seg.guanajuato.gob.mx

GTO Grandeza de México

Secretaría de Educación

Estadística en fenómenos naturales y procesos sociales

Guía de Estudio

PREPARATORIA ABIERTA
GUANAJUATO



Este material es propiedad de la Secretaría de Educación de Guanajuato y están protegidos por la LEY FEDERAL DEL DERECHO DE AUTOR bajo el Registro 03-2017-070611482200-01 y demás disposiciones internacionales en la materia. Esta guía tiene como único propósito que los usuarios lo consulten con finalidades educativas. Los usuarios no están autorizados para transmitir ninguno de estos materiales por vía electrónica, o reproducirlos en forma alguna por ningún medio, electrónico, mecánico, sistemas de recuperación, grabación, impresión, fotocopiado, etc.

Secretaría de Educación de Guanajuato
Yoloxochitl Bustamante Diez

Subsecretaría para el Desarrollo Educativo
José Abel Arteaga Sánchez

Dirección General de Media Superior
Ximena María Alarcón Verduzco

Dirección de Preparatoria Abierta y a Distancia
Alejandro Carretero Carretero

Departamento Académico de la Dirección de Preparatoria Abierta y a Distancia
Vania Ruth M. Castro Ortiz

Autoría
Yolanda del Pilar López Mejía

ÍNDICE

Unidad 1. La estadística descriptiva y los fenómenos naturales y procesos sociales 5

Tipos de eventos: determinísticos y aleatorios 7

Fenómenos naturales y procesos sociales determinísticos y aleatorios 8

Concepto de probabilidad 8

Asignación de la probabilidad a un evento aleatorio 9

Las variables como instrumento de medición de la incertidumbre y la variabilidad de un fenómeno aleatorio 13

Variables cuantitativas 14

Variables cualitativas 15

Variables categóricas nominales y ordinales 15

Clasificación de las variables por su grado de asociación 15

Muestreo: población, muestra y técnicas 16

Técnicas de muestro y su importancia en el análisis de la información 16

Censos y muestras 16

Concepto de muestra 17

Representatividad de una muestra 17

Métodos de muestreo 18

Selección de una muestra aleatoria simple 18

Muestreo aleatorio estratificado 19

Muestreo de conveniencia 20

Las distribuciones de probabilidad y su relación con las variables relacionadas con fenómenos naturales y procesos sociales 21

Distribuciones de frecuencias para variables categóricas 21

Histogramas de frecuencias 21

Histogramas para variables numéricas discretas 22

Histograma para una variable continua 23

Distribuciones de frecuencias para variables numéricas discretas. 25

Distribuciones de frecuencias para variables numéricas continuas 26

Distribuciones de probabilidad continua 26

Distribuciones de probabilidades teóricas 27

Distribución binomial 28

Distribución de Poisson 29

Unidad 2. Tratamiento estadístico de la información de fenómenos naturales y procesos sociales 31

La estadística descriptiva y su aplicación en la explicación de los fenómenos naturales y los procesos sociales 32

Medidas de tendencia central: media y mediana 32

La media 33

La mediana 34

Comparación de la media y la mediana 35

Medidas de dispersión 37

Desviación de la media 37

Varianza y desviación estándar 38

Las distribuciones de probabilidad como herramientas para estimar las posibilidades de eventos relacionados con fenómenos naturales y procesos sociales 40

Distribución de probabilidad normal 40

La distribución normal estándar 42

Otras distribuciones normales 46

Encontrando probabilidades de una distribución normal 47

Distribución de probabilidad binomial 48

Distribución de probabilidad de Poison 50

El modelo de regresión y el de correlación lineal como medidas para describir la asociación entre variables 52

Correlación 52

Coeficiente de correlación muestral de Pearson 52

Coeficiente de correlación de la población 55

El coeficiente de correlación de la población 56

Regresión lineal: ajuste de una recta a los datos bivariados 57

Características de las lineales y las relaciones lineales 57

El ajuste de una línea recta: principio de mínimos cuadrados. 57

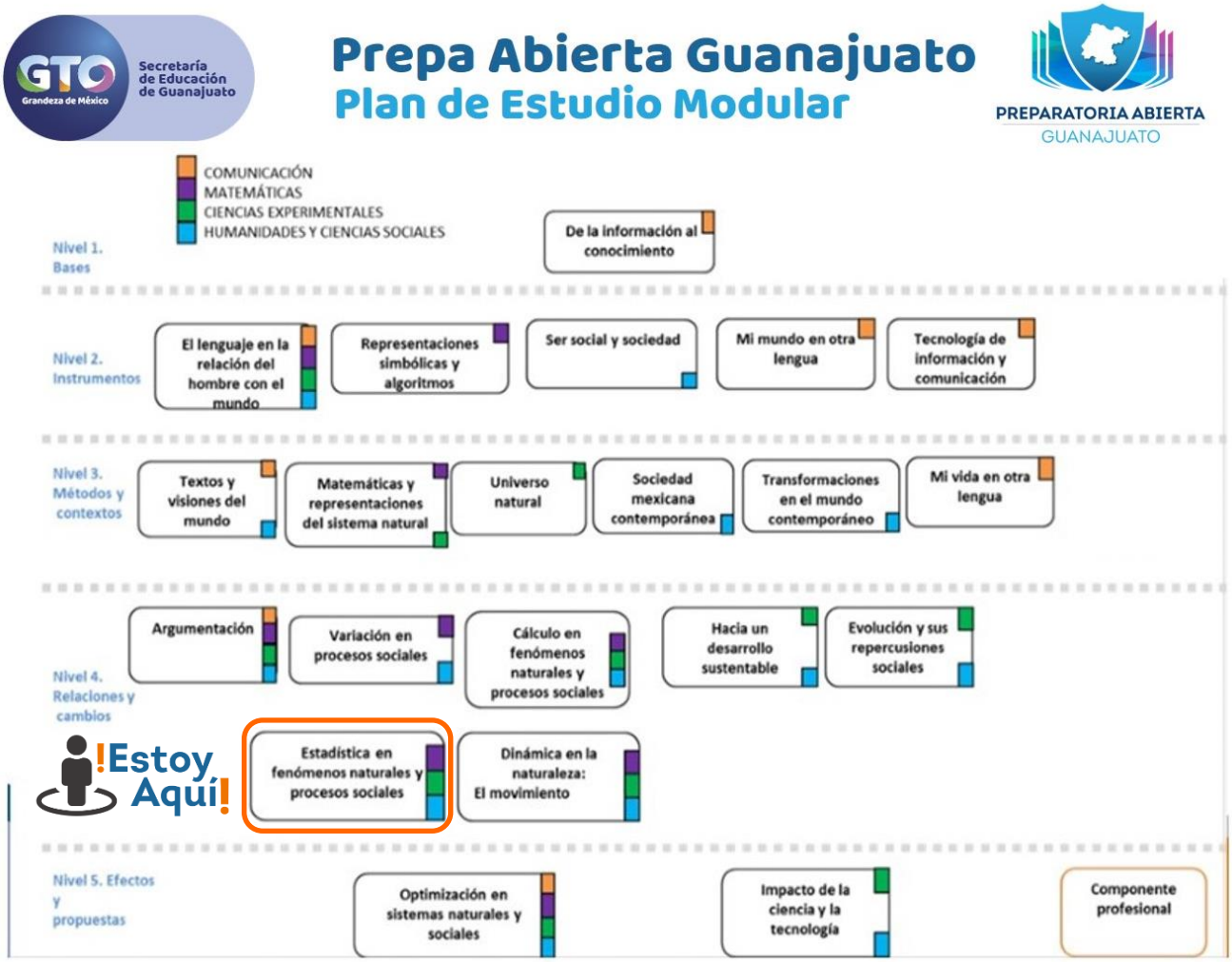
ANEXO: CORRECCIONES DE EDICIÓN AL LIBRO 63

Fuentes consultadas 64

Campo(s) disciplinar(es): Matemáticas, Ciencias experimentales, Humanidades y Ciencias sociales Nivel: IV (Relaciones y cambios)	Módulo: Estadística en fenómenos naturales y procesos sociales. Horas estimadas de estudio: 75 horas
Propósito formativo **Este módulo tiene como propósito, que desarrolles competencias que te permitan:** Registrar, organizar y analizar información factual relativa a los fenómenos naturales y procesos sociales, mediante herramientas estadísticas: tablas, gráficas y modelos matemáticos, para interpretar y explicar su comportamiento y posibilidad de ocurrencia, con el fin de facilitar la toma de decisiones individuales y colectivas en su comunidad, región, país y mundo.	**Competencias disciplinares básicas:** **Al finalizar el módulo deberás contar con las siguientes competencias:** Construye e interpreta modelos matemáticos mediante la aplicación de procedimientos aritméticos, algebraicos, geométricos y variacionales, para la comprensión y análisis de situaciones reales, hipotéticas o formales. Explica e interpreta los resultados obtenidos mediante procedimientos matemáticos y los contrasta con modelos establecidos o situaciones reales. Argumenta la solución obtenida de un problema, con métodos numéricos, gráficos, analíticos o variacionales, mediante el lenguaje verbal, matemático y el uso de las tecnologías de la información y la comunicación. Analiza las relaciones entre dos o más variables de un proceso social o natural para determinar o estimar su comportamiento. Elige un enfoque determinista o uno aleatorio para el estudio de un proceso o fenómeno, y argumenta su pertinencia. Interpreta tablas, gráficas, mapas, diagramas y textos con símbolos matemáticos y científicos. Identifica problemas, formula preguntas de carácter científico y plantea las hipótesis necesarias para responderlas. Obtiene, registra y sistematiza la información para responder a preguntas de carácter científico, consultando fuentes relevantes y realizando experimentos pertinentes.

