utiliza un histograma de frecuencias para graficar los datos de variables numéricas continuas y discretas.

Histograma para variantes numéricas discretas

En una muestra aleatoria de 30 mujeres de una población rural se encontraron los siguientes resultados referentes al número de hijos:

Número de hijos													
12	2	4	6	6	7	8	7	8	11	8	3	5	6	7
10	1	9	7	6	9	7	5	4	7	4	6	7	8	10

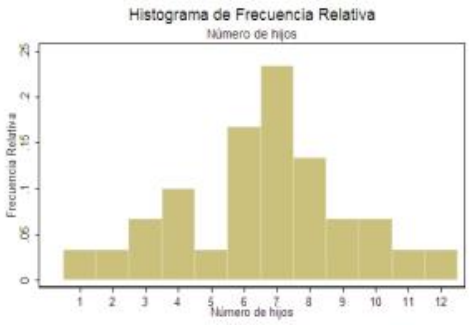
A partir de estos datos se generó una tabla de distribución de frecuencias:

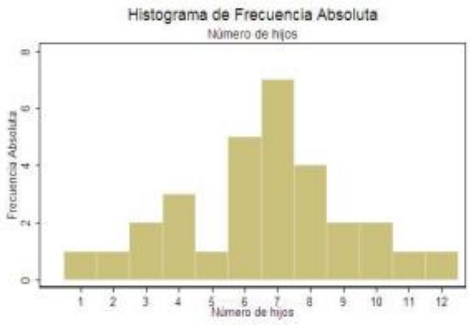
Distribución de frecuencias para el número de hijos		
Número de hijos	Frecuencia absoluta	Frecuencia relativa
1	1	0.033
2	1	0.033
3	1	0.033
4	3	0.100
5	2	0.067
6	5	0.167
7	7	0.233
8	4	0.133
9	2	0.067
10	2	0.067
11	1	0.033
12	1	0.033

Se pueden agrupar los datos y compactar la distribución de frecuencias:

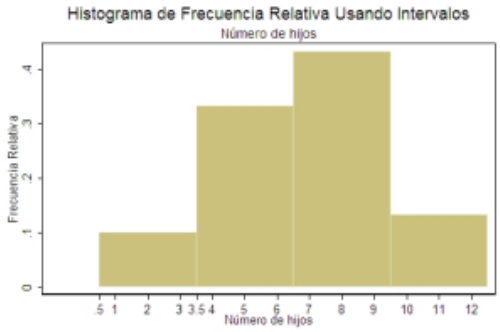
Distribución de frecuencias del número de hijos usando intervalos		
Número de hijos	Frecuencia absoluta	Frecuencia relativa
1-3	3	0.100
4-6	10	0.333
7-9	13	0.433
10-12	4	0.133

La representación para para la primera tabla de frecuencias de los histogramas de la frecuencia relativa y la frecuencia absoluta son los siguientes:





La representación de la segunda tabla de frecuencias agrupadas queda de la siguiente forma:



Histograma para una variable continua

Construir un histograma para una característica continua puede ser casi imposible por lo que se crean categorías con intervalos conformados convenientemente y donde cada intervalo es limitado por los símbolos **a** o **que**.

Ejemplo: El gobierno de un estado planea la distribución de recursos para la educación en los 50 municipios que lo conforman y donde consideran el porcentaje de estudiantes inscritos en escuelas públicas. Los datos recabados se presentan en la tabla.

Porcentaje de estudiantes matriculados en instituciones de educación pública por municipio										
96	86	81	84	77	90	73	53	90	93	76
86	78	76	88	86	87	64	89	86	80	66
70	90	89	82	73	72	56	55	75	77	82
83	79	75	43	50	64	80	82	75	96	60
81	59	73	58	73	59					

La observación más pequeña es de 43 y la más grande de 96, entonces el primer intervalo es 40 y el mayor es de 100. Con intervalos de 10 en 10, se obtienen las siguientes frecuencias, considerando el que:

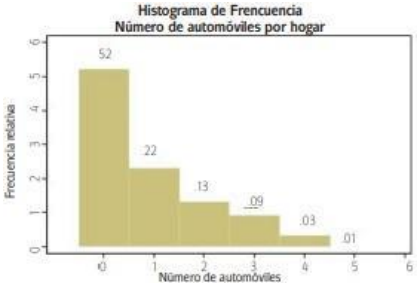
Distribución de frecuencias para el porcentaje de estudiantes matriculados en instituciones de educación pública por municipio		
Intervalos de clase	Frecuencia absoluta	Frecuencia relativa
40 a <50	1	0.02
50 a <60	7	0.14
60 a <70	4	0.08
70 a <80	15	0.30
80 a <90	17	0.34
90 a <100	6	0.12

Cuando existe una gran cantidad de datos, una distribución basada en 15 o 20 (o más) intervalos de clase se estima el número apropiado de intervalos bajo el siguiente criterio

$\sqrt{número\ de\ observaciones}$ la cantidad es utilizada como un estimador apropiado. De 25 observaciones $\sqrt{25} = 5$, de 100 observaciones, $\sqrt{100} = 10$ y así sucesivamente.

Distribuciones de frecuencias para variables numéricas discretas.

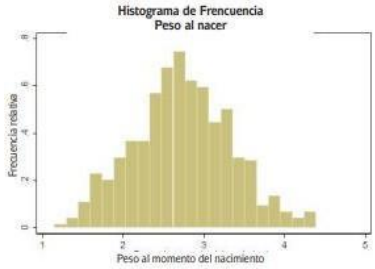
Ejemplo: en un estudio socioeconómico se quiere conocer la posibilidad que tienen los habitantes de una comunidad para contar con un automóvil particular. Se considera la variable x = número de automóviles por hogar que va de 0 a 5. Utilizando el histograma de frecuencias relativas, se obtiene la siguiente gráfica:



El valor más común de la población es x=0. En términos probabilísticos la probabilidad asociada con x=0 es de P(x=0) =0.52, esto es, 52% de los hogares no cuenta con automóvil y solo el 1% cuenta con 5 automóviles.

Distribución de frecuencias para variables numéricas continuas

Estudiar el peso (kg) al nacer para todos los bebés de una localidad, durante el año 2010 es un ejemplo de una variable continua. Los valores de **x** en la construcción de un histograma representan el rango de todos los posibles pesos al momento de su nacimiento.



1.8Distribuciones de probabilidad

Distribuciones de la probabilidad continua

Si en el ejemplo anterior se toma un mayor número de intervalos, la consecuencia es que se aumenta el número de observaciones de la variable **x** y las longitudes de los intervalos se vuelven más estrechas.

Se observa que los histogramas tienden a parecerse más a una curva suave.

A esta curva se le conoce como **distribución de probabilidad continua**. Además, se observa que se conservan las características importantes de la población (forma general, centro, rango de variación de sus valores, etc.). A la curva se le llama **curva de densidad** y tiene las siguientes propiedades:

- El área total bajo la curva es igual a1
- El área bajo la curva y por encima de cualquier intervalo particular se interpreta como la probabilidad de observar un valor en el intervalo correspondiente, cuando una persona u objeto es seleccionado al azar de la población.

Distribuciones de las probabilidades teóricas

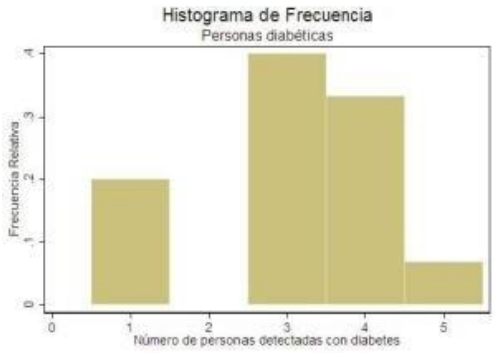
Así como el caso del peso de los recién nacidos, existen en la naturaleza fenómenos cuyas mediciones continuas tienden a comportarse de esta manera.

Una distribución teórica que formaliza este comportamiento continuo es la **distribución de probabilidad normal** ya que en ella la curva suave tiene forma de campana y es simétrica.

Distribución binomial

La **distribución binomial** nos permite encontrar el porcentaje en que es probable obtener un resultado entre dos posibles al realizar un número ***n*** de pruebas. La probabilidad de cada posibilidad no puede ser más grande que 1 y no puede ser negativa. Sirve para modelar variables discretas teóricas. Estudiaremos un ejemplo.

En un hospital comunitario se afirma que la prevalencia de diabetes entre los individuos de la comunidad es del 40%. Durante 15 días consecutivos se selecciona una muestra aleatoria de 6 personas a las que se les realizó la prueba de diabetes y determinando el número de personas con el resultado positivo. La variable se **discretiza** al definir como 0= no enfermo y 1= enfermo, obteniendo la variable discreta numérica X= 0.1

Al observar la distribución de frecuencias representada en el histograma, se concluye que ésta otorga evidencia que refuta el planteamiento, pues se puede observar que para la mayoría de las 15 muestras el dato observado más frecuente fue el de tres pacientes detectados con la enfermedad, entonces sí de cada seis pacientes diagnosticados, en la mayoría de los casos se observaron tres pacientes detectados con el padecimiento (50%), hay suficiente evidencia para afirmar que la prevalencia real del padecimiento entre la población es del 50% y no del 40% como le habían informado	

Revisa las páginas 72 y 73 de tu libro de texto donde se analizan nuevos experimentos binomiales que reafirman los resultados del experimento anterior y enfatizan la importancia y utilidad de la distribución binomial.

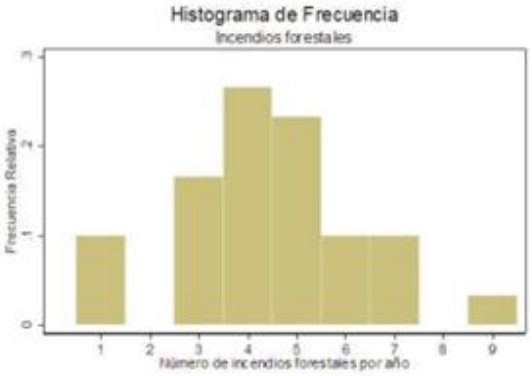
Distribución de Poisson

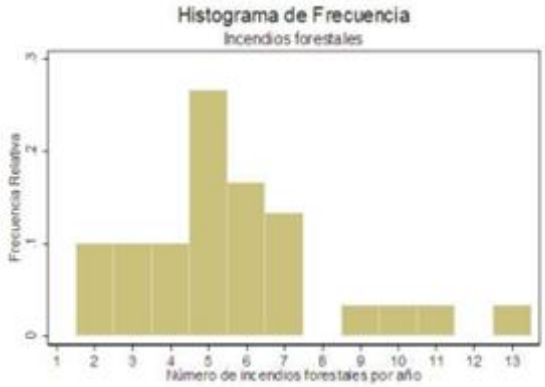
Esta **distribución de probabilidades** es muy utilizada para situaciones donde los sucesos son impredecibles o de ocurrencia aleatoria. Hace referencia a los sucesos relacionados con el espacio o el tiempo y satisface 3 condiciones:

- El número promedio de veces que ocurre un evento por unidad de tiempo o de espacio es constante.
- La probabilidad de más de un suceso en una unidad de tiempo o espacio es muy pequeña.
- El número de acontecimientos en intervalos ajenos de tiempo o espacio son

independientes unos de otros.

Un investigador forestal está alarmado porque debido al calentamiento global y la deforestación la tasa media anual de incendios forestales se ha incrementado en los últimos 30 años. Decide comparar los registros anuales del número de incendios ocurridos en los 30 años previos al año en que sospecha comenzó el incremento en el número de incendios. Define a la variable "x" como el número total de incendios observados en la región durante un periodo de un año y elabora un histograma de frecuencias de esta variable para los dos periodos considerados. Él observa en ambos histogramas que hubo un incremento en la aparición de incendios y que el número de estos alcanzó la cifra record de 10 o más por año. Después de su análisis concluye que tanto el calentamiento global como la deforestación influyeron para el aumento de los incendios, en la región, en los últimos 30 años. Si suponemos que el número promedio de veces que ocurrió un incendio durante un año permaneció constante, la probabilidad de haber observado uno en un periodo reducido es muy pequeña y el número de incendios en intervalos ajenos de tiempo (intervalos de tiempo que no se traslapan) son independientes unos de otros. La opción más pertinente para representar el fenómeno estudiado es el experimento de Poisson.





¡Ahora te toca a ti!

Ya conoces las características que ciertas variables deben cumplir para poder ser modeladas por la distribución de probabilidades teóricas: normal, binomial y de Poisson. Ahora tu tarea será utilizar las variables identificadas en la actividad tres y determinar si por sus características algunas de éstas pueden ser modeladas por algunas de las distribuciones teóricas descritas.

Actividad 1.5

Responde adecuadamente a cada una de las siguientes preguntas.

1. Durante el mes de julio, en una ciudad se han registrado las siguientes temperaturas máximas:

$$32, 31, 28, 29, 33, 32, 31, 30, 31, 31, 27, 28, 29, 30, 32, 31, 31,$$

$$30, 30, 29, 29, 30, 30, 31, 30, 31, 34, 33, 33, 29, 29.$$

a) Crea la tabla de frecuencias, esta debe conformarse por 4 columnas: en la primara los datos de la variable ordenados de menor a mayor; en la segunda, la frecuencia absoluta (cuántas veces aparece cada dato); la tercera columna representa la frecuencia acumulada (las frecuencias absolutas anteriores más la actual); por último, la frecuencia relativa (frecuencia absoluta dividido entre el total de variables).

2. Se registran los tiempos de las llamadas recibidas en un *call center*, y se obtiene la siguiente tabla de frecuencias con datos agrupados. Construir un histograma de frecuencia absoluta. Puedes utilizar frecuencia acumulada o porcentual para elaborar otros histogramas.

Tiempo de llamadas	Marcas de clase	Frecuencia absoluta	Frecuencia acumulada	Frecuencia porcentual
[0 - 10)	5	2	2	5%
[10 - 20)	15	6	8	15%
[20 - 30)	25	12	20	30%
[30 - 40)	35	10	30	25%
[40 - 50)	45	6	36	15%
[50 - 60]	55	4	40	10%
Total		**40**		**100%**

3.- Indica si se trata de una variable discreta o continua:

a). Longitud de 150 tornillos producidos en una fábrica. b). Número de pétalos que tiene una flor.
c). Tiempo requerido para responder las llamadas en un call center.
d). Número de páginas de una serie de libros de estadística.
e). Lugar que ocupa un nadador en una competencia.

4. ¿Cuáles son las similitudes y las diferencias de la distribución nominal y la distribución de Poisson?

Autoevaluación Unidad 1

I. Lee con atención la siguiente situación y selecciona la opción correcta.

Eres habitante de una ciudad costera cuya población es de 1 millón de habitantes, aproximadamente. La ciudad está enclavada en un corredor turístico ubicado en la costa del mar Caribe. Por su ubicación y desarrollo económico es un polo de atracción para habitantes de otras regiones y del extranjero. Además, sus condiciones climáticas posibilitan condiciones laborales y salariales que permiten una calidad de vida aceptable.

Identifica cuáles de los siguientes fenómenos pueden influir el entorno geográfico de la ciudad:

1. Turismo
2. Pesca
3. Huracanes
4. Emigración
5. Inmigración
6. Tormentas tropicales
7. Tsunamis
8. Marginación
9. Desarrollo económico
10. Comercio
11. Inundaciones
12. Epidemias

1. ¿Cuáles de los siguientes fenómenos descritos pueden ser clasificados como fenómenos naturales?

a) 1, 2, 5, 8 y 10
b) 3, 6, 7, 11 y 12
c) 1, 2, 8, 9 y 10
d) 4, 5, 8, 11 y 12

2. ¿Cuáles de los siguientes fenómenos, que son característicos de la región descrita en la situación anterior, pueden clasificarse como procesos sociales?

a) 1, 2, 4, 5, 8, 9 y 10
b) 3, 6, 7, 8, 10, 11 y 12
c) 2, 3, 4, 8, 10, 11 y 12
d) 1, 4, 5, 6, 7, 9 y 11

Los siguientes son eventos que por su naturaleza ocurren o pudieran ocurrir en el entorno geográfico de la ciudad descrita:

1. Día y noche.
2. Niveles de precipitación pluvial durante el verano.
3. Estaciones del año.
4. Niveles de ocupación hotelera durante un año.
5. Pérdidas económicas anuales.
6. Número de defunciones anuales.
7. Número de nacimientos anuales.
8. Generación de empleos anual.
9. Percepción mensual de un trabajador asalariado.

Con base en la lista de eventos responde las preguntas que se formulan a continuación:

3. ¿Cuáles de los siguientes eventos que por su naturaleza ocurren o pudieran llegar a ocurrir en el entorno geográfico descrito pueden ser clasificados como eventos determinísticos?

a) 2, 4 y 6
b) 4, 7 y 9
c) 3, 6 y 7
d) 1, 3 y 9

4. ¿Cuáles de los siguientes eventos que por su naturaleza ocurren o pudieran llegar a ocurrir en el entorno geográfico descrito pueden ser clasificados como eventos aleatorios?

a) 1, 2, 3, 4, 5 y 6
b) 3, 5, 6, 7, 8 y 9
c) 2, 4, 5, 6, 7 y 8
d) 1, 3, 6, 7, 8 y 9

II. Un investigador está interesado en realizar un estudio que le permita identificar las características ambientales que propician la propagación de parásitos que provocan enfermedades gastrointestinales. Así mismo, pretende recopilar información del entorno social con la finalidad de identificar los hábitos y las conductas sociales que propician la diseminación de estos padecimientos. También busca identificar las características de la población que resulta más afectada por la prevalencia de estas enfermedades. El investigador cuenta con un protocolo de investigación cuyos elementos metodológicos principales puedes identificar a partir de resolver acertadamente los siguientes reactivos:

El investigador en su protocolo de investigación identificó cuatro variables de interés:

a) Ingreso per cápita.
b) Nivel socioeconómico.
c) Tipo de padecimiento gastrointestinal.
d) Número de pacientes con algún tipo de padecimiento gastrointestinal.

Utilizando el siguiente criterio de clasificación de variables:

1. Variable numérica continua.
2. Variable numérica discreta.
3. Variable categórica ordinal.
4. Variable categórica nominal.

5. Selecciona la clasificación correcta de entre los cuatro posibles criterios de clasificación mostrados a continuación.

a) 1a; 2b; 3c; 4d
b) 1a; 2d; 3b; 4c
c) 1c; 2d; 3a; 4c
d) 1b; 2c; 3a; 4d

1. Población	a) Recuento de una población con la finalidad de conocer cuestiones inherentes a ella.
2. Método de muestreo	b) Subconjunto extraído de la población mediante la aplicación de algún método de muestreo, que permite obtener conclusiones válidas que se pueden generalizar hacia toda la población.
3. Censo	c) Conjunto de todos los elementos o individuos que son objeto de estudio.
4. Muestra	d) Técnicas basadas en la recolección de una fracción de la información de la población objeto de estudio.
5. Muestra Representativa	e) Es una muestra seleccionada mediante un método que asegura que cada muestra posible del tamaño deseado tiene la misma oportunidad de ser elegida.

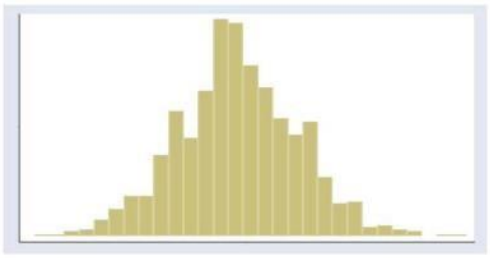
6. Para obtener la información el investigador debe decidir aplicar un muestreo o un censo. No puede realizarlo sin conocer los conceptos propios de estas metodologías. Identifica cuáles de los siguientes son propios de ellas relacionando el con su definición. Selecciona entre las cuatro opciones de respuesta desarrolladas al final la que consideres es la correcta.

a) 1a; 2b; 3c; 4d; 5e
b) 1c; 2e; 3a; 4d; 5b
c) 1c; 2d; 3a; 4b; 5e
d) 1e; 2d; 3c; 4b; 5a

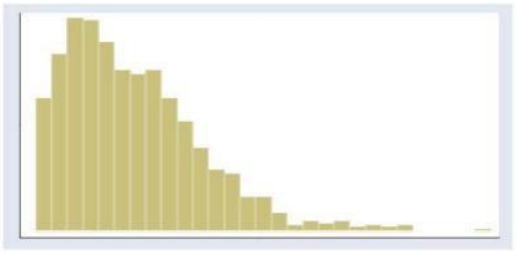
7. Con base en la información obtenida a partir de las mediciones de la variable "**ingreso per cápita**", el investigador realiza un histograma de frecuencia. La forma del histograma le permite concluir que el modelo de la distribución de probabilidad normal es la opción de modelo teórico adecuado para modelar esta variable continua. Selecciona entre las siguientes cuatro gráficas el histograma que más se asemeja a una distribución normal y lleva al investigador a seleccionar el modelo teórico.

En las siguientes cuatro gráficas el histograma que más se asemeja a una distribución normal y llevó al investigador a seleccionar el modelo teórico.

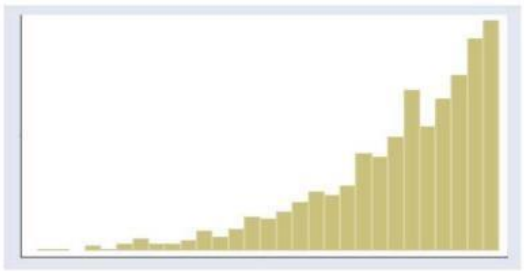
1



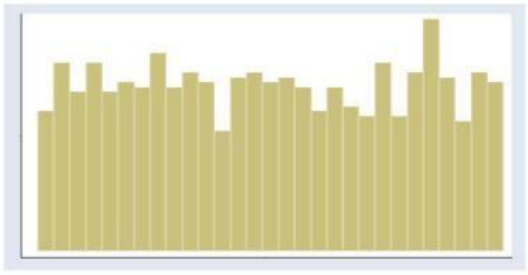
2



3



4



a) 4 **b) 2** **c) 1** **d) 3**

8. Para investigar la prevalencia de los padecimientos gastrointestinales un investigador médico decide seleccionar una muestra aleatoria de tamaño n = 100 a partir de la cual toma una medición de la variable "**número de pacientes**" con algún tipo de padecimiento gastrointestinal. Considera que hay suficientes elementos para considerar a este experimento como uno de tipo binomial. Identifica los supuestos que asumió el investigador para ser considerado experimento binomial.

1. El experimento consta de 100 intentos idénticos.
2. El número promedio de veces en que ocurre un evento por unidad de tiempo o de espacio es constante.
3. Los 100 intentos son independientes.
4. La probabilidad de más de un suceso en una unidad de tiempo o espacio es muy pequeña.
5. Cada intento da lugar exactamente a dos resultados: éxito y fracaso. Tales resultados pueden ser medidos por la variable x=0 correspondiente al fracaso y x = 1 correspondiente al éxito.
6. La probabilidad "p" de un éxito permanece constante de un intento a otro.

Selecciona la respuesta correcta de entre las cuatro opciones que se te presentan a continuación:

a) 1; 3; 5; b) 2; 4; 5; c) 1; 2; 3; 4 d) 1; 3; 4; 6

Unidad 2

Tratamiento estadístico de la información de fenómenos naturales y procesos sociales

¿Qué voy a aprender y cómo?

Está comprobado que la estadística tiene un papel fundamental en el avance de los conocimientos en muchos ámbitos, lo que ha permitido el desarrollo de la ciencia y la tecnología con las aplicaciones que conllevan.

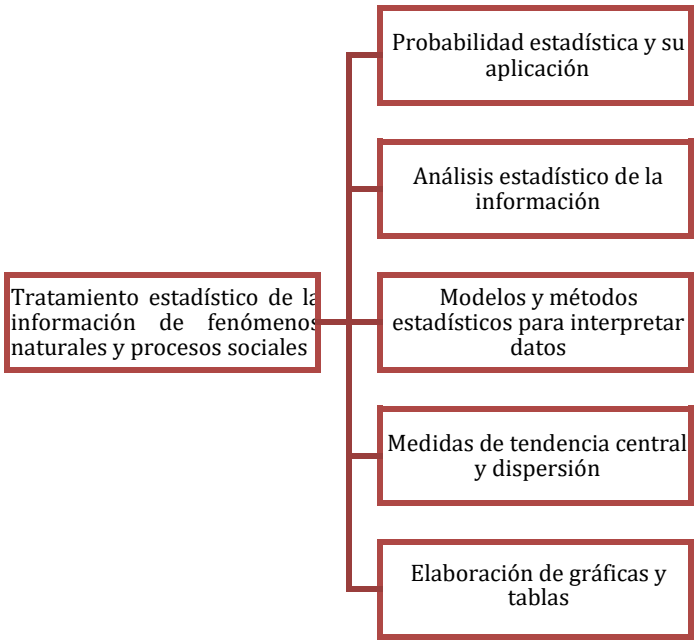
Para reflexionar:

¿Podrías considerar la importancia de la estadística en la literatura y la música? ¿Cómo?

¿Comprendes el alcance de la utilidad de los métodos estadísticos en la pintura, podrías ejemplificar?

:

Trabajarás los siguientes saberes.



1.9 La estadística descriptiva y su aplicación en la explicación de los fenómenos naturales y los procesos sociales.

Reunir, almacenar, ordenar y agrupar de forma clara y sencilla, los datos de una investigación en cuadros y tablas, y calcular parámetros básicos sobre el análisis de las tablas de acuerdo con las características del evento o población es la función de la estadística descriptiva.

A dos semanas de llegar a la fecha límite de inscripción de un programa del gobierno para repartir medicamentos gratuitos en Oaxaca, solo el 24% de los beneficiarios se habían inscrito.

Porcentaje de personas elegibles que se han inscrito al programa de medicamentos gratuitos en cada uno de los municipios del Estado											
24	60	12	38	21	26	23	33	19	19	26	60
16	21	28	20	21	41	22	16	29	26	22	16
48	11	19	13	22	22	30	20	21	34	26	20
25	19	17	21	27	19	27	60	20	52	20	12
14	18										

Por ser la variabilidad de porcentaje, muy elevada, se cuestiona si el 24% es una cifra representativa. ¿Cómo comprobar la **variabilidad numérica**?
¿Cómo cuantificar el grado de **dispersión** en el conjunto de datos?

Entendamos por **variabilidad numérica** la heterogeneidad presente en un conjunto de datos presentes y por **dispersión**, el grado de distanciamiento de un conjunto de valores respecto a un valor representativo del centro de datos.

Las medidas resumen de la variabilidad de datos son la variabilidad y la desviación estándar y las medidas representativas del centro de los datos son la media y la mediana.

2.1 Medidas de tendencia central: media, mediana y moda

El valor representativo donde están centrados o localizados los datos a lo largo de una recta numérica, se le llama **medida de tendencia central** y son la media y la mediana.

La media

La media, media aritmética o promedio muestral de un conjunto de datos numéricos es el resultado de la suma de las observaciones dividido entre el número de las observaciones mismas y bajo las siguientes consideraciones:

x = variable a partir de la cual hemos obtenido los datos muestrales.

n = número de observaciones en el conjunto de datos (el tamaño de la muestra).

x_1 = primera observación en el conjunto de datos.

x_2 = segunda observación en el conjunto de datos.

x_n = "n-ésima" (última) observación en el conjunto de datos

Ejemplo: si tenemos una muestra de tamaño n= 4 observaciones de la variable **x**=peso de un recién nacido en kg: X_1= 3.8 X_2= 2.9 X_3= 2.6 X_4= 3.4. Observa que el subíndice de la **x** no tiene relación con la magnitud. Se utiliza la letra griega sigma Σ para representar la suma de los valores de **x**.