	Hace explícitas las nociones científicas que sustentan los procesos para la solución de problemas cotidianos. Analiza las leyes generales que rigen el funcionamiento del medio físico y valora las acciones humanas de impacto ambiental. Sitúa hechos históricos fundamentales que han tenido lugar en distintas épocas en México y el mundo con relación al presente. Interpreta su realidad social a partir de los procesos históricos locales, nacionales e internacionales que la han configurado. Valora las diferencias sociales, políticas, económicas, étnicas, culturales y de género y las desigualdades que inducen. Establece la relación entre las dimensiones políticas, económicas, culturales y geográficas de un acontecimiento. Analiza con visión emprendedora los factores y elementos fundamentales que intervienen en la productividad y competitividad de una organización y su relación con el entorno socioeconómico Valora distintas prácticas sociales mediante el reconocimiento de sus significados dentro de un sistema cultural, con una actitud de respeto.

#Ubícate



Previamente

Curso propedéutico de Matemáticas y Módulo 3

Este módulo tiene

mayor relación temática con

Posteriormente

Principalmente con el módulo 20

Antes de empezar

ACTIVIDAD:

- ✓ Realiza la actividad, **¿Con qué saberes cuento?** que aparece en las págs. 15 a 19 del libro de texto correspondiente a este módulo, una vez que termines compara tus respuestas con las que aparecen en el apéndice del mismo libro en la pág.147.

¿Cuáles fueron tus resultados?

Qué tan sencillo te resultó resolver la evaluación diagnóstica ¿con qué saberes cuento?, si te sentiste como pez en el agua, estás listo para comenzar con el estudio de este módulo, recuerda estar monitoreando tu desempeño constantemente, basándote en las rúbricas de evaluación de las actividades y revisando el apéndice de las respuestas correctas que aparece al final de todos los libros de texto, ¡claro!, después de realizar por tí mismo los ejercicios.

Si al contrario, se te dificultó la resolución de este ejercicio de evaluación, será necesario que refresques algunos conocimientos y habilidades. Te recomendamos que revises el libro del **módulo 3 Representaciones simbólicas y algoritmos** de Preparatoria Abierta Guanajuato, que podrás consultar y descargar de manera gratuita en nuestra página de internet, www.seg.guanajuato.gob.mx/prepaabierta.

UNIDAD 1. LA ESTADÍSTICA DESCRIPTIVA Y LOS FENÓMENOS NATURALES Y PROCESOS SOCIALES

¿Qué voy a aprender y cómo?

¿Crees que es factible predecir con cierto nivel de certeza el clima para la próxima semana?, ¿conoces algún método que permita medir la probabilidad de que llueva el próximo jueves?, ¿podrías nombrar algún instrumento que posibilite medir el impacto económico o migratorio derivado de una sequía prolongada?

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.

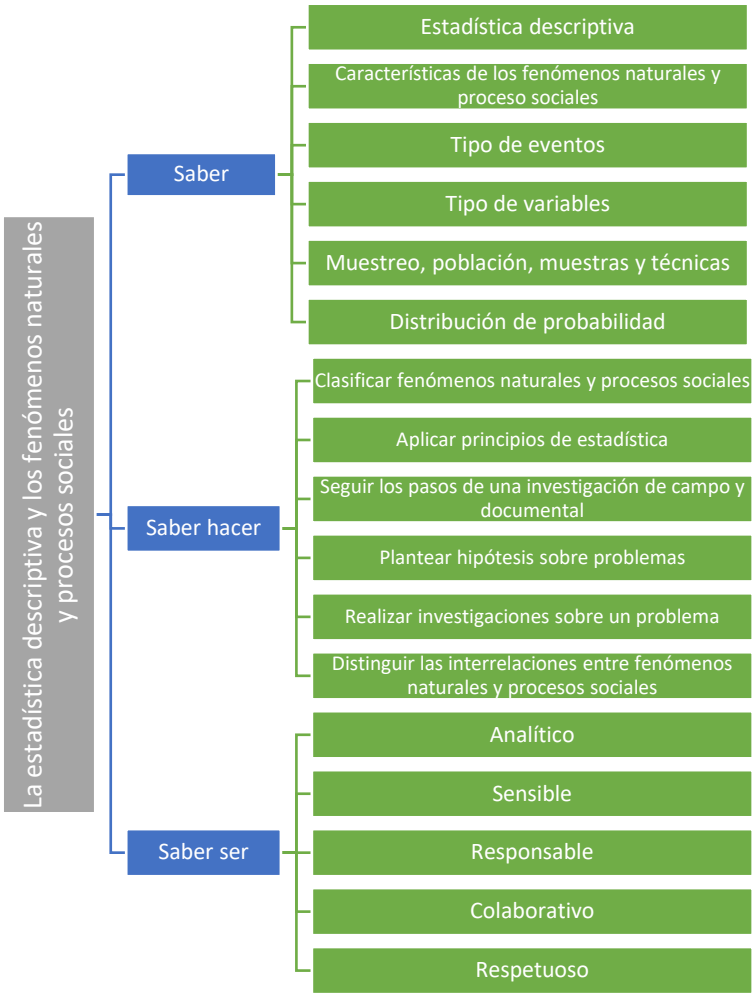
Es en la cotidianidad de lo natural y lo social en donde nos desenvolvemos y de la cual no somos ajenos. Su entendimiento conlleva no sólo a la búsqueda de información documental sino a la medición de su impacto a partir de la aplicación de métodos estadísticos y el cálculo de probabilidades. Su conocimiento profundo además de sensibilizarnos, permite nuestra participación de manera colaborativa ante la ocurrencia y el comportamiento de fenómenos naturales y procesos sociales.

¿Con qué propósito?

De manera concreta, el propósito de la unidad es que analices fenómenos naturales y procesos sociales de tu entorno mediante los conceptos básicos de la probabilidad y la estadística (**tipos de eventos, variables, muestreo, gráficas**) para reconocer los diferentes **tipos de distribución** de la información y con ello explicar su comportamiento en un contexto determinado.

¿Qué saberes trabajaré?

Para lograr el propósito formativo descrito es importante que adquieras y manejes los saberes relacionados a continuación:



El uso de la estadística es común en diversas ciencias y situaciones cotidianas pues permite obtener información detallada de todo tipo de variables sociales y naturales.
Es importante identificar los elementos de juicio que te permitan identificar y diferenciar los fenómenos.

Estadística: ciencia matemática que se sirve de los conjuntos de datos para obtener, a partir de ellos, conclusiones basadas en el cálculo de probabilidades.

Fenómenos naturales y procesos sociales:

Los **fenómenos naturales** son todos aquellos procesos constantes de movimientos o transformaciones que tienen lugar en la naturaleza.

Los **procesos sociales** son las distintas maneras en que la cultura y las organizaciones sociales cambian o son preservadas. Son cambios históricos consistentes dentro de una sociedad o una institución social, modos característicos de interacción social.





Tipos de fenómenos naturales

Según los procesos involucrados en su ocurrencia, los fenómenos naturales pueden ser clasificados de distintas formas:

Biológicos: son causados por organismos vivos, como plagas, epidemias, pandemias, marea roja, etc.

Físicos: son producto de fenómenos físicos, como movimientos de las mareas, evaporación del agua, erosión, arcoíris, aurora boreal.

Meteorológicos: son ocurridos en la atmósfera, como vientos, precipitaciones (lluvias, granizadas, nevadas), tormentas, tornados, huracanes, inundaciones, sequías, heladas, cambios repentinos de temperatura.

Hidrológicos: son aquellos que ocurren en el agua o las grandes masas de agua como, oleajes, tsunamis o inundaciones.

Geofísicos: son aquellos que se forman o surgen desde el interior del planeta o en su superficie: avalanchas, derrumbes, deslizamientos, terremotos, erupciones, hundimientos de tierra, etc.

Los fenómenos naturales son susceptibles de convertirse en desastres en relación con el grado de vulnerabilidad de la población que los experimenta.

Tipos de eventos: determinísticos y aleatorios

Algunos de los eventos asociados con cambios en la naturaleza son diarios u ocurren en periodos específicos; por ejemplo, eventos de la naturaleza como el día y la noche, las estaciones del año, se suscitan en periodos específicos, siendo completamente **predecibles** en cuanto a la hora, día, mes o año en que estarán presentes. Por otra parte, existen otro tipo de eventos asociados con los fenómenos naturales cuya característica de **ocurrencia es esporádica o no ocurre en periodos específicos**, lo que los convierte en **eventos completamente impredecibles o difíciles de precisar**; por ejemplo, los desastres naturales: tsunamis, maremotos, lluvias prolongadas, tornados, terremotos, etc.
Las observaciones con respecto a la ocurrencia de los eventos naturales y de los procesos sociales nos conducen a realizar una caracterización basada en la ocurrencia de sus eventos asociados.
Los procesos sociales mantienen una lógica parecida, donde se observa la ocurrencia de este tipo de eventos; por ejemplo, el ingreso de un trabajador asalariado es un **evento totalmente predecible**, la naturaleza de ocurrencia de éste lo hace totalmente pronosticable (siempre y cuando no se considere un periodo demasiado extenso donde seguramente por incrementos salariales o por ascenso de puesto el trabajador aumentaría

sus percepciones económicas) pues al tener el trabajador un salario fijo, se sabe con toda certeza cuánto será su salario mañana, pasado mañana o dentro de un mes. Por otra parte, el proceso social que considera como evento al número de inmigrantes y migrantes hacia o desde cierta región en un tiempo determinado, observó que por su ocurrencia los flujos de personas que migran hacia o desde determinada región no se pueden anticipar de manera exacta ni en un periodo determinado.
Las observaciones con respecto a la ocurrencia tanto de los eventos naturales como de los procesos sociales nos conduce a realizar una nueva caracterización de dichos fenómenos basada en la ocurrencia de sus eventos asociados. Podemos observar que la ocurrencia de estos eventos está caracterizada por la **certidumbre o incertidumbre** de su presencia; por un lado podemos obtener predicciones certeras y con ello tener la certidumbre de poder pronosticar su observación con exactitud para cualquier periodo determinado, mientras que existen otro tipo de eventos que por su ocurrencia incierta son más difíciles de pronosticar.