La media muestral de un conjunto de observaciones $x_1, x_2.x_3 \ldots x_n$ representada por x se define como:

$$\bar{x} = \frac{suma\ total\ de\ observaciones\ en\ la\ muestra}{número\ de\ observaciones\ en\ la} = \frac{x1, x2.\ x3\ \ldots\ xn}{n} = \frac{\sum x}{n}$$

Por lo tanto, el peso promedio de los 4 recién nacidos ($\bar{x}$) quedaría como:

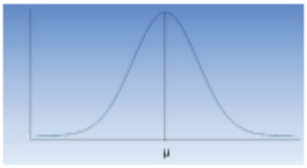
$$\bar{x} = \frac{\sum n =}{n} = \frac{12.7}{4} = 3.18\ kg$$

Analiza detenidamente el ejemplo (cuadro amarillo) de la página 93 de tu libro de texto donde se aplica el concepto de media estadística.

Es importante considerar que las características de la muestra se representan con letras del abecedario ($\bar{x}$) y las características de la población se designan con letras griegas por ejemplo, la media poblacional.

La media poblacional, representada por μ (se lee mu) es el promedio de todos los valores de la x de la población entera.

En una distribución de probabilidad normal, el valor de la media μ es el que divide en dos partes iguales (0.5 del área bajo la curva a cada lado de μ) a esta distribución de la probabilidad.



En una población de 300 niños recién nacidos el peso promedio podría ser podría ser μ= 3.2 kg mientras que $\bar{x}$ =3.14 para una muestra en particular $\bar{x}$ = 3.52 para otra y $\bar{x}$= 2.9 en otra.
Una posible desventaja potencial de la media es que su valor puede verse afectado por la presencia de valores extremos o atípicos, tal como se muestra en el ejemplo del libro de texto, páginas 94 y 99.

La mediana

- Se define como la cantidad que divide al conjunto de datos en dos partes iguales.
- Una vez que los valores de los datos se han enumerado de mayor a menor, la mediana es el valor medio de la lista y es el que divide la lista en dos partes iguales.
- La mediana puede ser ligeramente diferente si el tamaño de la muestra *n* es par o impar.
-

La mediana de la muestra se obtiene al ordenar un conjunto de *n* observaciones tomando como criterio de orden el iniciar con el valor más pequeño y terminarlo con el más grande (con los valores repetidos incluidos, de tal forma que cada muestra aparece en una lista ordenada). La mediana muestral se define como el valor medio si *n* es impar y como el promedio de los dos valores medios si *n* es par.

Ejemplo con una muestra de 40 datos (par)

Lista ordenada de los datos											
0	0	0	0	0	0	3	4	4	4	5	5
7	7	8	8	8	12	12	13	13	13	14	14
16	18	19	19	20	20	21	22	23	26	36	36
37	42	84	331								

La mediana se determina como

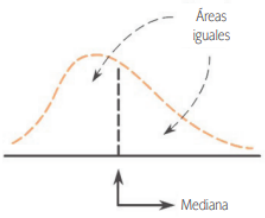
$$mediana = \frac{13 + 13}{2} = 13$$

El valor de la mediana parece ser más típico que el valor de la media que es de $\bar{x}$ = 23.10 pues se encuentra más al centro del conjunto de datos. La media es sensible a valores que se disparan, la mediana no, es por esto que se justifica el uso de esta última como medida de tendencia central.

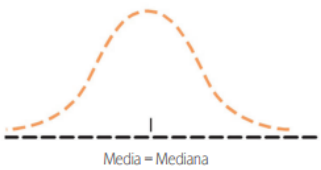
Comparación de la media y la mediana

Otra medida alternativa de la tendencia central es la **moda muestral** que se define como el valor que ocurre con mayor frecuencia en un conjunto de datos. Ejemplo en la muestra 2, 2, 5, 7, **9, 9, 9**,10,10,11,12 y 18, la moda es 9. Puede no existir o ser **bimodal**, significa un histograma donde aparecen dos picos.

Gráficamente, la mediana es el valor en el eje de medición que separa a la curva en dos partes, con 0.5 (50%) del área bajo la curva en cada parte. Válido para



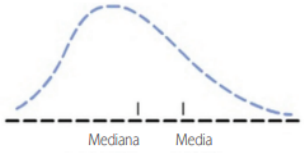
Si el histograma es simétrico, como es el caso de la curva de distribución normal, el punto de simetría divide la curva en dos áreas iguales y



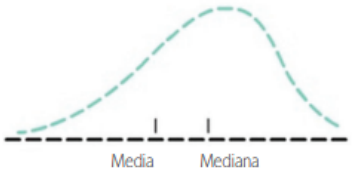
muestras y población.

además es punto de equilibrio. En este caso la media y la media son iguales.

Cuando el histograma tiene un solo pico (es unimodal), con una cola más larga (a esta característica se le conoce como sesgo positivo) los valores extremos de la cola superior jalan a la media, por lo que generalmente ésta es mayor que la mediana.

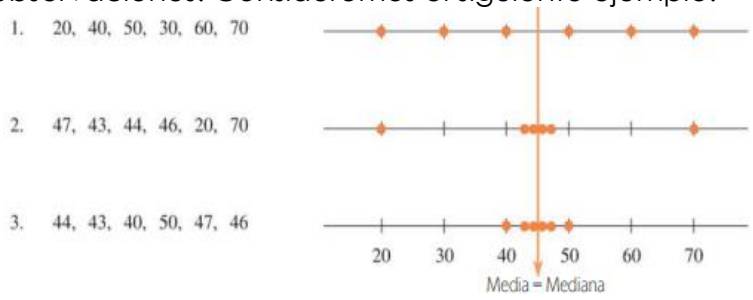


Del mismo modo, cuando un histograma unimodal está negativamente sesgado (con una cola inferior más larga), como se muestra en la siguiente gráfica, la media es generalmente menor que la mediana.



2.2 Medidas de dispersión

Una medida de tendencia central no permite cuantificar que tan diferentes son las observaciones. Consideremos el siguiente ejemplo:



Los 3 conjuntos tienen una media y mediana de 45 pero la dispersión de datos es completamente distinta.

Rango muestral= observación más grande - la observación más pequeña. Es una resta.

Desviación de la media

Es una de las medidas de variabilidad más utilizadas. Mide el grado en que cada observación de la muestra se desvía de su correspondiente media muestral $\bar{x}$. Al restar $\bar{x}$ de cada observación se proporciona un conjunto de desviaciones de la media.

> Dado una muestra de tamaño n, las n desviaciones de la media muestral son las diferencias $(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots (x_n - \bar{x})$

Una desviación en particular será positiva si el valor de x correspondiente es mayor que $\bar{x}$ y negativo si el valor de x es menor que $\bar{x}$.

Ejemplo:

Un laboratorio multinacional produjo vacunas contra el virus de la influenza. En la tabla aparece el costo, en dólares, de cada dosis.
Se observa poca variabilidad en los precios y donde $\bar{x}$= $3.36

País	PRECIO DE LA VACUNA POR DOSIS
Argentina	3.02
Brasil	4.67
Chile	3.28
Colombia	3.51
Costa Rica	3.42
Perú	2.70
Uruguay	2.87

La tabla adjunta muestra las desviaciones estándar (después de restar 3.36 a cada dato registrado). Tres de las desviaciones son positivas, es decir mayor que x. Las desviaciones negativas son más pequeñas que x. Algunas de las desviaciones son muy grandes en magnitud absoluta* (1.31 y 0.60, por ejemplo), indicando con esto las observaciones que están más lejos de la media muestral.

País	Precio de la vacuna por dosis	Desviaciones de la media
Argentina	3.02	-0.34
Brasil	4.67	1.31
Chile	3.28	-0.08
Colombia	3.51	0.15
Costa Rica	3.42	0.06
Perú	2.76	-0.60
Uruguay	2.87	-0.49

*Una **magnitud absoluta** es la medición que no considera o ignora el signo negativo de la cantidad asignada. Por ejemplo, la magnitud absoluta de -2 es 2.

Las desviaciones se combinan en una sola medida de variabilidad bajo la fórmula $\sum x - \bar{x}$ y luego dividirla por n.

Contribuciones

Valor numérico asociado con una desviación media.

En el ejemplo anterior el valor de las sumas de las desviaciones es aproximadamente cero, es decir que en la mayoría de los casos el valor calculado de la suma $\sum x - \bar{x}$ estará muy cercano a cero o incluso será cero. Esto limita la aplicabilidad de la desviación media como medida general de la variabilidad de un conjunto de datos muestrales, ya que las **contribuciones** (por ejemplo, una contribución de -3 significa que la diferencia correspondiente a la desviación media entre un valor x y su media $\bar{x}$ es -3) que tengan la misma magnitud absoluta pero diferente signo como -2 y 2 serán

anuladas al momento de sumarlas, por lo que su contribución a la variabilidad no se considerará.

Varianza y desviación estándar

Una manera de evitar las contribuciones negativas con las positivas es elevarlas al cuadrado antes de sumarlas ya que en el caso de -2 y 2 no se anularán ya que cada una equivale a positivo y contribuirán a la variabilidad. El cuadrado de estas desviaciones para muestras de tamaño *n* se representan con la siguiente sumatoria

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 = \sum(x - \bar{x})^2$$

Al dividir el resultado de la suma por el tamaño de la muestra *n* se obtiene la desviación cuadrada promedio. Aunque ésta parece ser una medida $\sum (x - \bar{x})^2$ razonable de la variabilidad utilizaremos un divisor ligeramente menor a *n,* como lo sugiere la teoría de la estimación estadística.

La **varianza muestral** representada como S^2, es la suma de los cuadrados de las desviaciones de la media divididas entre n-1. Esto es: $S^2 = \frac{\sum(x-\bar{x})_2}{n-1}$

La **desviación estándar muestral** es la raíz cuadrada positiva de la varianza muestral y se representa como *S*

Una mayor variabilidad dará como resultado un valor grande en S^2 y *S* dará como resultado un valor relativamente grande, mientas que un valor de cero o cercano a cero de S^2 y *S* indica una menor variabilidad. Es importante enfatizar que las unidades de medición se elevarán a cuadrado de acuerdo con la fórmula y se simplificarán al obtener la raíz cuadrada.

Siguiendo con los datos del ejemplo anterior, las desviaciones de la media y sus cuadrados quedarían de esta manera:

País	Precio de la vacuna por dosis	Desviaciones de la media	Desviaciones de la media cuadrada
Argentina	3.02	-0.34	0.1156
Brasil	4.67	1.31	1.7161
Chile	3.28	-0.08	0.0064
Colombia	3.51	0.15	0.0225
Costa Rica	3.42	0.06	0.0036
Perú	2.76	-0.6	0.3600
Uruguay	2.87	-0.49	0.2401
			$\sum(x+\bar{x})^2 = 2.4643$

A partir de este valor obtenemos los valores de la varianza y la desviación estándar de la siguiente forma: $S^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{2.4643}{7-1} = \frac{2.4643}{6}$ 0.4107 $\quad y \quad S = \sqrt{0.4107} = 0.641$

(En el Apéndice 2 se describen los pasos para realizar estos cálculos en la hoja electrónica).

Hay medidas de variabilidad para toda la población, son análogas a S^2 y S para una muestra. Estas medidas reciben el nombre de varianza de la población y la desviación estándar de la población y se simbolizan con σ^2 y σ $(sigma\ cuadrada\ y\ sigma)$

S^2 Varianza muestral
σ^2 Varianza poblacional
S Desviación estándar muestral
σ Desviación estándar poblacional

¡Ahora te toca a ti!

Retoma el caso planteado al inicio de la unidad y empieza a medir el impacto del fenómeno migratorio en la comunidad El Encino. Ubica fuentes de información donde puedas obtener dos muestras representativas de 100 mediciones cada una, correspondientes a dos poblaciones distintas.

Actividad 2.1:

Realiza un análisis descriptivo de poblaciones utilizando las medidas de tendencia central y dispersión: media, mediana y moda.

1. Sea distribución estadística que viene dada por la siguiente tabla:

Muestra x_i	Frecuencia absoluta f_i
61	5
64	18
67	42
70	27
73	8
	n=100

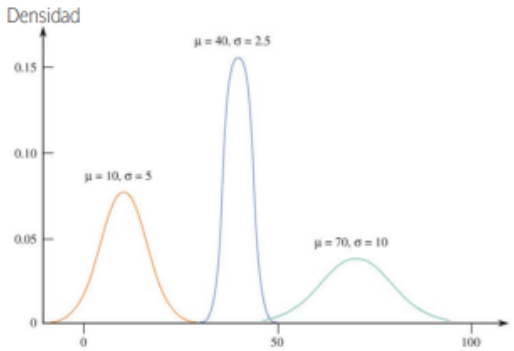
a) Moda
b) Mediana
c) Media
d) Rango
e) Varianza
f) Desviación estándar

2.3 Las distribuciones de probabilidad como herramientas para estimar las probabilidades de eventos relacionados con fenómenos naturales y procesos sociales.

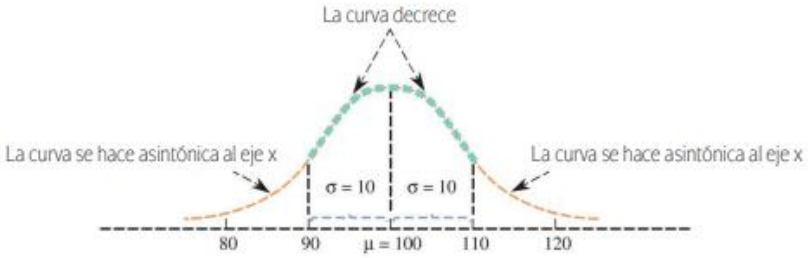
Distribución de probabilidad normal

Las distribuciones normales tienen una forma de campana con curvas normales. Los tipos de distribuciones normales se distinguen entre sí por los valores de su media μ y su desviación estándar σ. La μ describe dónde está centrada la curva correspondiente, la desviación estándar σ define la dimensión de la curva cuando se extiende alrededor de ese centro.

La figura a la derecha muestra (página 107 del texto) 3 distribuciones normales donde se observa la curva de la gráfica en relación a μ y σ; cuánto es menor la desviación estándar, cuánto más alta y más estrecha se vuelve la curva correspondiente, ya que al haber menor variabilidad la mayoría de los valores de la correspondiente distribución normal se concentra a una distancia menor de la media.

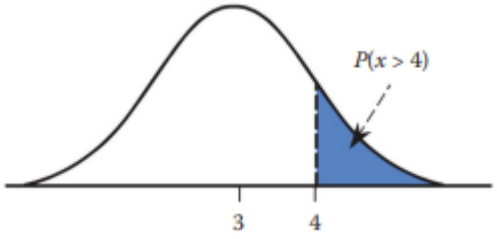


El valor de la media es el número que en el eje de la medición ($eje\ x$) se extiende debajo de la parte superior de la campana. El valor de la desviación estándar también se puede determinar a partir de la gráfica.



Asintónica se refiere al hecho de que la curva se aproxima a una línea recta horizontal pero sin llegar jamás a tocar el eje x

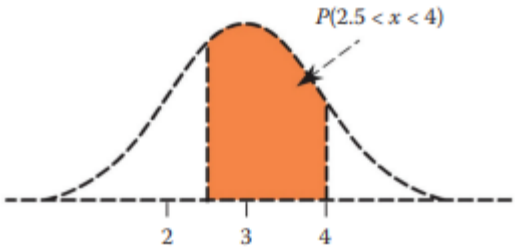
Retomado el ejemplo del peso de los recién nacidos, se pueden usar áreas bajo la curva normal con μ =3 y σ= 1. La probabilidad de que el peso de un recién nacido sea mayor a 4kilogramos,$P(x > 4)$ corresponde al área



Por otra parte, como se observa en la gráfica de abajo a la izquierda, la probabilidad $P(2.5 < x < 4)$ de que un recién nacido pese entre 2.5 y 4 kg, corresponde al área sombreada.

El cálculo directo de estas probabilidades es complejo por lo que se

sombreada en la gráfica de arriba a la derecha

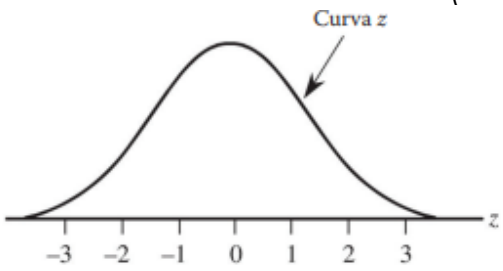


utiliza una tabla de áreas para la distribución normal de referencia, distribución normal estándar.

La distribución normal estándar es una distribución normal con μ=0 y σ=1. La curva de densidad correspondiente también es denominada curva z o curva normal estándar. Se acostumbra utilizar la letra z como la variable para representar los posibles valores que puede tomar una distribución normal.

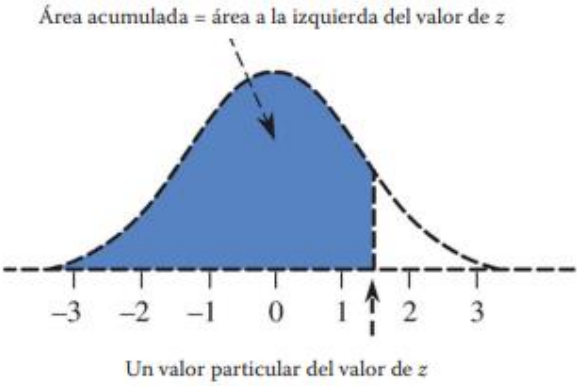
La distribución normal estándar

La curva normal estándar (z) que se muestra en la figura, está centrada en μ= 0 y σ =1; este último valor se emplea como medida del grado en que posibles valores de la variable z se extienden alrededor de la media (en este caso, 0).



Por ejemplo, alrededor del 95% del área (probabilidad) se asocia con valores que están dentro de dos desviaciones estándar de la media (entre -2 y 2) y casi toda el área en la zona de los valores que están dentro de tres desviaciones estándar alrededor de la media (entre -3 y 3)

Para calcular la curva normal estándar se utiliza una tabla con los valores acumulados bajo el área de la curva z para valores $z*$
distintos de la variable z, que puedes consultar en el Apéndice 4 (página 231 de tu libro de texto) y del tipo mostrado en la gráfica de la derecha de la derecha.

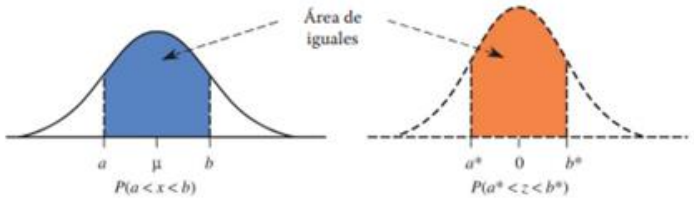


Analiza detenidamente la información que aparece en la columna Gestión del aprendizaje de la página 109 de tu libro de texto, consultando al mismo tiempo la tabla de la distribución

normal estándar del Apéndice 4 página 231para que identifique la probabilidad que corresponde a cada valor de $z*$ dentro del área de la curva z.
Revisa también los ejercicios de las páginas 110 y 111 y 112 de tu libro de texto para localizar el área de la curva z y las probabilidades del evento, en la tabla que aparece en la página 110.

Otras distribuciones normales

Las áreas de la curva z también pueden emplearse para el cálculo de probabilidades de cualquier distribución normal.
La letra z se utiliza para las variables que tienen una distribución normal estándar, la letra x para cualquier variable cuya distribución se describe mediante una curva normal con media μ y la desviación estándar σ. Si se va a calcular $P(a < z < b)$, la probabilidad de que la variable x se encuentre en un rango determinado; esta probabilidad corresponde al área bajo la curva normal y por encima del intervalo de a a b.



La notación (*) se utiliza para distinguir a y b (los valores de la distribución normal original con la media μ y la desviación estándar σ), de a^*y b^*, los valores correspondientes a la curva de z. Para encontrar a^* y b*, se calculan las puntuaciones z para los puntos finales del intervalo para los que se desea calcular la probabilidad. Este proceso se llama normalización de los criterios de valoración. Ejemplo si la variable x tiene una distribución normal con media μ=100 y la desviación estándar $\sigma = 5$ para encontrar $P(98 < z < 107)$

Se convierte este problema en un problema equivalente a la distribución normal estándar; se inicia calculando la puntuación z en el punto final inferior a =98, para lo cual a partir de dicho valor se le resta su media y posteriormente se le divide entre su desviación estándar, convirtiendo el punto final inferior a=98 en el punto estandarizado, dando por resultado -0.40

$$a^* = \frac{98-100}{5} = \frac{2}{5} = -0.40$$

Y convirtiendo el punto final $b = 107$. Para transformarse en:

$$b^* = \frac{107-100}{5} = \frac{7}{5} = 1.40$$

Entonces: $P(98 < z < 107) = P(-0.40 < z < 1{,}40).$

La probabilidad puede ahora evaluarse utilizando la tabla del Apéndice 4, página 231.

La puntuación z correspondiente a un valor particular es:

$$z = \frac{valor - media}{desviación\ estándar}$$

La puntuación z indica que tantas desviaciones estándar está el valor de la media. Es positivo o negativo si el valor se encuentra por encima o por debajo de la media.

Encontrando probabilidades de una distribución normal

A fin de calcular las probabilidades para cualquier distribución normal se deben estandarizar los valores correspondientes al punto inferior y superior de la distribución normal original y luego usar la tabla de áreas de la curva z. De manera específica, si x es una variable cuyo comportamiento está descrito por una distribución normal con media y desviación estándar, entonces (analiza las fórmulas de la derecha):

$$P(x < b) = P(z < b^*)$$

$$P(x < a) = P(z < a^*)$$

$$P(a < x < b) = P(a^* < z < b^*)$$

donde z es una variable cuya distribución es normal estándar y $a^* = \frac{a-\mu}{\sigma}$ $b^* = \frac{b-\mu}{\sigma}$

Ejemplo:

En *Los parámetros de crecimiento fetal y peso al nacer: su relación con la composición corporal neonatal* se sugiere que una distribución normal con media $\mu = 3500$ *gramos* y una desviación estándar $\sigma = 600$ gramos representan un modelo razonable para la distribución de probabilidad de variable numérica continua x = peso al nacer de un bebé seleccionado al azar. ¿Qué proporción de pesos de recién nacidos está entre 2900 y 4700 gramos?

$P(2900 < x < 4700)$

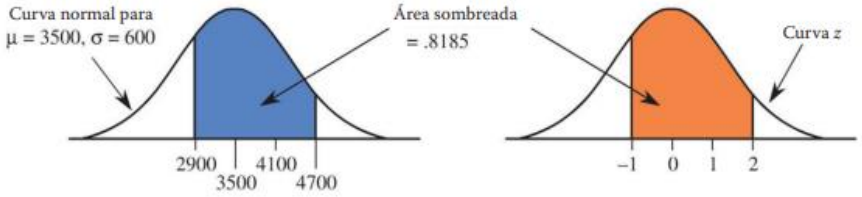
Se transforman los extremos a $a = 2900$ y $b = 4700$ del intervalo correspondiente a la distribución de la variable x a los extremos $a^* y\ b^*$ del intervalo equivalente a la distribución normal estándar para a^* y b^*:

$$a^* = \frac{a-\mu}{\sigma} = \frac{2900-3500}{600} = -1.00$$

$$b^* = \frac{b-\mu}{\sigma} = \frac{4700-3500}{600} = 2.00$$

Entonces:

$$P(2900 < x < 4700) = P(-1.00 < z < 2.00)$$
$$= (\text{área de la curva } z \text{ a la izquierda de})$$
$$= 0.9772 - 0.158$$
$$= 0.8185$$



El resultado significa que, si se midieran los pesos al nacer muchos bebés de esa población, el peso del 82% de ellos se situaría entre 2900 y 4700 gramos.

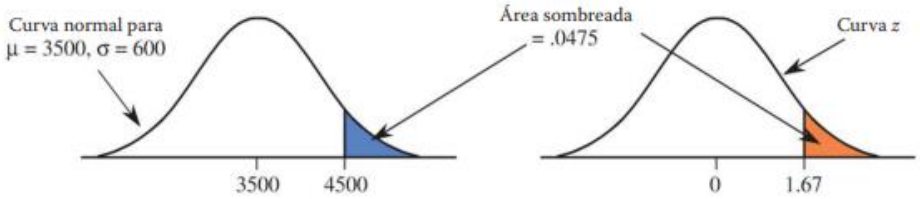
¿Cuál es la probabilidad de que un niño elegido al azar tenga un peso al nacer mayor de 4500 gramos? Es decir: $P(x > 4500)$

$$a^* = \frac{a-\mu}{\sigma} = \frac{4500-3500}{600} = 1.67$$

Entonces:

$$P(x > 4500) = P(z > 1.67)$$
$$= \text{área de la curva z a la derecha de } 1.67$$
$$= 1 - (\text{área de la curva z a la izquierda de } 1.67$$
$$= 1 - 0.9525$$
$$= 0.0475$$

Las probabilidades de x y z se muestran en la siguiente gráfica.



Distribución de probabilidad binomial

Recordemos que la probabilidad **binomial** se refiere a la probabilidad de x éxitos en n intentos

repetidos en un experimento que tiene dos resultados posibles.
Ejemplo:
Si se desea registrar el sexo de los próximos 5 niños nacidos en el hospital de una comunidad, ¿cuál es la probabilidad de que uno de los 5 niños sea mujer? ¿cuál de que entre 2 y 4 sean mujeres?
Analizando el experimento se encuentra que presenta las siguientes características:

1. n= 5 ensayos (intentos) idénticos; cada intento consiste en registrar el sexo del recién nacido.
2. Cada intento da lugar exactamente a dos resultados, hombre o mujer.
3. Los 5 intentos son independientes pues el resultado obtenido de cada registro es independiente de los resultados obtenidos en los otros registros.
4. La probabilidad de que al registrar el sexo de un recién nacido sea niña no varía de un registro a otro. Esta probabilidad a la cual denotamos como p representa la probabilidad de que en cualquier ensayo particular se obtenga un éxito (en este caso que el sexo del recién nacido sea niña).