La certidumbre e incertidumbre que caracteriza la ocurrencia de estos eventos nos lleva a clasificarlos como **determinísticos** o **aleatorios**.

Fenómenos naturales y procesos sociales determinísticos y aleatorios



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



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

Una vez identificada la característica que permite diferenciar a los fenómenos abrimos paso a la discusión sobre la complejidad que representa tratar con la **incertidumbre** derivada del estudio de los eventos aleatorios. La incertidumbre presente en la ocurrencia de un fenómeno natural o un proceso social es una problemática a la que se enfrentan distintos investigadores de las áreas sociales y naturales.

La dificultad de eliminar la incertidumbre sobre la ocurrencia de eventos relacionados con fenómenos naturales y procesos sociales aleatorios hace importante contar con herramientas que otorguen los elementos para entender y pronosticar el comportamiento de un fenómeno natural o proceso social, a partir del estudio del comportamiento de la incertidumbre relacionada con un evento aleatorio. A este conjunto de herramientas, de métodos y técnicas cuyo objetivo es analizar y predecir dichos eventos, es a lo que a partir de este momento denominaremos **Estadística**.

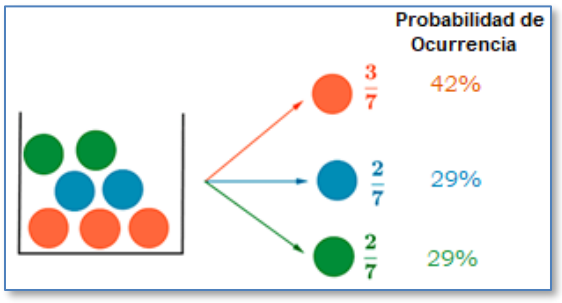
Estadística: ciencia referente a la recolección, análisis e interpretación de datos, que busca explicar condiciones regulares en fenómenos de tipo aleatorio.

Concepto de probabilidad

En cualquier evento aleatorio relacionado con un fenómeno natural o un proceso social siempre existe la incertidumbre pues no se sabe si ocurrirá o no. Por eso es importante

contar con una medida que permita cuantificar la oportunidad o posibilidad de ocurrencia de un evento aleatorio.

La **probabilidad** asigna a la oportunidad o posibilidad de ocurrencia de un evento aleatorio un valor numérico que se encuentra entre 0 y 1 (o entre 0 y 100 si hablamos en términos porcentuales).



- Si el valor numérico asociado a la ocurrencia de un evento aleatorio es 1, representa la certeza de que es totalmente seguro que éste ocurra, o que su probabilidad de ocurrencia sea del 100%.
- Un valor de 0 indica que es imposible su realización o equivale a un 0% de probabilidad de que ocurra.

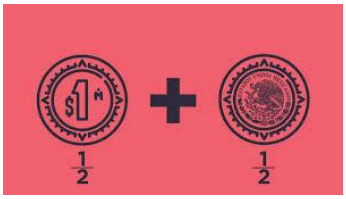
Asignación de la probabilidad a un evento aleatorio

Para asignar la probabilidad a un evento aleatorio se considera a ésta como la razón entre el número de casos (sucesos) favorables y el número total de casos (sucesos posibles). Esto es, si denotamos a "A" como un evento aleatorio, la probabilidad del evento aleatorio "A" a la que denotamos por P(A) se obtiene como esta razón, donde "k" es el número de casos favorables que caracterizan el evento y "n" es el número de casos posibles que caracterizan el fenómeno bajo estudio.

La técnica de asignación de la probabilidad a un evento aleatorio conocida como **enfoque clásico** es válida cuando no existen evidencias de que algunos eventos aleatorios asociados con el fenómeno bajo estudio puedan tener mayores oportunidades o posibilidades de ocurrencia sobre los demás.

Enfoque clásico: enfoque que determina la probabilidad de un evento con base en el supuesto de que todos los eventos asociados con la ocurrencia de un fenómeno tienen la misma probabilidad de ocurrir.

Cuando todos los eventos de un espacio muestral tienen la misma probabilidad de ocurrencia se denominan **eventos equiprobables**.



Ejemplo:

Calcular la probabilidad asociada con el evento que llueva una sola vez durante tres días consecutivos de los datos de proporcionados en la siguiente tabla que contienen la combinación de todos los eventos posibles:

Continuación…

Evento	Día 1	Día 2	Día 3
1	Llueve	Llueve	Llueve
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

Partiendo de un ***enfoque clásico***:

- El número de casos favorables que corresponden al evento de que llueva una sola vez durante tres días consecutivos es **3** → Los casos **5, 6 y 7**.
- El número total de casos son **8** como se muestra en la tabla.
- La probabilidad de este evento queda determinada por $3/8$, o equivalentemente se tiene un **37.5% de probabilidad que llueva una vez durante tres días consecutivos.**

Existen otros dos enfoques para asignación de la probabilidad: el **enfoque frecuentista** y el **enfoque subjetivo**.
Hay cierto tipo de fenómenos naturales y procesos sociales cuya ocurrencia puede tener un impacto directo en nuestro contexto, es así que una vez identificados seamos capaces de cuantificar su impacto hacia nuestro entorno, generando conciencia sobre ello. Si bien es cierto que la simple percepción posibilita que tengamos un aceptable nivel de discernimiento para intuir de alguna manera el grado de ocurrencia de un evento, es importante contar con una o varias medidas objetivas que además de cuantificar su impacto nos sirvan como instrumentos para validar o refutar nuestras percepciones.

Para medir el impacto de un fenómeno natural o un proceso social sobre el entorno es necesario estudiar la frecuencia con la que se presenta, se consideran como medidas objetivas del impacto de un evento a aquellas que involucren el cálculo de su frecuencia observada, estas medidas son: la **frecuencia absoluta** y **frecuencia relativa**.
Para definir estos términos debemos considerar tener un fenómeno aleatorio bajo estudio en el que contemos con todos los registros de ocurrencia de sus eventos asociados dentro de un tiempo específico o un espacio geográfico determinado.

Frecuencia absoluta es el número de veces que un evento es observado dentro de un periodo específico o en un espacio geográfico determinado.

Ejemplo:
Al observar la precipitación pluvial durante 30 días en cierta región observamos que sólo llovió 10 días, por lo tanto, la ***frecuencia absoluta*** de este fenómeno queda determinada por el número de días en los que se observó la precipitación pluvial, que para este ejemplo es de **10 días**.

Frecuencia relativa es el cociente entre la frecuencia absoluta del evento y el número total de eventos observados dentro de un periodo específico o un espacio geográfico determinado. En términos simbólicos, la *frecuencia relativa* de un evento se define como **n/N** (número de casos observados entre el número total de casos), donde **n** y **N** representan *el número de veces que el evento fue observado y el número total de eventos observados en un periodo específico de tiempo o en un espacio geográfico determinado.*

Ejemplo:
Al observar la precipitación pluvial durante 30 días en cierta región observamos que sólo llovió 10 días. Determina la frecuencia relativa.
Se considera la frecuencia absoluta de 10 días.

$$\text{Frecuencia relativa} = \frac{\#\text{ de veces que el evento fue observado}}{\#\text{ total de eventos observados}}$$

$$\text{Frecuencia relativa} = \frac{\text{Frecuencia absoluta}}{\#\text{ de días observados}}$$

$$\text{Frecuencia relativa} = \frac{10}{30} = 0.33 \quad \text{o} \quad 33.3\%$$

Ejemplo:
En cierta comunidad rural se inicia un programa para la prevención del contagio y transmisión de enfermedades respiratorias entre hombres y mujeres de la tercera edad de 70 o más años durante el periodo invernal. Se pretende informar a dicha población sobre cuáles son los cinco tipos de enfermedades respiratorias más frecuentes.
Para ello se cuenta en el hospital comunitario con un registro del número de individuos en este rango de edades a los que se les diagnosticó alguna de las cinco enfermedades respiratorias consideradas. En la tabla se muestra el registro.

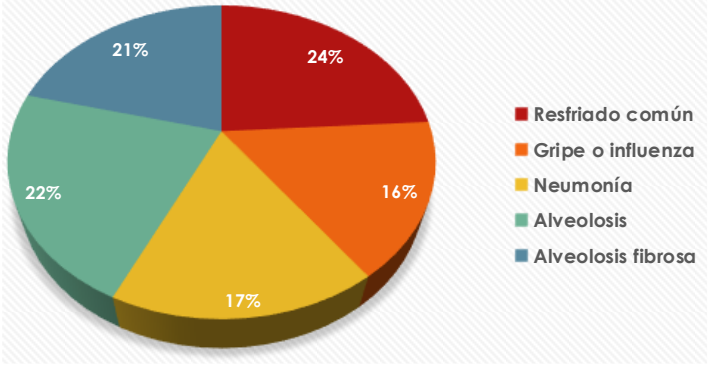
Enfermedad	Número de casos registrados (n)
Resfriado común	36
Gripe o influenza	24
Neumonía	26
Alveolitis	33
Alveolitis fibrosa	31
Número Total de Casos "N"	150

Continuación...