Fórmula de la distribución binomial

Sea n= número de ensayos independientes en un experimento binomial. p = la probabilidad constante de que en cualquier ensayo particular se obtenga un éxito. Entonces su fórmula es:

$$P(x) = P(x\ éxitos\ en\ n\ ensayos) = \frac{n!}{x!\,(n-x)!}\,p^x(1-p)^{n-x}$$

$$x = 0, 1, \dots, n$$

La expresión $\binom{n}{x}$ algunas veces se utiliza en lugar de $\frac{n!}{x!(n-x)!}$, cuya expresión representa el número de maneras de elegir elementos de un conjunto de n elementos. El símbolo $n!$ (se lee n factorial) se define como $n! = n(n-1)(n-2)\dots(2)(1)$ y $0! = 1$.
La fórmula de la probabilidad binomial puede entonces expresarse como:

$$P(x) = \binom{n}{x}p^x(1-p)^{n-x} \qquad x = 0, 1, \dots, n$$

Si suponemos que la probabilidad de éxito es $p = 0.4$ y queremos calcular la probabilidad de que uno de los próximos cinco recién nacidos sea mujer, basta con sustituir en la fórmula los valores $n = 5$ y $x = 1$ para obtener:

$$P(1) = P(x = 1) = \binom{5}{1}(.4)^1(.6)^4$$
$$(.6)^4 = .26$$

Y si ahora queremos calcular la probabilidad de que entre dos y cuatro recién nacidos sean mujeres obtenemos $P(2 \leq x \leq 4) = P(x = 2\ o\ \ x = 3\ \ o\ x = 4)$
Dado que estos resultados son **disjuntos**, esto es igual a

$$P(x = 2\ \ o\ \ \ x = 3\ \ \ o\ \ x = 4) = P(x = 2) + (x = 3) + (x = 4)$$
$$= P(2) + P(3) + P(4)$$
$$= \binom{5}{2}(.4)^2(.6)^3 + \binom{5}{3}(.4)^3(.6)^2 + \binom{5}{4}(.4)^4(.6)^1$$
$$= \frac{5!}{2!\,3!}(.4)^2\,(.6)^3 + \frac{5!}{3!\,2!}\,(.4)^3(.6)^2 + \frac{5!}{4!\,1!}(.4)^4(.6)^1$$
$$= (10)(.4)^2(.6)^3\ + (10)(.4)^3(.6)^2 + (5)(.4)^4(.6)^1$$
$$= .3456 + \ .2304 + \ .0768 \qquad Resultado = .65$$

Disjuntos

Dos eventos que no suceden al mismo tiempo, que no tienen resultados en común

Y si ahora queremos calcular la probabilidad de que entre dos y cuatro recién nacidos sean mujeres obtenemos $P(2 \leq x \leq 4) = P(x = 2\ o\ x = 3\ o\ x = 4)$
Dado que estos resultados son **disjuntos**, esto es igual a

$$P(x = 2\ \ o\ \ \ x = 3\ \ \ o\ \ x = 4) = P(x = 2) + (x = 3) + (x = 4)$$
$$= P(2) + P(3) + P(4)$$
$$= \binom{5}{2}(.4)^2(.6)^3 + \binom{5}{3}(.4)^3(.6)^2 + \binom{5}{4}(.4)^4(.6)^1$$
$$= \frac{5!}{2!\,3!}(.4)^2\,(.6)^3 + \frac{5!}{3!\,2!}\,(.4)^3(.6)^2 + \frac{5!}{4!\,1!}(.4)^4(.6)^1$$
$$= (10)(.4)^2(.6)^3\ + (10)(.4)^3(.6)^2 + (5)(.4)^4(.6)^1$$
$$= .3456 + \ .2304 + \ .0768 \qquad Resultado = .65$$

Refuerza tus conocimientos, analizando el ejercicio de la página 119.

Distribución de probabilidad de Poisson

Mediante esta herramienta se determina la probabilidad de que un evento suceda en un determinado tiempo o espacio.

Por ejemplo: ¿cuál es la probabilidad de que en una hora lleguen exactamente dos pacientes al servicio de emergencias?, ¿cuál es la probabilidad de que en una hora lleguen al menos dos pacientes?

Revisa las condiciones del evento:

1. El número promedio de pacientes que acude al hospital por unidad de tiempo (una hora para este evento) es constante.

2. La probabilidad de que más de un paciente acuda en cualquier intervalo de tiempo menor es casi cero.
3. El número de pacientes que acude al hospital en intervalos ajenos de tiempos son independientes unos de otros.

Si el evento cumple con estas hipótesis, entonces se trata de un **experimento de Poisson**.

Si $\boldsymbol{x}$ es la variable que representa el número de sucesos resultantes de este experimento, entonces ésta representa una variable discreta que puede tomar números enteros (0, 1, 2, 3, 4, 5) utilizando la siguiente fórmula:

Fórmula de la distribución de Poisson

Sea

$\lambda = número\ medio\ de\ observaciones\ por\ unidad\ de\ tiempo$

$e = base\ de\ los\ logaritmos\ neperanios.\ Su\ valor\ aproximado\ es\ de\ 2.71828$

$$P(x) = \frac{\lambda\, e}{x!} \qquad x = 0, 1, 2, \ldots$$

Ejemplo:

Calcular la probabilidad de que en una hora lleguen exactamente 2 pacientes $x = 2,\ \lambda = 3$ (número promedio de pacientes que llegan al servicio de emergencias por unidad de tiempo) y sustituir los valores en la fórmula de Poisson

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad P(2) = \frac{3^2 e^{-3}}{2!} \qquad = \frac{(9)(0.4979)}{2} \qquad = .22$$

La probabilidad de que en una hora lleguen al menos dos pacientes es

$$P(x \geq 2) = P(x = 2\ o\ x = 3\ o\ x = 4\ o\ \ldots\ldots)$$

Para sortear la dificultad de la infinidad de eventos, se calcula la probabilidad a partir de su complemento:

$$P(x \geq 2) = P(x < 2) = 1 - P(x < 2)$$

$$= 1 - Px = 0\ o\ x = 1$$

$$= 1 - P(x = 0) - P(x = 1)1$$

$$= 1 - P(0) - P(1)$$

$$= 1 - \frac{(3)^0 e^{-3}}{0!} - \frac{(3)^1 e^{-3}}{1!}$$

$$= 1 - e^{-3} - 3e^{-3}$$

$$= .80$$

Complementos

Un evento A es complementario a un evento B si A está compuesto por los eventos que no están en el evento B. Por ejemplo, en el fenómeno natural de temperatura ambiente, el evento frío (F) es complementario del evento calor (C).

En términos probabilísticos:

$$P(F) = 1 - P(C)$$

¡Ahora te toca a ti!

Ahora aplica los modelos de distribuciones de probabilidad teórica para estimar las probabilidades de ocurrencia de cierto tipo de eventos, que has aprendido en esta sección. Para realizar la siguiente actividad, retoma la investigación del impacto migratorio en la comunidad del Encino.

Actividad 2.2

Resuelve los siguientes ejercicios de distribución: normal, binomial y Poisson.

1. Distribución normal

En una ciudad se estima que la temperatura máxima en el mes de junio sigue una distribución normal, con media de 23° y desviación típica de 5°.
Calcula el número de días del mes en los que se espera alcanzar máximas entre 21° y 27°

2. Distribución binomial

Un laboratorio afirma que un medicamento causa efectos secundarios en una proporción de 3 de cada 100 pacientes. Para contrastar esta afirmación, otro laboratorio elige al azar a 5 pacientes a los que aplica el medicamento. ¿Cuál es la probabilidad de los siguientes sucesos?
a) Ningún paciente tenga efectos secundarios.
b) Al menos dos tengan efectos secundarios.
c) ¿Cuál es el número medio de pacientes que espera laboratorio que sufran efectos secundarios si elige 100 pacientes al azar?

3. Distribución de Poisson

La veterinaria de Jorge recibe un promedio de $\lambda = 4$ pacientes por día. Sabiendo que el número de pacientes que llegan en un día sigue una distribución de Poisson, calcula:

a) la probabilidad de que lleguen 3 pacientes en un día.
b) la probabilidad de que lleguen 5 pacientes en un día.

2.4 El modelo de regresión y el de correlación lineal como medidas para describir la asociación entre variables.

El estudio de los métodos que complementan las gráficas de dispersión para describir las relaciones entre dos variables numéricas y que también permitan evaluar la fortaleza de esta relación tales como el coeficiente de correlación muestral de Pearson, el coeficiente de correlación de la población y la regresión lineal son los temas por revisar en esta sección.

Correlación

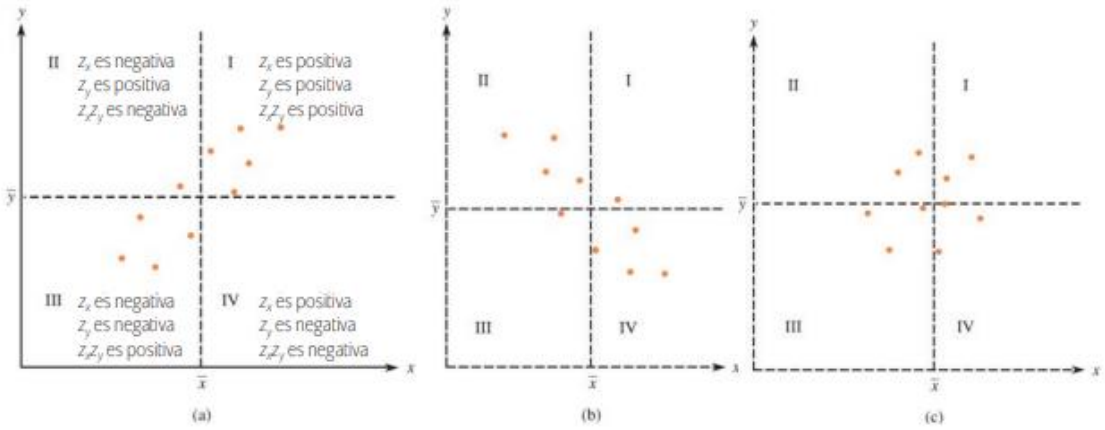
La correlación nos ayuda a visualizar la relación que hay entre dos variables; para argumentar de forma precisa y sanado conclusiones acertadas se utiliza el coeficiente de correlación muestral, cuyo valor da una evaluación numérica de la tendencia y la fuerza de la relación entre los valores de x y y en un conjunto de datos con observaciones pareadas de la forma (x, y). El coeficiente más utilizado es el de Pearson.

Coeficiente muestral de Pearson

El coeficiente muestral de Pearson mide la fuerza de una relación lineal entre dos variables numéricas mediante el uso de las puntuaciones z. Lo hace al reemplazar cada valor de x por su correspondiente puntuación z y z_x (obtenido de restar $\bar{x}$ y x y dividir este resultado por S_x, la desviación estándar de los valores de x); de manera similar, se reemplaza cada valor y por su puntaje z.

Los valores de x mayores generan puntuaciones z_x positivas y los menores un efecto negativo en las puntuaciones z_x . También los valores y más grandes que $\bar{y}$ generan puntuaciones z_y positivas y los más pequeños dan como resultado puntuaciones z_y negativas. El coeficiente de correlación muestral de Pearson se basa en la suma de los productos de z_xy z_y para cada observación en el conjunto de observaciones pareadas. Su notación algebraica es $\sum z_x z_y$ ....

Analiza detenidamente las gráficas que aparecen a continuación, especialmente la información de la gráfica (a) complementa con la información de la página 126 de tu libro de texto, en caso que así lo consideres.



El coeficiente de correlación muestral de Pearson se obtiene dividiendo $\sum z_x z_y$ por $(n-1)$

El coeficiente de correlación muestral de Pearson está dado por $\frac{\sum z_x z_y}{(n-1)}$

Hay varios coeficientes de correlación. Éste es el más usado y se conoce simplemente como coeficiente de correlación.

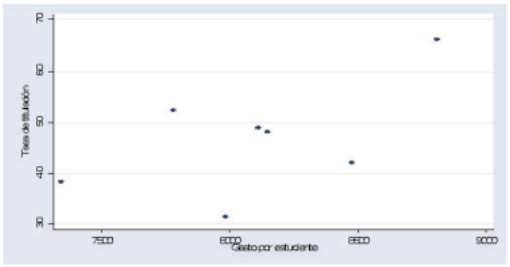
Es posible realizar el cálculo del coeficiente de correlación muestral en una hoja de cálculo electrónica de manera sencilla. Basta seguir los pasos descritos en el Apéndice 2 desde la página 216 hasta las 218.

Ejemplo:

En cierto estado del país existen siete universidades públicas con una matrícula registrada en conjunto de entre 10,000 y 20,000 estudiantes. Las universidades están solicitando presupuesto para financiarse. La tasa de titulación, así como el costo en pesos invertido por estudiante durante el último año en cada una de las siete universidades, su registro aparece en la tabla de la derecha

Universidad	Tasa de titulación (%)	Gasto por estudia
1	66.1	8,810
2	52.4	7,780
3	48.9	8,112
4	48.1	8,149
5	42.0	8,477
6	38.3	7,342
7	31.3	7,984

La gráfica de dispersión de estos datos se observa a la derecha(costo por estudiante en x y tasa de titulación en y)



$\bar{x} = 8093.43$ $S_x = 472.39$ $\bar{y}$ 46.73 $S_y = 11.15$

Para calcular el coeficiente de correlación se empieza por el cálculo de las puntuaciones z para cada par (x, y) en el conjunto de datos. Por ejemplo, la primera observación es (8810, 66.1). Los correspondientes valores z son:

$$z_x = \frac{8810 - 8093.43}{472.39}\ 1.52 \qquad z_y = \frac{61.1 - 46.73}{11.15}$$

La tabla de la derecha muestra las puntuaciones z, el producto $z_x z_y$ para cada observación, así como la suma total de los productos $z_x z_y (\Sigma z_x z_y)$:

y	x	zx	zy	zx zy
66.1	8,810	1.52	1.74	2.64
52.4	7,780	-0.66	0.51	-0.34
48.9	8,112	0.04	0.20	0.01
48.1	8,149	0.12	0.13	0.01
42.0	8,477	0.81	-0.42	-0.34
38.3	7,342	-1.59	-0.75	1.20
31.3	7,984	-0.23	-1.38	0.32
				$\sum z_x z_y = 3.52$

Con la información anterior es posible calcular el valor del coeficiente de correlación r

$(\Sigma z_x z_y) = 3.52$

Entonces con $n = 7$

$$r = \frac{\Sigma z_x z_y}{n-1}\ \frac{3.52}{6} = .587$$

Con base en el diagrama de dispersión y las propiedades del coeficiente de correlación ilustradas en el ejemplo anterior, es posible concluir que existe una moderada relación lineal positiva entre el gasto por estudiante y la tasa de titulación en estas siete universidades.

Propiedades del coeficiente de correlación (r)

1. El valor de r no depende de la unidad de medida de cada variable; así, si x es la altura, la correspondiente puntuación z es la misma; es decir, que si la altura se expresa en centímetros, metros o kilómetros el valor del coeficiente de correlación no se afecta.

2. El valor de r no depende de cuál de las dos variables se considera x . Por lo tanto, si en el ejemplo anterior tuviéramos que x = tasa de graduación e y = gastos por estudiante, se habría obtenido el mismo valor, $r = .587$.

3. El valor de r está entre -1 y 1. Un valor cercano al límite superior 1, indica una fuerte relación positiva. En términos de la relación entre la variable x y la variable y se traduce en que al aumentar el valor de la variable x existe una fuerte tendencia

lineal a que aumente el valor de la variable y. Mientras un valor de r cercano al límite inferior, -1,

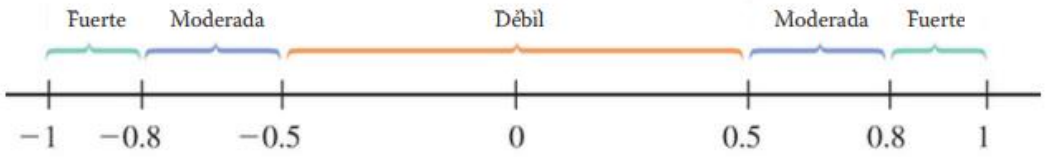
sugiere una fuerte relación negativa, lo que indica que al aumentar el valor de la variable x existe una fuerte tendencia lineal a que disminuya el valor de la variable y.

4. Un coeficiente de correlación de $r = 1$ sólo se produce cuando todos los puntos en una gráfica de dispersión se encuentran exactamente en una línea recta que tiene pendiente ascendente. Del mismo modo, $r = -1$ sólo se da cuando todos los puntos están exactamente en una línea descendente. Únicamente cuando hay una relación lineal perfecta entre x y y en la muestra, r toma alguno de sus dos valores extremos posibles. Encontrar estos niveles de relación entre dos variables es muy raro.

5. El valor de r es una medida del grado en que x y y están relacionadas linealmente; es decir, el grado en que los puntos de una dispersión está cerca de una línea recta. Un valor cercano a 0 no descarta una relación fuerte entre x y y, pues podría haber una fuerte relación que no es lineal entre x y y. De hecho, la imposibilidad de detectar relaciones no lineales entre dos variables constituye una seria desventaja de esta medida de asociación.

La figura muestra los 3 niveles de asociación en que se clasifican los posibles valores que puede tomar el coeficiente de correlación:

- Correlación débil si el coeficiente de correlación toma valores entre -0.5 y 0.5.
- Correlación moderada si sus valores están entre 0.5 y 0.8 o entre -0.5 y -0.8.
- Correlación fuerte si los valores del coeficiente se encuentran entre 0.8 y 1 o entre -0.8 y -1

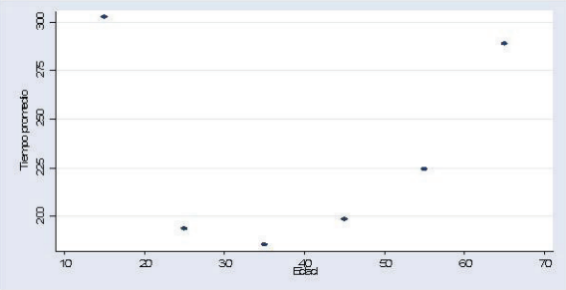


Refuerza la información repasando los dos ejemplos de las páginas 130, 131 y 132 de tu libro de texto.

Un ejemplo mas

Un grupo de médicos del deporte, contratado para llevar el seguimiento médico de los participantes en una carrera de 20 km, está interesado en medir la relación que pueden tener el grupo de edad de las mujeres participantes, con el tiempo en minutos que tardan en cruzar la meta. Para ello se toma la medición del tiempo promedio que le tomó a cada grupo de edad llegar a la meta. Se muestran en la tabla las mediciones de tiempo promedio en minutos y grupo de edad para las mujeres participantes y en la gráfica la dispersión del tiempo promedio de recorrido (en minutos) *versus* la edad representativa de cada grupo.

Grupo de edad	Edad representativa	Tiempo promedio
10–19	15	302.38
20–29	25	193.63
30–39	35	185.46
40–49	45	198.49
50–59	55	224.30
60–69	65	288.71



El coeficiente de correlación de la población

El coeficiente de correlación muestral r—que mide la fuerza de la relación de los valores x y los y en una muestra de parejas ordenadas que están linealmente relacionadas entre sí—no es una medida análoga de la fuerza con que x y y están relacionadas en toda la población de parejas ordenadas de la cual se obtuvo la muestra. Al coeficiente que mide la fuerza con que x y y están relacionadas en toda la población de parejas ordenadas se le conoce como el coeficiente de correlación de la población y se define como ρ-rho- (nótese de nuevo el uso de una letra griega para nombrar una característica de la población y una letra romana para designar una característica de la muestra). Aunque nunca tendrás que calcular ρ de una población total de pares ordenados, es importante saber que ρ satisface propiedades paralelas a las de r:

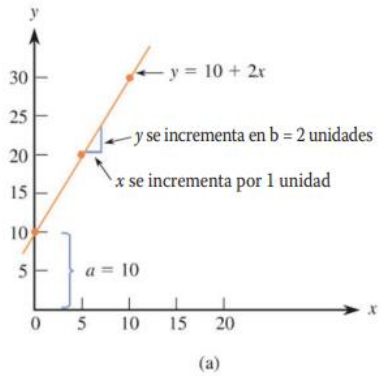
1. ρ es un número entre -1 y 1, que no depende de las unidades de medición de x y y, ni de cuál variable que se denomine como x y cuál se denomine como y.
2. $\rho = -1$ o $\rho = 1$ si y sólo si todos los pares (x, y) en la población se encuentran exactamente sobre una línea recta, por lo que mide el grado en que existe una relación lineal entre la población.

Regresión lineal: ajuste de una recta a los datos bivariados.

El objetivo del análisis de regresión es el uso de la información sobre una variable x, con la finalidad de obtener alguna conclusión acerca de una segunda variable, y. Por ejemplo, podríamos tratar de predecir y = la proporción de enfermos de un mal respiratorio, cuando la temperatura ambiente promedio se encuentra en $x = 2°C$. Las dos variables en un análisis de regresión desempeñan funciones diferentes: se denomina a la variable y dependiente o respuesta; y a la variable x variable independiente, predictora o explicativa. Cuando se observa que una gráfica de dispersión presenta un patrón lineal se analizar la relación entre las variables mediante la búsqueda de una línea que esté lo más cerca posible de los puntos descritos en la gráfica de dispersión.

Características elementales de las lineales y las relaciones lineales

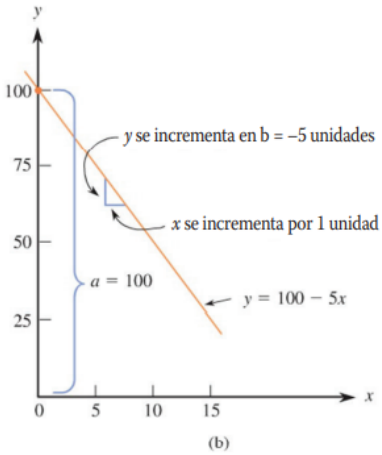
La ecuación de una recta es $y = a + bx$. Una línea en particular se determina eligiendo los valores de a y b. Por ejemplo, una recta es $y = 10 + 2x$ y otra es $y = 100 - 5x$. Si se dan valores a la variable x y se calcula $y = a + bx$ para cada uno de estos valores, las parejas ordenadas resultantes (x, y), pertenecen a una línea recta.



(a)

En la ecuación de una línea recta
$y = a + bx$
A la cantidad b, se le llama la pendiente la línea recta. Representa la cantidad en que se incrementa la variable y cuando x se incrementa en 1 unidad. El valor a se conoce como la intersección de la línea recta con el eje y y representa la altura que toma la línea por encima del valor x = 0.

La recta y = 10 + $2x$ tiene pendiente $b = 2$ por lo que cada incremento de 1 unidad en x está asociado con un aumento de 2 unidades en y. Cuando x = 0, y = 10, por lo que la altura a la que esta línea cruza el eje vertical (cuando x = 0) es 10. Esto se ilustra en la gráfica (a). La pendiente de la línea y =100-5x es igual a -5, así que y decrece en 5 cuando x se incrementa en 1. La altura de la línea cuando x = 0, se encuentra en 100. La línea de resultante se representa en la gráfica (b)

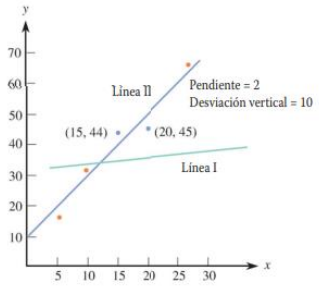


(b)

Seleccionamos dos valores de x, cualesquiera, y los sustituimos en la ecuación para obtener los valores correspondientes de la variable y. Si representamos los valores resultantes como parejas de puntos (x, y), la línea deseada es la que pasa a través de estas dos parejas de puntos. Para la ecuación y = 10 + 2 x, si sustituimos los valores x = 5 y x = 10 obtenemos los valores para la variable y = 20 y y = 30, respectivamente. Así, las parejas ordenadas resultantes son (5, 20) y (10, 30), por lo que la línea recta representada por esta ecuación es la que pasa por estos dos puntos.

El ajuste de una línea recta: principio de mínimos cuadrados.

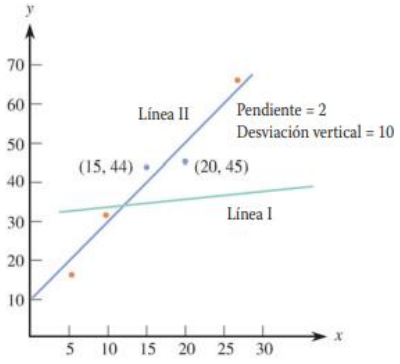
En la figura de la derecha se muestra una gráfica de dispersión con dos líneas superpuestas entre la nube de puntos. La línea II se ajusta mejor a los datos que la línea I. Para medir el grado en que una línea particular proporciona un buen ajuste a los datos, nos centramos en las desviaciones verticales de la línea. Por ejemplo, la línea II tiene como ecuación $y = 10 + 2x$ y los puntos tercero y cuarto a la izquierda en el diagrama de dispersión son (15, 44) y (20, 45). Para estos dos puntos, las desviaciones verticales de esta línea son:



Tercera desviación $= y3 -$ *altura de la línea por encima de* $x3 = 44 - [10 + 2(15)] = 4$

y Cuarta desviación $= 45 - [10 + 2(20)] = -5$

Una desviación vertical positiva corresponde a un punto situado por encima de la línea elegida; se produce una desviación vertical negativa si el punto correspondiente está situado por debajo de dicha línea. Una línea en particular se dice que es un buen ajuste de los datos, si las desviaciones de la línea son de pequeña magnitud. La línea I de la gráfica que se muestra a la derecha se adapta mal, porque todas las desviaciones de la línea son más grandes que las de la línea II. Para evaluar el ajuste global de una línea (bondad de ajuste) en un conjunto de n mediciones, necesitamos una manera de combinar las n desviaciones en una sola medida de ajuste. El método estándar consiste en tomar los cuadrados de las desviaciones (para obtener un número no negativo) y luego considerar la suma de los cuadrados de estas desviaciones.



La medida más utilizada de la bondad del ajuste de una línea a un conjunto de datos bivariados $(x_1y_1), (x_2y_2), \ldots\ldots, (x_n,y_n)$
es la suma de los cuadrados de las desviaciones alrededor de la línea
$\sum[y - (a + bx)]^2 = [y_1 - (a + bx_1)]^2 + [y_2 - (a + bx_2)]^2 + \ldots + [y_n - (a + bx_n)]^2$

La línea de mínimos cuadrados, también llamada línea de regresión muestral, es la que minimiza la suma de los cuadrados de las desviaciones.

La ecuación de la recta de mínimos cuadrados se puede obtener sin necesidad de calcular las desviaciones de cualquier línea en particular. En el siguiente cuadro están las fórmulas relativamente simples para la pendiente y la intersección de la línea de mínimos cuadrados.

La pendiente de la línea de los mínimos cuadrados es=

$$\frac{\sum(x-\bar{x})\ (y-\bar{y})}{\sum(x-\bar{x})^2}$$

Y la intersección es:

$$a-\bar{y}\text{-}b\bar{x}$$

Se escribe la ecuación de mínimos cuadrados como la línea

$$\hat{y}=a+bx$$

donde la ˆ arriba de y (que se lee $y\ gorro$) indica que $\hat{y}$ es la predicción de y que resulta de sustituir un valor particular de x en la ecuación

Cuando los cálculos para obtener la pendiente y la intersección de la línea de mínimos cuadrados no se resuelvan en hojas de cálculo o calculadoras, las fórmulas para obtenerlo son las siguientes:

Fórmula para calcular la pendiente de la línea de los mínimos cuadrados

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

Analiza los ejemplos de las páginas 137 a la 140 de tu libro de texto para reforzar tu habilidad en la resolución de este tipo de cálculos estadísticos.

¡Ahora te toca a ti!

El coeficiente de correlación proporciona información adicional que, junto con la evidencia visual, permite aceptar o desechar la relación causa-efecto planteada entre dos variables y que además te permiten aceptar o refutar las hipótesis planteadas.

Actividad 2.3

Resuelve adecuadamente el siguiente ejercicio, sobre la recta de regresión y el coeficiente de correlación lineal.

4. Se registran en una tabla las horas trabajadas en un taller (x) y las unidades producidas (y) Determina la recta de regresión de (y) sobre (x), el coeficiente de correlación lineal e interprétalo.

Horas (x)	Producción (y)
80	300
79	302
83	315
84	330
78	300
60	250
82	300
85	340
79	315
84	330
80	310
62	240

Autoevaluación Unidad 2

1. Para aplicar adecuadamente las medidas de tendencia central, la comunidad debe tener presente las definiciones y características de estas medidas. Vincula los conceptos con sus definiciones y características relacionando las columnas. Selecciona entre las cuatro opciones la respuesta correcta.