Para medir el impacto de estas enfermedades, se calculan sus frecuencias relativas, las cuales se muestran en la siguiente tabla junto a sus correspondientes frecuencias absolutas, en la siguiente tabla:

Enfermedad	Frecuencia absoluta	Frecuencia relativa
Resfriado común	36	=36/150= 0.24
Gripe o influenza	24	=24/150= 0.16
Neumonía	26	=26/150= 0.17
Alveolosis	33	=33/150= 0.22
Alveolosis fibrosa	31	=31/150= 0.21
Total	150	1

Estos datos se pueden representar en una gráfica de pastel como la siguiente:



Como ejemplo de cálculo de la frecuencia relativa, se muestra la manera en que se obtuvo el valor correspondiente a la frecuencia relativa asociada con la *prevalencia* del resfriado común. Para ello se consideró a N = 150 y n = 36, de donde se obtiene que la frecuencia relativa correspondiente es de 36/150 o del 24%. Dicha cantidad demuestra que durante un año el resfriado común fue la enfermedad respiratoria que más impacto tuvo entre los adultos mayores de 70 o más años. Tal demostración significa que es alta la probabilidad de que el fenómeno se repita.

Prevalencia: número total de casos de enfermos para un tipo específico de enfermedad en un momento y lugar particular.

Revisa más de este tema en:

Estructura de informes es un tema que podrás revisar también en el **Módulo 13** página 150.

ACTIVIDAD:

- ✓ Realiza la actividad 2 de la página 39 de la versión electrónica correspondiente a este módulo, además realiza el **informe** escrito. Corrobora tus respuestas en la pág. 167 del Apéndice 1.

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

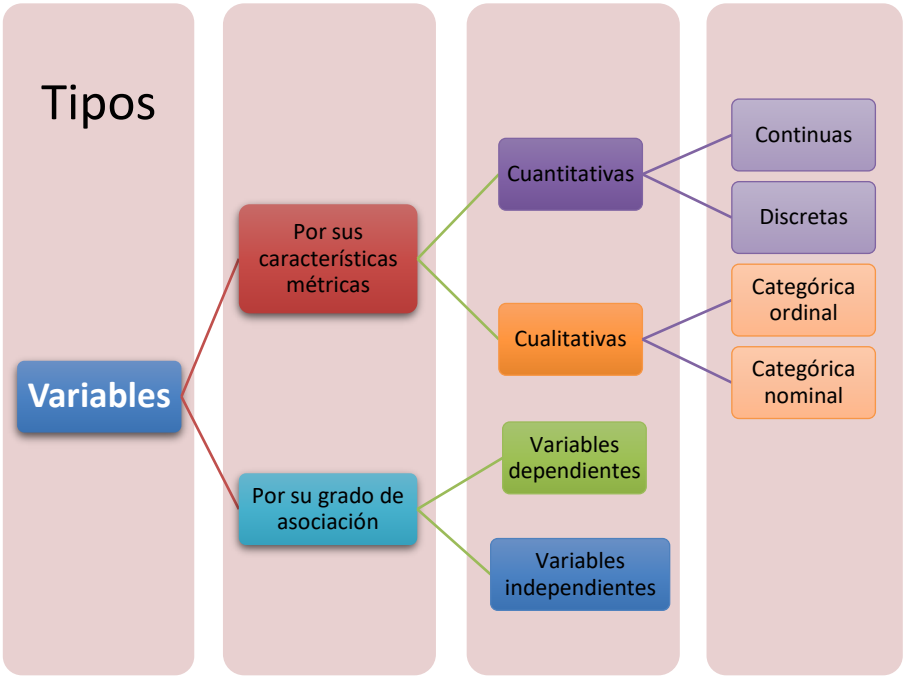
Los fenómenos naturales y los procesos sociales poseen características que por su naturaleza pueden ser medidas y estudiadas. Estas características se denominan **variables**.

Variable: símbolo que representa el valor de una cantidad cuyo posible valor observado no es fijo porque puede variar.

A partir de la naturaleza aleatoria de estos fenómenos es fácil observar que las características consideradas están sujetas a variaciones en el tiempo o en el espacio. La incertidumbre derivada de la aleatoriedad de estos fenómenos dificulta la obtención de medidas idénticas para cada una de las características, para distintos periodos y diferentes espacios.

Variable es cualquier característica cuyo valor puede cambiar al ser medida en distintos periodos del espacio o tiempo.

Ejemplo: Existen muchos tipos de variables como; la temperatura, que se mide en grados centígrados o Fahrenheit, la edad que se mide en años, estatura que se mide en centímetros/metros, o el peso que se mide en kilogramos o libras. También hay otras variables de tipo social que pueden medirse como la pobreza, desigualdad o calidad de vida que se miden de manera cualitativa.



Para estudiar esa variabilidad se recurre a los **métodos estadísticos,** herramientas y técnicas que también posibilitan controlar y predecir el comportamiento de fenómenos. El desarrollo de los métodos estadísticos dependen de las **características métricas** de las variables en consideración, para lo cual es básico distinguir los tipos de variables definidas en función de sus características métricas.

Métodos estadísticos: secuencia de procedimientos útiles para el manejo de datos cualitativos y cuantitativos de la investigación.
Características métricas: elementos que forman parte de objetos, seres vivos o fenómenos que pueden ser sujetos de medición.

Variables cuantitativas

En el estudio de los fenómenos naturales y procesos sociales aleatorios es muy común encontrarse con mediciones de diferente naturaleza, definidas como **variables continuas.**

Variables continuas cumplen la propiedad de que entre dos valores observados hay una infinidad de posibles valores observables. Otra característica de las variables continuas es que, en teoría, se pueden obtener una infinidad de valores distintos asociados con su medición. ***Por ejemplo*:** la temperatura, entre los 35 y 37 grados hay valores como 35.1. 35.2…36.9.

Siempre es posible identificar el rango de variación de todos sus valores posibles, es denominado **intervalo de variación** de todas las posibles mediciones asociadas con una variable continua. Por esto, a las variables continuas también se les llama **variables intervalares**.

Existen otro tipo de variables que no son de tipo continuo y que aparecen con mayor frecuencia en áreas sociales y humanísticas.

A partir de las observaciones de la naturaleza de las mediciones, se da la definición de la **variable discreta**.

Variable discreta es aquella que sólo puede tomar valores dentro de un conjunto numerable, generando con ello una separación entre sus valores observables sucesivos. Estas variables sólo pueden adoptar valores numéricos enteros, de tal modo que entre dos valores consecutivos hay por lo menos un valor no observable, dicho de otro modo, los valores intermedios que pudieran ser considerados entre dos valores enteros consecutivos, carecen de sentido. ***Por ejemplo:*** el número de personas obesas en una comunidad, el número de contribuyentes morosos que no han pagado sus impuestos el número de personas que padecen cierta enfermedad y que están internadas en un hospital comunitario.

A diferencia de las variables continuas, las discretas pueden tomar sólo un número finito de valores.

Variables cualitativas

Existen otro tipo de mediciones sobre las cuales no se puede asignar de manera natural, un valor numérico, a partir del cual se obtendrá como resultado una característica y no una medida numérica. A partir de esas consideraciones, es posible definir la **variable categórica** como aquella cuyos **niveles de respuesta** son de tipo cualitativo pues están relacionados con características de personas u objetos como color, raza, credo, filiación política, nivel educativo, etc.

Medición cualitativa: medición de naturaleza no numérica obtenida a partir de la observación directa de alguna característica específica de un objeto, ser vivo o fenómeno y que busca describir algunas de sus cualidades.

Variables categóricas nominales y ordinales

Es necesario asignar un valor numérico a cada uno de los niveles de respuesta de las variables categóricas. La asignación de valores numéricos a los niveles de respuesta de una variable categórica tomará sentido cuando, por su naturaleza, las categorías de su respuesta guarden de manera natural una jerarquía. En este caso la **variable categórica** se denomina **ordinal**. Por ejemplo**:** primero, segundo, tercero, etc.

Hay ocasiones en las que la característica de la respuesta no es de tipo ordinal, por lo tanto, una **variable categórica** se define como **nominal** si sus niveles de respuestas no guardan ninguna relación de jerarquías ente ellos. Por ejemplo: estado civil: soltero ó casado, el sexo: hombre o mujer.

Clasificación de las variables por su grado de asociación

Los vínculos entre los fenómenos naturales y los procesos sociales abren paso al planteamiento de **hipótesis** sobre las posibles relaciones de causa y efecto que pudieran existir entre las variables asociadas con fenómenos naturales y variables asociadas con procesos sociales, dando lugar a la identificación de variables que se puedan definir como dependientes o independientes.

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

CASO PRÁCTICO:
Un Investigador está interesado en conocer la relación que existe entre el consumo de tabaco y ansiedad, su hipótesis es que pese a lo que creen los fumadores el consumo de tabaco aumenta la ansiedad en lugar de disminuirla.

La dependencia o independencia ente dos variables está basada en la relación de causa efecto entre ambas; esto es, si el valor observado de una de estas tiene efecto directo sobre el valor observado de la otra o viceversa.

Toda hipótesis planteada debe ser sujeta a comprobación, existen herramientas que permiten validad este tipo de hipótesis.

Una **variable independiente** se define como la característica que se supone es la causa del fenómeno estudiado. Mientras que, **la variable dependiente** lo hace como la característica cuyos valores están afectados por los valores que pueda tomar la variable independiente.