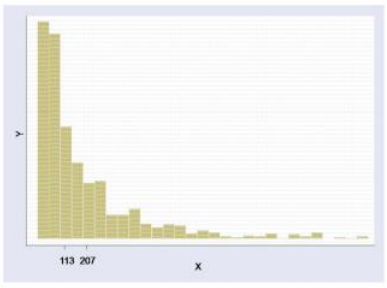
1. Desventaja de la media muestral 2. Mediana muestral 3. Medida de tendencia central 4. Media muestral 5. Es una característica de la mediana muestral	a) Describe aproximadamente donde están localizados o centrados los datos a lo lago la recta numérica. b) Su valor puede verse muy afectado por la presencia de valores extremos o atípicos. c) Se define como la cantidad que divide al conjunto de datos en dos partes iguales. d) Se define como el resultado del cociente de la suma de las observaciones dividida entre el número de observaciones. e) Su valor no se ve afectado por la presencia de valores extremos.

a) 1a; 2b; 3c; 4d; 5e

b) 1c; 2e; 3a; 4d; 5b

c) 1c; 2d; 3a; 4b; 5e

d) 1b; 2c; 3a; 4d; 5e



3 Para decidir qué medida de tendencia central utilizar en el estudio de la distribución del ingreso de las mujeres, la comunidad calculó el valor de la media y la mediana a partir de los valores obtenidos en una muestra representativa. Los valores obtenidos para la media y la mediana fueron de $207.00 y $113.00, respectivamente. Junto a estos valores graficó su respectivo histograma de frecuencias el cual se muestra a tu izquierda

Con base en el análisis de la gráfica selecciona una respuesta sobre la medida de tendencia central más adecuada a utilizar como medida representativa de la información disponible.

a) La medida de tendencia central más adecuada es la media muestral.
b) La medida de tendencia central más adecuada es la mediana muestral.
c) Es indistinto, cualquiera de las dos medidas puede ser tomada como representativa de la información.
d) Ninguna de las dos medidas.

4 Para aplicar adecuadamente las medidas de dispersión la comunidad elabora un cuadro con los conceptos y las definiciones básicas. Identifica estas definiciones y características relacionando las siguientes columnas, vinculando los conceptos con sus definiciones respectivas. Selecciona entre las cuatro opciones de respuesta desarrolladas al final la que consideres es la correcta.

1. Desventaja del rango muestral como medida de dispersión	a. Medida que permite cuantificar que tanto las observaciones difieren unas de otras.
2. Desviación estándar muestral	b. Se define como el resultado de la diferencia entre la observación más grande y la observación más pequeña de un conjunto de datos.
3. Medida de dispersión	c. Para una muestra de tamaño n se define como la suma de los cuadrados de las desviaciones de la media divididas entre n-1.
4. Varianza muestral	d. Se obtiene como resultado de la raíz cuadrada de la varianza.
5. Rango muestral	e. Tiene la desventaja de no considerar en su cálculo la contribución de cada observación a la variabilidad.

a) 1e; 2d; 3a; 4c; 5b
b) 1c; 2e; 3a; 4d; 5b
c) 1c; 2d; 3a; 4b; 5e
d) 1b; 2c; 3a; 4d; 5e

4. En una región de difícil acceso, la comunidad de investigación se vio forzado a tomar dos muestras aleatorias de tamaño n=10 del ingreso diario en pesos de mujeres y hombres, con la finalidad de realizar un comparativo de la variabilidad en el ingreso de hombres y mujeres de este distrito. Los valores obtenidos del ingreso para ambas muestras aleatorias son presentados en las siguientes dos tablas:

Ingreso diario en pesos de 10 mujeres seleccionadas al azar				
$55.00	$38.00	$41.00	$54.00	$48.00
$48.00	$48.00	$52.00	$63.00	$57.00

Ingreso diario en pesos de 10 hombres seleccionados al azar				
$47.00	$75.00	$35.00	$78.00	$46.00
$63.00	$85.00	$73.00	$59.00	$40.00

Para cada conjunto de datos calcula la desviación estándar y selecciona la respuesta que da una adecuada interpretación a los resultados obtenidos; puedes utilizar una hoja de cálculo.

a) A partir de los valores calculados para la desviación estándar de ambas muestras se puede concluir que existe evidencia que nos hace suponer que entre los hombres existe mayor variabilidad en sus ingresos diarios que entre las mujeres.

b) A partir de los valores calculados para la desviación estándar de ambas muestras se puede concluir que existe evidencia que nos hace suponer que entre las mujeres existe mayor variabilidad de ingresos diarios que entre los hombres.

c) A partir de los valores calculados para la desviación estándar de ambas muestras no se puede obtener algún tipo de conclusión.

d) A partir de los valores calculados para la desviación estándar de ambas muestras se puede concluir que existe evidencia que nos hace suponer que la variabilidad en las percepciones es la misma para hombres y mujeres.

5. La variable x_1 = precipitación pluvial fue identificada como una variable cuya distribución de frecuencia se puede modelar de forma adecuada con una distribución de probabilidad normal con media $\mu = 121.9\ millones\ de\ m^3$y desviación estándar $\sigma = 50.3\ millones\ de\ m^3$. Utilizando esta distribución de probabilidad, ¿cuál es el valor de la probabilidad $P(71.6 \leq x_1 \leq 172.2)$?

a) .1587 b) .3174 c) .6826 d) .5000

6. Para poder medir el impacto social que una tormenta tropical dejo a su paso por la región, el grupo de investigadores seleccionó una muestra de 50 habitantes de la región a partir de la cual midió la variable x_2 =número de habitantes damnificados. El modelo probabilístico adecuado para modelar la distribución de frecuencia de esta variable es una distribución binomial con $n = 50\ p = .3$. Aplicando esta distribución de probabilidad, ¿cuál es el valor de la probabilidad $P(25 \leq x_2)$? Recuerda que puedes utilizar la hoja de cálculo para resolver este ejercicio.

a) .9976 b) .0032 c) .9968 d) .0024

7. Para poder medir el impacto que los huracanes tienen sobre la región. El equipo de investigadores recabó la información necesaria para suponer que la variable x_3 = número de huracanes por año, puede moldearse de manera adecuada por medio de una distribución de probabilidad de Poisson con $\lambda = 15$.

Identifica los supuestos o hipótesis que el grupo de investigadores tomó como base seleccionando de la siguiente lista aquellos que hacen que el evento que da origen a las observaciones de esta variable pueda ser considerada como un experimento de Poisson. 1. El número promedio de huracanes que ocurren en la región por unidad de tiempo (para este evento la unidad de tiempo será considerada un año) es constante.

a) La variable x_3 es una variable numérica discreta
b) La probabilidad de que más de un huracán ocurra en la región en cualquier intervalo de tiempo breve es casi cero. Esto significa que la probabilidad de que dos o más huracanes lleguen a ocurrir en la región durante un intervalo de tiempo breve por ejemplo de un día es muy pequeña o casi cero
c) 4. La variable x_3 toma un número finito de valores

El número de huracanes que suceden en la región en intervalos ajenos de tiempo son independientes unos de otros. Es decir que el número de huracanes que ocurrió en un intervalo concreto de tiempo no influye sobre el número de huracanes que serán observados en la región durante un próximo intervalo de tiempo. Selecciona la respuesta correcta de entre las cuatro opciones de respuesta que se te muestran a continuación:

a) 1; 3; 5 b) 2; 4; 5 c) 1; 2; 3 d) 1; 3; 4

8. Utilizando esta distribución de probabilidad de Poisson con $\lambda = 15$, ¿cuál es el valor de la probabilidad $P(12 \leq x_3 \leq 14)$? Puedes utilizar una hoja de cálculo para realizar esta operación.

a) .0829 b) .1785 c) .2809 d) .0024

Un grupo de investigación pretende identificar la relación entre las actividades del ser humano y el cambio climático, a partir del vínculo entre fenómenos como la sequía, aridez de una región o la degradación de recursos naturales y su efecto socioeconómico. Por la naturaleza de la investigación planteada, tendrán que recurrir a las gráficas de dispersión, coeficiente de correlación y líneas de regresión lineal, como herramientas que les permitan indagar la relación entre variables. Resuelve las siguientes preguntas para saber qué es lo que el grupo de investigación debe saber acerca de los conceptos y aplicaciones de la gráfica de dispersión, coeficiente de correlación y línea de regresión lineal para adecuarlos a su problema de investigación.

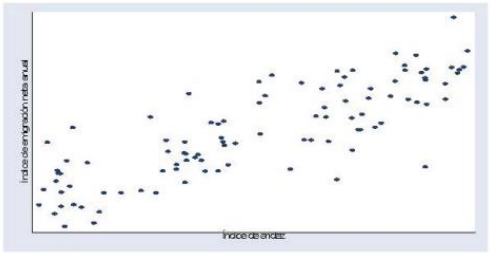
9. Para la aplicación correcta de la gráfica de dispersión, el coeficiente de correlación y la línea de regresión es importante que el investigador conozca las definiciones y características de éstas. Identifícalas relacionando las siguientes columnas. Al final selecciona la opción que consideres correcta.

1. Gráfica de dispersión 2. Coeficiente de correlación muestral 3. Desventaja del coeficiente de correlación como medida de asociación 4. Línea de regresión 5. Tendencia crecientes 6. Tendencia decreciente	a. Tiene la característica de ser una línea recta que está lo más cerca posible de los puntos descritos en una gráfica de dispersión. b. Tendencia que se caracteriza por observar incrementos en la variable Y como resultado de incrementos en la variable X. c. Coeficiente que sirve para medir la posible relación lineal entre dos variables. d. Tendencia que se caracteriza por observar decrementos en la variable Ycomo resultado de incrementos en la variable X. e. Técnica que utiliza un conjunto de registros apareados de la forma (X, Y), con X como variable independiente y Y como variable dependiente, a partir de los que se elabora una gráfica conformada por parejas ordenadas en el plano cartesiano con base en la identificación del valor de la variable X respecto al eje horizontal mientras que el valor de la variable Y se identifica con el eje vertical. f. Tiene la desventaja de no poder detectar relaciones de asociación no lineales entre dos variables

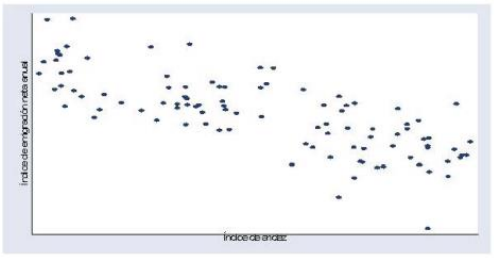
a) 1a; 2b; 3c; 4d; 5e; 6f
b) 1e; 2c; 3f; 4a; 5b; 6d
c) 1e; 2c; 3a; 4b; 5d; 6f
d) 1b; 2c; 3a; 4d; 5e; 6f

10. Durante el desarrollo de la investigación el equipo determinó la presencia de una tendencia creciente entre las variables índice de aridez e índice de emigración neta anual. Detectó la presencia de esta tendencia a partir de la construcción de una gráfica de dispersión junto con el cálculo del coeficiente de correlación muestral asociado. A partir de las gráficas de dispersión y de los coeficientes de correlación calculados selecciona de entre las cuatro opciones que se te presentan, la respuesta correcta que combine la gráfica de dispersión y el coeficiente de correlación que permitió detectar esta tendencia creciente entre las variables.

a.



b.

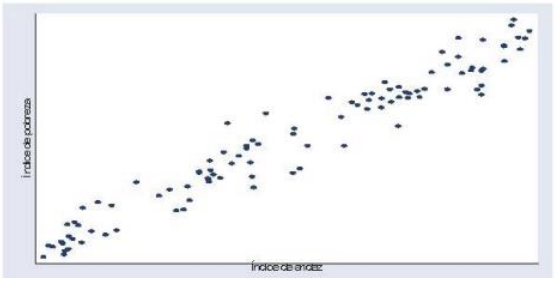


1. r = -0.76
2. r = 0.82

a) 1; b
b) 2; b
c) 1; a
d) 2; a

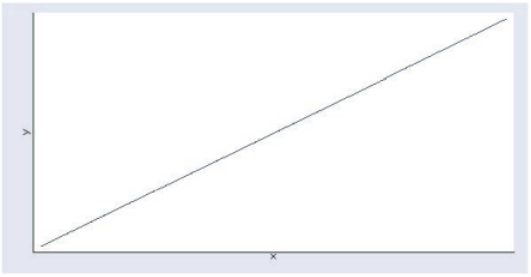
11. Para ajustar una línea de regresión que permita predecir el índice de pobreza como función del índice de aridez, con los datos recabados para ambas variables, el equipo de investigación realizó un diagrama de dispersión con el fin de definir qué tipo de línea de regresión es la más adecuada para representar la tendencia que se observa en el diagrama de dispersión.

La gráfica del diagrama de dispersión elaborada por el equipo de investigación se muestra a la derecha.

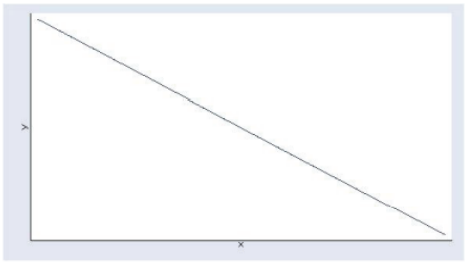


Con base en la gráfica de dispersión determina cuál de las siguientes líneas es la más adecuada para representar la línea de regresión de los datos mostrados en la gráfica de dispersión. Observa las siguientes gráficas:

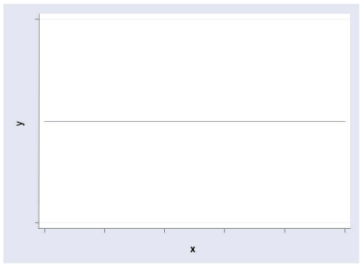
1.



2.



3.



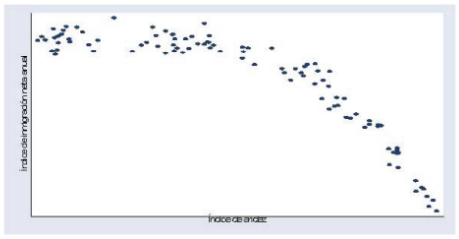
Selecciona la respuesta correcta:

a) 1.
b) 3.
c) 2.
d) Ninguna de las anteriores.