Ejemplo: El calentamiento global es una variable dependiente del efecto invernadero.
El sobrepeso es una variable dependiente de los malos hábitos alimenticios y la falta de ejercicio.



Revisa más de este tema en:
Variables dependientes e independientes es un tema que podrás revisar también en el **Módulo 14** página 25 y **Módulo 20 página 157**.



Revisa más de este tema en:
Estructura de informes es un tema que podrás revisar también en el **Módulo 13** página 150.

ACTIVIDAD:

- ✓ Realiza la actividad 3 **sobre tipos de variables** de la página 47 de la versión electrónica correspondiente a este módulo, además de realizar el **informe.** Corrobora tus respuestas en el Apéndice 1.

Muestreo: población, muestra y técnicas

Técnicas de muestro y su importancia en el análisis de la información



Revisa más de este tema en:
Muestreo es un tema que podrás revisar también en el **Módulo 20** página 148.

Un instrumento de medición carece de utilidad si no se tiene posibilidad de utilizarse para obtener información, se requieren las fuentes de información de donde obtener sus valores o mediciones.

Las **fuentes de información** son el insumo sin el cual el estudio de los fenómenos naturales y sociales no se podrá llegar a buen término. Una vez definida una variable de interés, es importante cerciorarnos si tenemos la posibilidad de acceder a la información que nos permita obtener las mediciones relacionadas con ella.

Una vez identificadas las fuentes de información, sigue la **recolección de mediciones o datos** que permita estudiar el comportamiento del fenómeno. En este punto, la calidad de la información recolectada juega un papel muy importante, pues de ésta dependerá la confiabilidad de los resultados que obtengamos del fenómeno a estudiar.

Censos y muestras

Una de las fuentes más confiables para obtener información es el **censo**, el cual arroja información valiosa para realizar mediciones sobre las variables de interés.

Censo: recuento de una población (entendiendo por población al conjunto de todos los elementos o individuos que son objeto del estudio) con la finalidad de conocer cuestiones inherentes a ella.

Existen alternativas metodológicas que permiten al investigador obtener información confiable de una variable sin tener que realizar el censo, que son los **métodos de muestreo**. Estas técnicas se basan en la recolección de información de una fracción de la población de estudio.

Métodos de muestreo: conjunto de técnicas diseñadas para la obtención de una muestra.

Concepto de muestra



Se define como muestra al subconjunto extraído de la población mediante la aplicación de algún método de muestreo, que permite obtener conclusiones validas que se pueden generalizar hacia la población.

Muestra: colección de mediciones seleccionadas de una población de interés.



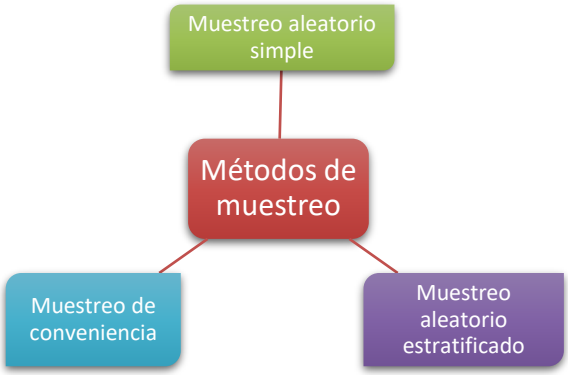
CASO PRÁCTICO:
Un investigador quiere conocer los hábitos de los niños con sobrepeso que acuden a un centro de salud, en total ha registrado que de manera mensual se atienden a 100 niños que tienen esta característica. El investigador ha decidido entrevistar a 60 niños por considerar una muestra representativa de los hábitos alimenticios que pudieran tener todos los niños, incluidos los que no se entrevistarán.
Población= 100 niños, Muestra=60 niños (60 muestras)

Representatividad de una muestra

La naturaleza de los métodos de muestreo permite encontrar generalizaciones a partir de una muestra de la población correspondiente siempre y cuando la muestra sea representativa de la población.
Una muestra mal tomada puede introducir un sesgo en los resultados, el **sesgo** en una muestra se define como la tendencia a diferir en sus resultados de la correspondiente población.

Métodos de muestreo





El **muestreo aleatorio simple** es un método que asegura que cada muestra posible del tamaño deseado tiene la misma oportunidad de ser elegida. Por ejemplo: una tómbola.



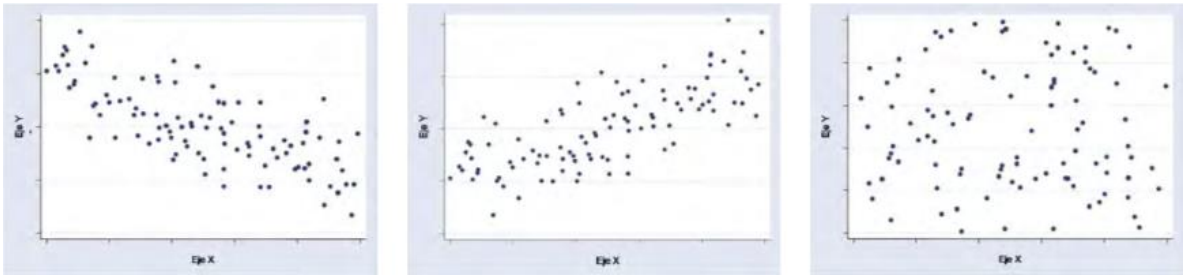
Muestrea aleatoria simple: Si representamos por “n” al número de objetos o individuos en la muestra, entonces una muestra aleatoria simple de tamaño “n” se define como aquella que se selecciona a partir de un procedimiento que asegura que, todas las muestras posibles diferentes del tamaño deseado tienen la misma probabilidad de ser seleccionadas. La definición de muestra aleatoria simple implica que cada uno de los miembros de la población tiene la misma probabilidad de ser seleccionado en la muestra.

Selección de una muestra aleatoria simple

Se pueden utilizar diferentes métodos para seleccionar una muestra aleatoria simple. Un método muy utilizado para la obtención de una muestra aleatoria simple es hacer primero una lista, cada elemento de ella puede ser identificado con un número y utilizar una tabla de número aleatorios o un generador de números aleatorios para seleccionar la muestra.

Un generador de números aleatorios es un procedimiento que produce una secuencia de números que satisface las propiedades asociadas con la noción de aleatoriedad.

La técnica del **diagrama de dispersión** utiliza un conjunto de registros apareados de la forma (X, Y), con X como variable independiente e 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. La forma que toma la nube de puntos resultante es la que permite identificar si existe alguna tendencia que indique algún grado de asociación entre las variables.

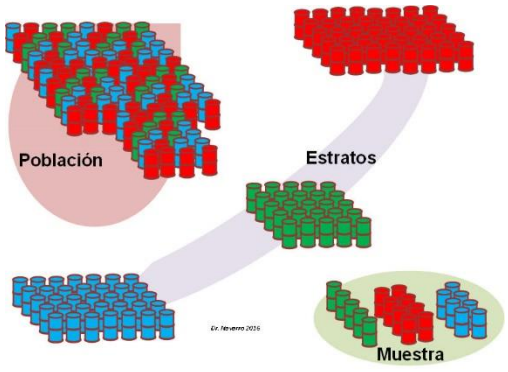


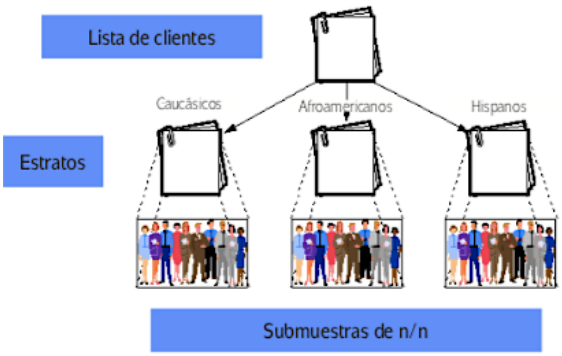
La primera de estas gráficas muestra que no hay relación entre las variables, mientras que la segunda y la tercera muestran la asociación entre ellas, ya que en el primer caso el valor de la variable Y aumenta conforme lo hace el valor de la variable X, mientras que en el segundo este disminuye.

A pesar de contar con un número menor de observaciones como resultado de la selección de una muestra aleatoria simple, es posible identificar que los datos muestreados representan adecuadamente la tendencia observada por los datos totales. Con ello concluye que es posible generalizar el comportamiento de la población con el simple hecho de haber observado el comportamiento de la muestra.

Muestreo aleatorio estratificado

Cuando toda la población puede ser dividida en una serie de subgrupos que no se superponen, el método más adecuado para el análisis es el conocido como **muestreo aleatorio estratificado**. En este se seleccionan diferentes muestras aleatorias de cada subgrupo, de manera independiente.





Cuando la población se divide de esta manera, los subgrupos se llaman **estratos** y cada subgrupo individual se llama **estrato**. El muestreo estratificado se puede utilizar en lugar del muestreo aleatorio simple siempre y cuando sea importante obtener información sobre las características de los estratos individuales, así como de toda la población.

A menudo nos permite hacer **inferencias** más concretas de una población que un muestreo aleatorio simple.

Inferencia: proceso que puede ligar dos o más proposiciones.

CASO PRÁCTICO:
El investigador que tomó la muestra de 60 niños con obesidad, obtuvo a través de sus encuestas que 50 de estos 60 niños consumen bebidas azucaradas de 2 a 3 veces al día, y realizan poca actividad física; también encuentra que la variedad de alimentos que se ingiere no es constante en todos los entrevistados; por lo que **infiere** que la obesidad infantil puede estar mayormente asociada al consumo de alimentos azucarados y falta de actividad física que a otros hábitos.

Muestreo de conveniencia

Una forma común de muestreo por conveniencia es la conocida muestra de respuesta voluntaria, que se basa en la información obtenida de personas que se ofrecen como voluntarios para formar parte de la muestra, a menudo respondiendo un anuncio, llamando a un número telefónico o al iniciar sesión en un sitio de internet.
Es muy poco probable que los individuos que participan en tales encuestas voluntarias sean representativos de una población más grande de interés.



Revisa más de este tema en:
Presentaciones en Microsoft Power Point es un tema que podrás revisar también en el **Módulo 6** página 61.

ACTIVIDAD:
- ✓ Realiza la actividad 4 que integra conceptos del tema **censos y muestras** de la página 58 de la versión electrónica correspondiente a este módulo, corrobora tus respuestas en las págs., 170 y 171 del Apéndice 1. Además, realiza una presentación en computadora.

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



Conjeturar que se cuenta con la posibilidad de medir cualquier característica de interés a cada uno de los elementos de la población, implica que se pueden obtener mediciones poblacionales de variables numéricas (cuantitativas) o categóricas (cualitativas) y con ello conocer la manera cómo se distribuyen dichas medidas dentro de la población.

Distribuciones de frecuencias para variables categóricas

Para realizar una investigación lo primero que se debe hacer es recurrir a la descripción de cómo se distribuyen las categorías (**distribución de frecuencias**) de tal manera que el impacto sea medido a partir de la frecuencia relativa que ambas respuestas tienen entre el número total de la población.

Distribución de frecuencias: instrumento tabular que presenta los valores de los datos y la frecuencia con que se presentan facilitando con ello se comprensión e interpretación.

La **frecuencia relativa** se calcula dividiendo la frecuencia absoluta entre el número total de observaciones en el conjunto de los datos. La distribución de frecuencias puede resumirse en una gráfica de barras cuya frecuencia relativa (proporción) queda representada por la altura de cada rectángulo correspondiente a cada valor de la variable.

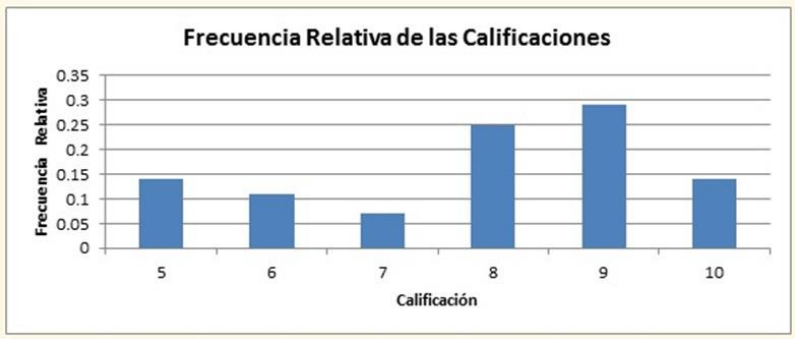


Figura 2

Frecuencia relativa: cociente entre la frecuencia absoluta y el tamaño de la muestra.

Histogramas de frecuencias

Histograma: gráfica de barras que muestra el número de datos observados o su frecuencia.

Esta técnica gráfica resulta de gran utilidad para la representación de la distribución de frecuencias para las variables categóricas. El **histograma de frecuencias** adopta ciertas particularidades en su elaboración dependiendo de la característica de la variable bajo estudio.

Histograma de frecuencias: gráfica de barras que se utiliza para representar la frecuencia de los datos.

Histogramas para variables numéricas discretas

Un investigador social está interesado en estudiar la distribución de frecuencias relacionada con el número de hijos que cada mujer procreó en cierta comunidad rural. Para ello toma una muestra aleatoria de 30 mujeres y registra para cada una el número de hijos vivos que ha procreado. En la siguiente tabla se muestran los registros obtenidos para las 30 mujeres.

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

Como paso previo a la elaboración de un histograma es necesario generar una tabla de distribución de frecuencias. Con este tipo de variables es posible crear distintos tipos de tablas; uno de ellos es listar los valores posibles junto con sus correspondientes frecuencias absoluta y relativa.

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

Otra forma de construir una tabla de frecuencias es elaborar una distribución de frecuencias más compacta agrupando los datos y algunos de los posibles valores en intervalos. Un agrupamiento similar de los otros valores da como resultado la siguiente tabla de distribución de frecuencias relativas.

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

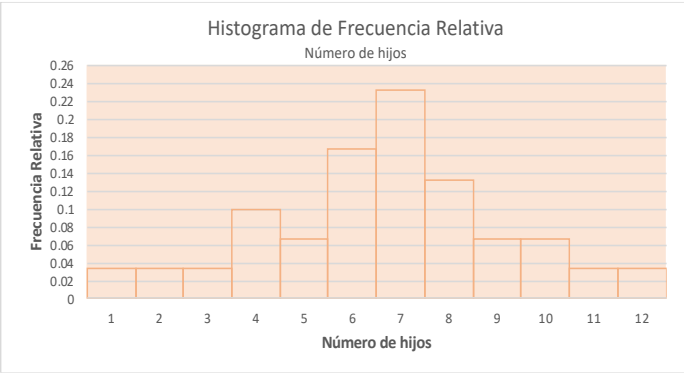
Elaborada la tabla de distribución de frecuencias se desarrollan los siguientes pasos para la elaboración de un histograma:

1. Se dibuja una escala horizontal y se marcan los valores posibles de la variable.

2. Se establece una escala vertical con la marca de la frecuencia o frecuencia relativa.

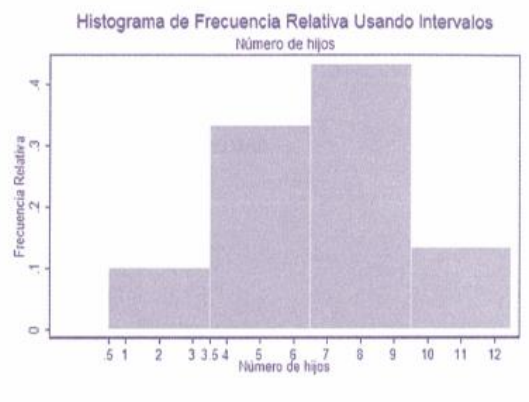
3. Por encima de cada posible valor se dibuja un rectángulo centrado en el valor mismo, de modo que el rectángulo para 1 esté centrado en 1, el del 5 lo esté en 5 y así respectivamente. La altura de cada rectángulo está determinada por la respectiva frecuencia o frecuencia relativa. A menudo, los posibles valores son números enteros consecutivos, en cuyo caso el ancho de la base de cada rectángulo es 1.

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





Se puede construir de manera similar un histograma, con base en la segunda tabla de frecuencias. En este caso el rectángulo representa la frecuencia absoluta o frecuencia relativa de cada intervalo. Para el intervalo 1-3 el rectángulo se extiende de 0.5 a 3.5 por lo que no hay diferencias entre los rectángulos del histograma.



Histograma para una variable continua

Cuando la característica a medir es de tipo continua, el proceso de construcción del histograma es más laborioso pues ciertas medidas son tan distintas que no es posible asignar la misma medida a un grupo de individuos de una población. La salida de este problema es definir la tabla de frecuencias con nuestras propias categorías, cada observación puede clasificarse en alguno de estos intervalos, denominados intervalos de clase.

Las observaciones que caen en el límite de clases se pueden marcar con el símbolo < (menor que), por ejemplo 40 < 50.
Para construir un histograma de frecuencias se deben seguir los pasos que se enumeran a continuación:

1. Marcar los límites de los intervalos de clase en un eje horizontal.

2. Utilizar la frecuencia o frecuencia relativa como escala del eje vertical.

3. Dibujar un rectángulo para cada clase directamente por encima del intervalo correspondiente (de modo que los bordes se encuentren en los límites de clase del intervalo). La altura de cada rectángulo es la frecuencia o la frecuencia relativa del intervalo de clase correspondiente.

No hay reglas para construir el número de intervalos de clase o la longitud de los intervalos, se pueden utilizar pocos intervalos relativamente amplios para agrupar demasiados datos en unas cuantas categorías o un gran número de intervalos relativamente estrechos que pueden diseminar los datos de más, por lo que habría intervalos conteniendo muy pocas observaciones. Ninguna de estas dos estrategias de agrupación dará información de cómo

los valores se distribuyen en todo el rango de medición, y ciertas características podrían perderse. En general, con una pequeña cantidad de datos, tal vez entre 5 y 10 intervalos de clase deberían ser utilizados. Cuando existe una gran cantidad de datos, una distribución basada en 15 a 20 (o más) intervalos de clase debería ser usada. La cantidad

$$\sqrt{número\ de\ observaciones}\ ,$$

es frecuentemente utilizada como un estimador del número apropiado de intervalos de clase: 5 intervalos para 25 observaciones, 10 intervalos cuando el número de observaciones es 100, y así sucesivamente.

El histograma de frecuencias es una herramienta gráfica de gran utilidad porque permite percibir de manera visual, a partir de la frecuencia relativa, el impacto que tiene la presencia de cada variable dentro de una población.