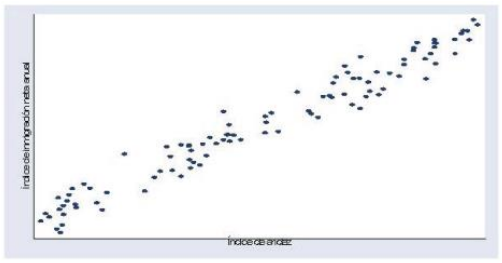
10. Al estudiar la relación entre las variables índice de aridez e índice de inmigración neta anual, el grupo de investigación se percató que el diagrama de dispersión correspondiente a las observaciones pareadas de ambas variables presentó una tendencia decreciente no lineal. Identifica el tipo de gráfica de dispersión que el grupo de investigación observó, seleccionando cuál de las cuatro gráficas de dispersión mostradas muestra una tendencia decreciente no lineal

a) 3 b) 2 c) 1 d) 4

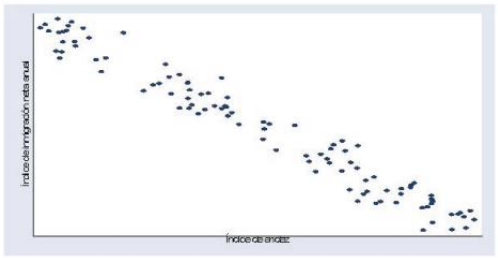
1.



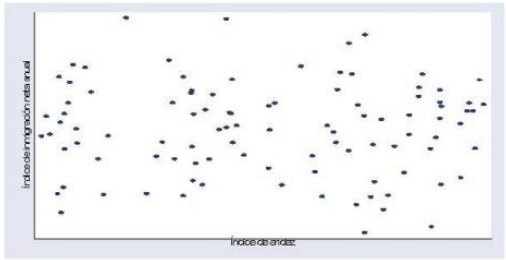
2.



3.



4.



Respuestas de autoevaluaciones

Respuestas de autoevaluación Unidad 1

1. b) 3,6,7,11 y 12	5. b) 1a; 2d; 3b; 4c
2. a) 1,2,4,5,8,9 y 10	6. c) 1c; 2d; 3ª; 4b; 5e
3. d) 1, 3 y 9	7. c) 1
4. c) 2,4,5,6,7 y 8	8. a) 1; 3; 5; 6

Respuestas de autoevaluación Unidad 2

9. d) 1b; 2c; 3c; 4d; 5e	15. a) 1; 3; 5
10. b) La mediana muestral.	16. c) .2809
11. a) 1e; 2d; 3a; 4c; 5b	17. b) 1e; 2c; 3f; 4a; 5b; 6d
12. a)	18. d) 2; a
13. c)	19. a) 1
14. d) .0024	20. c) 1

Soluciones de actividades

Unidad 1

Actividad 1.1

Fenómeno	Clasificación
Huracanes	Fenómeno natural
Inseguridad	Proceso social
Lluvias	Fenómeno natural
Escolaridad en la comunidad	Proceso social
Obesidad en la comunidad	Proceso social
Inundaciones	Fenómeno natural
Dengue en la comunidad	Fenómeno natural
Temperatura	Fenómeno natural
Pobreza y marginación	Proceso social
Sequía	Fenómeno natural
Crecimiento demográfico	Proceso social
Epidemias	Fenómeno natural
Migración	Proceso social
Inmigración	Proceso social
Enfermedades en la comunidad	Fenómeno natural
Natalidad en la comunidad	Proceso social
Servicios de salud en la comunidad	Proceso social
Virus del Covid 19	Fenómeno natural

2.

Fenómeno Natural	Proceso Social
Temperatura	Servicios de salud en la comunidad
Lluvias	Escolaridad en la comunidad
Huracanes	Crecimiento demográfico
Enfermedades en la comunidad	Pobreza y marginación

Actividad 1.2

Color	Frecuencia absoluta	Frecuencia relativa
Negro	4	0.20
Azul	5	0.25
Amarillo	5	0.25
Rojo	6	0.30
Total	20	1

Actividad 1.3

Variables Cuantitativas		Variables Cualitativa	
Continuas	**Discretas**	**Ordinales**	**Nominales**
Temperatura media anual en la comunidad Porcentaje anual de habitantes que se contagiaron de dengue durante el verano Porcentaje de habitantes que se contagiaron de alguna enfermedad respiratoria durante el invierno Porcentaje de familias cuyo sustento económico es obtenido en actividades derivadas de la comercialización de productos del mar Porcentaje anual de la población que ha migrado hacia la comunidad Porcentaje anual de la población que ha emigrado fuera de la comunidad Índice de analfabetismo en la comunidad Presión sanguínea sistólica registrada en las personas que acuden a la clínica de la comunidad	Número de altas semanales en la clínica de la comunidad Número diario de camas disponibles en la clínica de la comunidad Número anual de huracanes y tormentas tropicales que han impactado a la comunidad Número diario de personas detectadas con diabetes en el hospital de la comunidad.	Escolaridad de los habitantes mayores de 18 años	Sexo de los habitantes de Santiago Tipo de enfermedades más frecuentes Tipo de accidente por el que ingresan los pacientes a la sala de emergencia en la clínica de la comunidad

Variables Independientes	Variable Dependiente
Número anual de huracanes y tormentas tropicales que han impactado a la comunidad	Porcentaje anual de habitantes que se contagiaron de dengue durante el verano.
Temperatura media anual en la comunidad	Porcentaje de habitantes que se contagiaron de alguna enfermedad respiratoria durante el invierno.
Sexo de los habitantes de Santiago	Tipo de enfermedades más frecuentes

Actividad 1.4

1.

2.a) muestreo aleatorio simple

b) muestreo aleatorio estratificado

c) muestreo aleatorio simple

d) muestreo por conveniencia

e) muestreo aleatorio estratificado

f) muestreo por conveniencia.

3. a) Se numera la lista de alumnos del 1 al 20. Se generan números aleatorios para seleccionar los 4 alumnos. (Si cuentas con Excel puedes utilizarlo).

b) Por ejemplo, si se seleccionan los números 10, 1, 11, 20, la muestra la conforman Victoria, Juan, María, Marcelo (3 mujeres y 1 hombre). Trabajan: Victoria, Juan, Marcelo No trabajan: Mar. Entonces:

El parámetro o porcentaje de alumnos que trabajan en la población de tamaño N=20 alumnos, es decir: P = no. de personas que trabajan / N = 7 / 20 = 0.35 ó 35%

El estadístico o porcentaje de alumnos que trabajan en la muestra de tamaño n=4 alumnos, es decir: P = no. de personas que trabajan / n = 3 / 4= 0.75 ó 75%

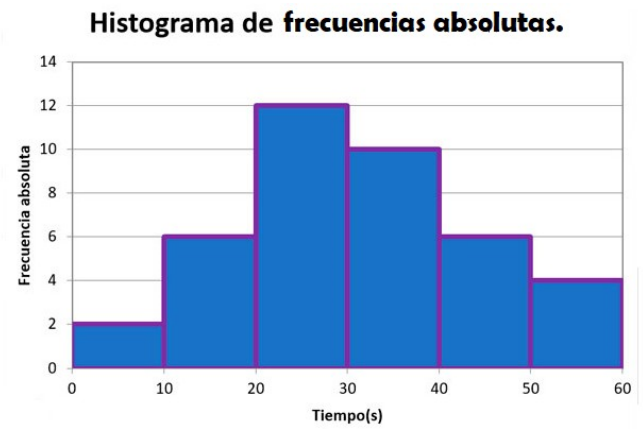
c) Dos criterios de estratificación para esta muestra pueden ser dividir en dos grupos: hombres y mujeres y quienes trabajan en empresa privada y quienes en pública.

Actividad 1.5

1

x_i	f_i	F_i	N_i
27	1	1	0.032
28	2	3	0.097
29	6	9	0.290
30	7	16	0.516
31	8	24	0.774
32	3	27	0.871
33	3	30	0.968
34	1	31	1
	31		

2



3.

a) Variable cuantitativa continua

b) Variable cuantitativa discreta

c) Variable cuantitativa continua

d) Variable cuantitativa discreta

e) variable cualitativa ordinal (no incluida)

4.

Similitudes	Diferencias
Ambas distribuciones se pueden usar para modelar el número de ocurrencias de algún evento. En ambas distribuciones, se supone que los eventos son independientes.	En una distribución binomial, hay un número fijo de intentos (por ejemplo, lanzar una moneda 3 veces). En una distribución de Poisson, podría haber cualquier número de eventos que ocurran durante un cierto intervalo de tiempo (por ejemplo, ¿cuántos clientes llegarán a una tienda en una hora determinada?).

Unidad 2

Actividad 2.1

a) Moda: (el valor que tiene mayor frecuencia absoluta).
En la columna f_i, el mayor valor es 42 y corresponde a x_i= 67. Moda = 67
b) Mediana: N=100/2 = 50. El valor más cercano es 67. Mediana= 67

c) Media: $\bar{\bar{x}} = \frac{\sum x}{n} = \frac{6745}{100} = 67.45$

d) Rango (diferencia entre el mayor y el menor de los valores)

$$r = 73 - 61 = 12$$

e) Varianza

$$\sigma^2 = x_1^2 f_1 = \frac{455803}{n} - media^2 \qquad \sigma^2 = \frac{455803}{100} - 67.45^2 = 8.53$$

f) Desviación estándar (raíz cuadrada de la varianza. $\sigma = \sqrt{8.53} = 2.95$

x_i	f_i	F_i	$x_i f_i$	$\lvert x_i - \bar{x}\rvert$	$\lvert x_i - \bar{x}\rvert f_i$	$x_i^2 f_i$
61	5	5	305	6.45	32.25	18605
64	18	23	1152	3.45	62.10	73728
67	42	65	2814	0.45	18.90	188538
71	27	92	1890	2.55	68.85	132300
73	8	100	584	5.55	44.40	42632
	100		6745		226.50	455803

Actividad 2.2

1. Distribución normal estándar

$$Z = \frac{X - \mu}{\sigma} \qquad P(21 \le X \le 27) = P\left(\frac{(21-23)}{5} \le Z \le \frac{(27-23)}{5}\right) = P(Z \le 0.8) - P(Z \ge -0.4)$$

$=P(Z \le 0.8) - (1 - P\ (Z \le 0.4))$. Se buscan los valores correspondientes en la tabla de distribución normal:

$$P(21 \le X \le 27) = P\left(\frac{(21-23)}{5} \le Z \le \frac{(27-23)}{5}\right) = (30)\big(0.7881 - (1 - 0.6554)\big) = \quad (30)(0.4435) = 13$$

Significa que, en todo el mes, solo 13 días alcanzarán temperaturas entre 21° y 27°.

2. Distribución binomial

a) $B(100, 0.03) \qquad p = 0.03 \qquad q = 0.97$

$$p(x=0) = \binom{5}{0} \cdot 0.97^5 = 0.8587$$

b) $p(x \geq 2) = 1 - p(x < 2) = 1 - [p(x=0) + p(x=1)]$

$$= 1 - \left[\binom{5}{0} 0.97^5 + \binom{5}{1} 0.03 \cdot 0.97^4\right] = 0.00847$$

c) $\mu = 100 \cdot 0.03 = 3$

3. Distribución de Poisson: $P(x) = \frac{\lambda x e - \lambda}{x!}$

a) variable aleatoria: ***X*** = *número de pacientes que llegan en un día.* 3 en este caso f (3)

$$\lambda = 4 \qquad f(3) = P(X=3) = \frac{4^4.e^{-4}}{3!} \qquad f(3) = P\,(X=3) = \frac{0.01832.64}{3.2.1}$$

f(3) = P(X = 3) = 0.1954 o 19.54%

b) Probabilidad de que lleguen 5 pacientes en un día.

$$f(5) = P(X=5) = \frac{4^5.e^{-4}}{5!} \qquad f(5) = P(X=5) = \frac{0.01832.1024}{5.4.3.2.1} \qquad f(5) = P(X=5) = 0.1563 = 15.63\%$$

Actividad 2.3

1.

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
80	300	6400	90000	24000
79	302	6241	91204	23858
83	315	6889	99225	26145
84	330	7056	108900	27720
78	300	6084	90000	23400
60	250	3600	62500	15000
82	300	6724	90000	24600
85	340	7225	115600	28900
79	315	6241	99225	24885
84	330	7056	108900	27720
80	310	6400	96100	24800
62	240	3844	57600	14880
936	3632	73760	1109254	285908

Se calculan los promedios: $\bar{x} = \frac{936}{12} = 78$ $\bar{y} = \frac{3632}{12} = 302.67$

Se calcula la covarianza, la varianza y las desviaciones estándares.

$$\sigma_{xy} = \frac{285908}{12} - (78)(302.76) = 217.41$$

$$\sigma_x^2 = \frac{737660}{12} - (78)^2 = 62.67 \quad \Rightarrow \quad \sigma_x = \sqrt{62.67} = 7.92$$

$$\sigma_y^2 = \frac{1109254}{12} - (302.67)^2 = 827.7 \quad \Rightarrow \quad \sigma_y = \sqrt{828.7} = 28.8$$

El coeficiente de correlación está dado por $r = \frac{\sigma_{xy}}{\sigma^2}$

y

$$r = \frac{\sigma_{xy}}{\sigma_x \, . \sigma_y} = \frac{217.41}{(7.92)(28.8)} = 0.95$$

0.95 indica una correlación positiva muy fuerte.

La recta de regresión de (x) sobre (y) es aquella que pasa por el punto $(\bar{x}, \bar{y})$ y tiene pendiente $\frac{\sigma_{xy}}{\sigma_x \, . \sigma_y}$ $y - 302.67 = 3.47(x - 78)$ despejando (y) se obtiene la

recta de regresión, $y = 3.47x + 32.01$

Nos complace anunciarte que has llegado al final de tu módulo, ¿crees estar preparado para el siguiente reto?

Pon a prueba tus conocimientos, compara las respuestas de tus actividades con las soluciones que ofrece la última sección de esta guía. Si tu resultado no es aprobatorio, ¡no te preocupes!, puedes regresar a los recursos del libro para reforzar los contenidos que necesites volver a retomar y así acreditar el examen oficial.

Felicidades por llegar hasta aquí, siendo un aprendizaje independiente el éxito es tuyo.


EL GOBIERNO DEL NUEVO
2021
2027
NUEVO LEÓN