EJEMPLO:

El gobierno de un estado busca planear la distribución de recursos que en materia de educación se destinarán a cada uno de los 50 municipios que conforman su territorio, en el año venidero. Para hacerlo, el gobierno busca el porcentaje de estudiantes que están inscritos en las escuelas públicas, en cada municipio. Los datos recabados se muestran en la siguiente 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					

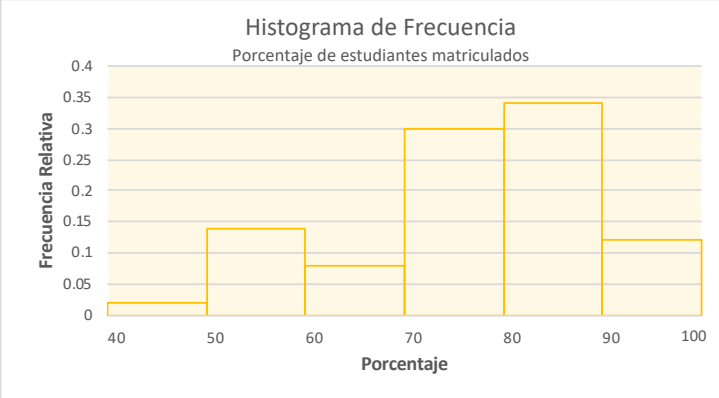
La observación más pequeña es 43 y las más grande 96. Es razonable comenzar el primer intervalo de clase en 40 y tomar cada intervalo de ancho 10. Esto genera intervalos de clase de 40 a <50, 50 a <60, 60 a<70, 70 a<80, 80 a<90 y 90 a<100.

La siguiente tabla despliega una tabla de distribución de frecuencias resultante junto con las frecuencias relativas.

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

Continuación:

Con base en la información de la tabla anterior se construye un histograma de frecuencias, el histograma resultante es el siguiente:



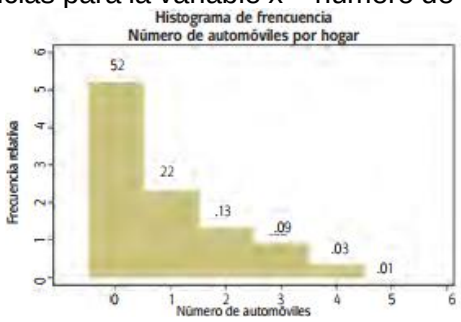
Distribuciones de frecuencias para variables numéricas discretas.

Una forma de describir la distribución de los valores de "x" para esta población es la construcción de un histograma, la escala de medición se divide en intervalos de clase. Cada valor del conjunto de datos es clasificado en uno de los intervalos.

Ejemplo:

En un estudio socioeconómico se quiere conocer la posibilidad que tienen los habitantes de cierta comunidad de contar con automóvil particular. Para tal efecto se considera la variable x = número de automóviles por hogar, los valores posibles para x son 0, 1, 2, 3, 4 y 5 (en esta población no existen hogares con más de cinco vehículos).

Para resumir la distribución de frecuencias para esta variable utilizaremos un **histograma de frecuencias relativas**. Para ello supóngase que se cuenta con registros censales del número total de vehículos para cada hogar de la localidad, por lo que se tiene la posibilidad de calcular la frecuencia relativa asociada con cada una de las respuestas de la variable "x". Esta frecuencia relativa es representada en el siguiente histograma de frecuencias, el cual representa la distribución de frecuencias para la variable x = número de automóviles por hogar.



Continuación:
A partir de esta gráfica se puede observar que el valor más común de la población es $x = 0$. En términos probabilísticos la probabilidad asociada con $x = 0$ es $P(x=0)=0.52$ esto es, el 52% de los hogares no cuentan con automóvil propio. Mientras que a partir de la misma gráfica se puede observar que sólo el 1% de los hogares tiene $x = 5$ automóviles de su propiedad; es decir, este valor tiene una probabilidad asociada de $P(x=5)=0.01$.

Distribuciones de frecuencias para variables numéricas continuas

Supongamos que se quiere estudiar el peso al nacer (en kilogramos) para todos los bebés nacidos en el año 2010 en cierta localidad. La variable x = peso de recién nacido para la población de todos los bebés recién nacidos en esta localidad es un ejemplo de una variable continua. Una forma de describir la distribución de los valores de "x" para esta población es la construcción de un histograma, la escala de medición (en este caso, es el rango que incluye a todos los posibles pesos al momento del nacimiento) se divide en intervalos de clase. Cada valor del conjunto de datos (en este caso, cada peso al nacer) es clasificado en uno de los intervalos. La gráfica correspondiente a este histograma está representada por la siguiente figura.



Este histograma representa la distribución del peso de los recién nacidos en esta localidad, y puede ser visto como la distribución de población de la variable x=peso de un recién nacido. La probabilidad de observar un valor de la variable "x" está asociada con el área de su rectángulo correspondiente.

Distribuciones de probabilidad continua

La distribución de probabilidad continua conserva las características importantes de la población y puede servir como modelo para la distribución de frecuencias de los valores de la población y sustituir el cálculo de las probabilidades de eventos, debido a que el área bajo de curva y por encima de cualquier intervalo particular se puede interpretar como la probabilidad de observar un valor en ese intervalo cuando un individuo es seleccionado al azar de la población.



Una **distribución de probabilidad continua** es una curva suave, la cual se denomina **curva de densidad**, que sirve como modelo para la distribución de frecuencias de la población para una variable continua y que cuenta con las siguientes propiedades:

1. El área total bajo la curva es igual a **1**.
2. 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.

Distribución de probabilidad continua: distribución de probabilidad asociada con una variable de tipo continuo.

Figura 1

Distribución del peso al nacer. Municipio de Río de Janeiro, 2000.

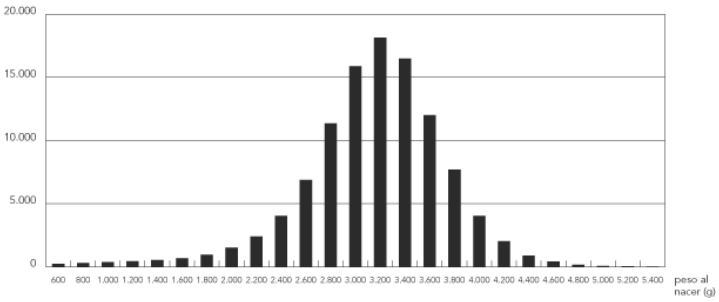
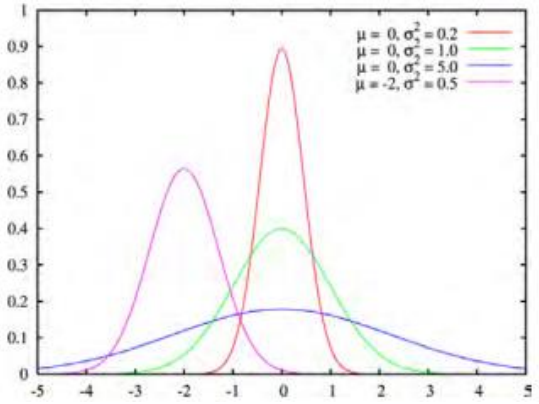


Figura 3

Distribuciones de probabilidades teóricas

CASO PRÁCTICO:
Observa la figura 3 se ha formado a partir del pesaje de las personas que han nacido en Río de Janeiro, si eligiéramos una persona que nace en Río de Janeiro antes de pesarla sabríamos que la mayor probabilidad de peso se encuentra entre 2.8 kgs., y 3.6 kgs., con mayor posibilidad de que su peso sea de 3.2 kgs.

Una distribución teórica que formaliza el comportamiento continuo es la **distribución de probabilidad** normal ya que en ella la curva suave tiene forma **de campana y es simétrica**. Las distribuciones normales también son referidas como curvas normales que se caracterizan al representarse en gráficas.



Distribución de probabilidad: función que asigna cada suceso definido sobre una variable la probabilidad de que dicho suceso ocurra.

Cuando hay necesidad de medir alguna variable de tipo continuo es recomendable organizar las mediciones en un histograma de frecuencias, ya que este brinda evidencia visual de un comportamiento de tipo normal. Esto es verificar si el histograma que hemos elaborado muestra tendencia a ser simétrico y tener forma de campana.

El uso de la distribución normal es frecuente para el estudio de fenómenos reales, y se utiliza como herramienta para el cálculo de probabilidades de eventos naturales y sociales asociados con variables que siguen una distribución de **probabilidad normal**.

Distribución de probabilidad normal: distribución de probabilidad continua cuya gráfica se asemeja a una campana.

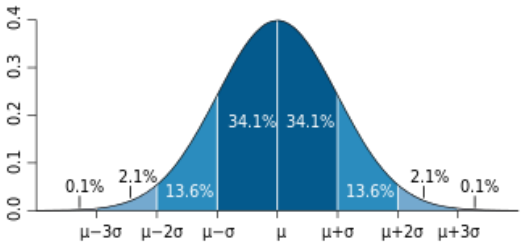


Figura 4

Distribución binomial

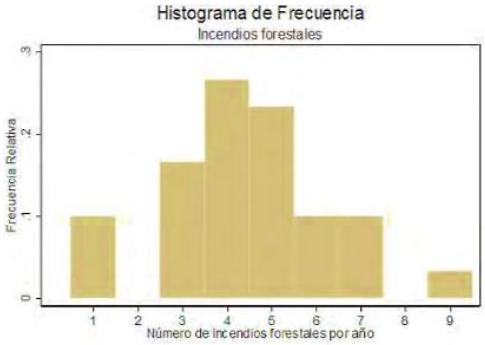
Un **experimento binomial** es aquel que posee las cuatro propiedades siguientes:

1. El experimento consiste de "n" intentos idénticos.
2. Cada intento da lugar a exactamente a dos resultados, llamados éxito y fracaso (pueden ser medidos como X o Y, respectivamente),
3. Los "n" intentos son independientes.
4. La probabilidad "p" de un éxito permanece constante de un intento a otro.

Por lo tanto, si "x" es una variable discreta que toma solo los valores 0 y 1 y el número de éxitos (x=1) es obtenido a partir de la realización de un **experimento binomial**, entonces a la distribución de probabilidad para el número de éxitos se le denomina **distribución binomial**.

CASO PRÁCTICO:

En un hospital comunitario se afirma que la presencia de diabetes es de 4 de cada 10 (40%), para verificarlo se toma una muestra de 6 personas (las cuales representarán al 100% de la población). Durante 15 días **(n intentos idénticos)** evalúa a 6 personas diferentes **(intentos independientes)** respecto a la presencia **(éxito=1)** ó ausencia **(fracaso=0)** de diabetes. La distribución de frecuencias para los pacientes que dieron positivo fue la siguiente:



Para la mayoría de las muestras tomadas el caso más frecuente fue 3, con lo que el investigador concluye que es el 50% (3 de cada 6) y no el 40%la presencia de diabetes en esta comunidad.

Experimento binomial: experimento que se caracteriza por la realización de un número finito de ensayos considerados independientes, para cada uno de los cuales sólo se tienen dos posibles resultados.
Distribución binomial: distribución de probabilidad asociada con la realización de un experimento binomial.

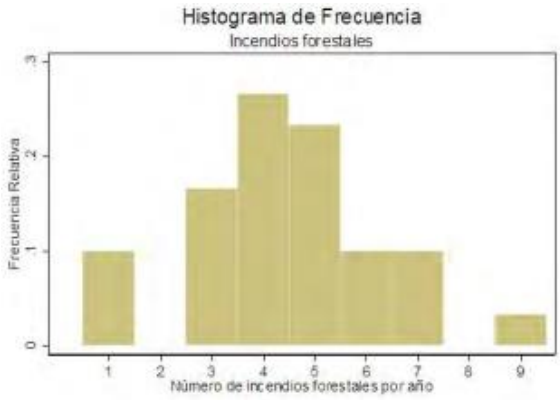
Distribución de Poisson

El experimento de Poisson hace referencia a procesos relacionados con el espacio o el tiempo y satisface las siguientes condiciones:

1. El número promedio de veces que ocurre un evento por unidad de tiempo o de espacio es constante.
2. La probabilidad de más de un seceso en una unidad de tiempo o espacio es muy pequeña.
3. El número de acontecimientos en intervalo ajenos de tiempo o de espacio son independientes unos de otros.

Si la variable "x" representa el número de sucesos resultantes de un experimento de Poisson se dice que esta variable sigue una distribución de Poisson.

CASO PRÁCTICO:
Un investigador alarmado por el incremento de incendios en los último 30 años (por decir, 1986 al 2016) debido al calentamiento global y la deforestación, decide comprar los registros anuales del número de incendios ocurridos en los 30 años previos al año que sospecha comenzó el incremento en el número de incendios (por decir, 1956 al 1986, periodo 1).



- Revisa los registros de número de incendios anules ocurridos durante los 30 años a partir del año en que sospecha comenzó el incremento (por decir, 1986 al 2016, periodo 2).



- Define la variable "x" como número total de incendios observados en la región en un año.
- Elabora un histograma de esta variable para los dos periodos.
- El investigador observa que en ambos periodos hubo incremento de incendios y la cifra más alta fue de 10 ó más por año (periodo 2).
- Concluye que el calentamiento global y la deforestación influyeron en el aumento de los incendios en los últimos 30 años (periodo 2).
  Para ambos periodos observados, 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.

ACTIVIDAD:

- ✓ Realiza la actividad **"Evalúa tu aprendizaje"** de las páginas 80 a 83 del libro correspondiente a este Módulo.
- ✓ Realiza la actividad 5 sobre distribuciones **normal, binominal o Poisson** de la página 74 de la versión electrónica correspondiente a este módulo, corrobora tus respuestas en el Apéndice 1.

UNIDAD 2. TRATAMIENTO ESTADÍSTICO DE LA INFORMACIÓN DE FENÓMENOS NATURALES Y PROCESOS SOCIALES

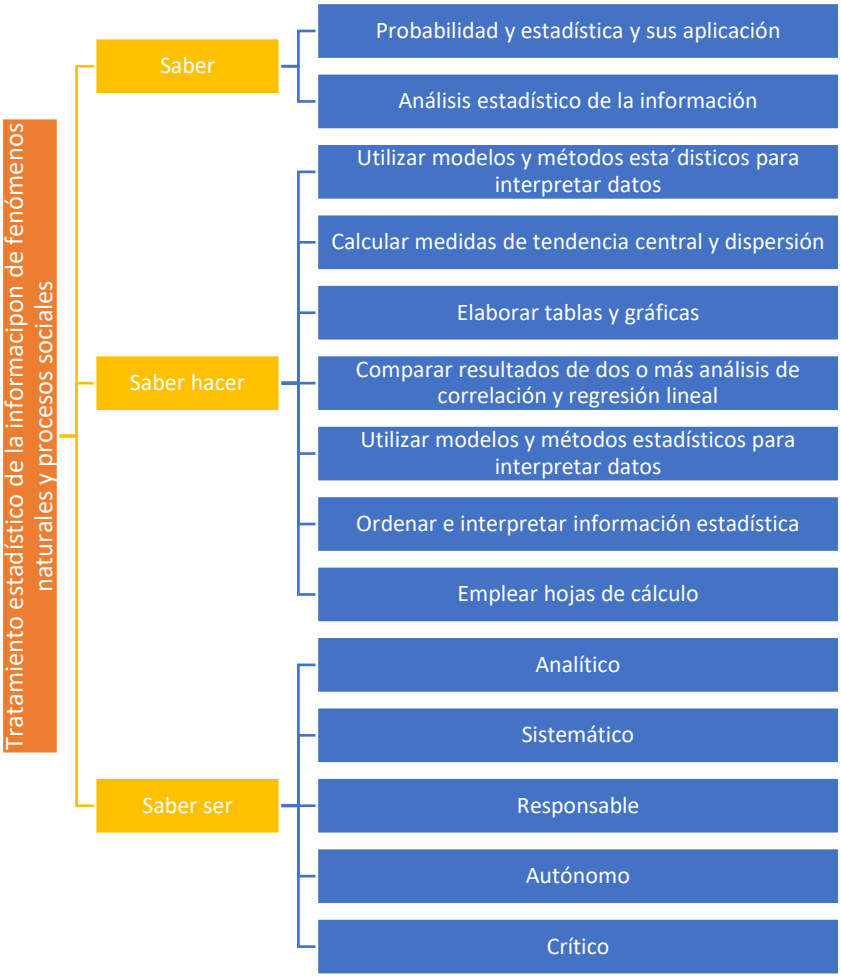
¿Qué voy a aprender y cómo?

¿Sabías que la estadística se puede aplicar a la literatura y la música?, ¿alguna vez habías oído que también puede ser útil en la pintura? Aunque algunas personas lo consideren poco creíble, está comprobado que la estadística tiene un papel fundamental en el avance del conocimiento en muchos ámbitos, desde la investigación sobre medicamentos, salud pública, mercados, control de calidad y hasta las artes. La aplicación de la estadística, como verás en esta unidad, en campos diversos de la ciencia y la vida posibilita tomar decisiones de manera individual y colectiva.

¿Con qué propósito?

El objetivo de esta unidad es el desarrollo de modelos de estadística descriptiva que permite analizar, ordenar, interpretar y explicar la información obtenida a partir de los fenómenos naturales y los procesos sociales para predecir y entender estos hechos en su contexto. En este sentido, el desarrollo del texto continua con el espíritu establecido desde la primera unidad de aplicar tus aprendizajes en cualquier entorno específico, permitiendo así la contextualización dentro de todo ámbito, ofreciendo de esta manera un criterio para la toma de decisiones individuales, de modo que puedas incidir en tu comunidad, región y país.

¿Qué saberes trabajaré?



Migración: movimiento poblacional que se realiza del campo a la ciudad y viceversa, de una población a otra o de un país a otro, provocando por cataclismos, guerras, violentos cambios sociales o persecuciones, o bien motivado por razones de tipo económico o cualquier otra que induzcan a los habitantes de un país o región a mudar su residencia.
Inmigración: es la llegada de personas a un país o región procedentes de otro país o región para residir en él.
Tasa de inmigración neta anual: personas que entran a un país o una región durante un año.
Movilidad social: movimientos o desplazamientos que efectúan los individuos, las familias o los grupos dentro de un determinado sistema socioeconómico.
Población económica activa (PEA): conjunto de personas de más de 12 años que desempeñan una ocupación.
Índice de pobreza humana: indicados que mide el nivel de vida en un país o en una región.
Producto Interno Bruto (PIB): valor de todos los bienes y servicios finales producidos dentro de una nación o región en un año determinado.

Revisa más de este tema en:
Población Económica Activa es un tema que podrás revisar también en el **Módulo 20** página 106.

PIB podrás revisar también en el **Módulo 10** página 101.

Revisa más de este tema en:
Migración es un tema que podrás revisar también en el **Módulo 4** página 55, **Módulo 10** página 71, **Módulo 11** página 283, **Módulo 14** página 69 y **Módulo 20** página 104.

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

La **estadística descriptiva** es la rama de las matemáticas que recolecta, presenta y caracteriza un conjunto de datos (por ejemplo, edad de una población, altura de los estudiantes de una escuela, etc.) con el fin de describir apropiadamente las diversas características de ese conjunto.
La **Estadística inferencial,** es una parte de la estadística que comprende los métodos y procedimientos que por medio de la inducción determina propiedades de una población estadística, a partir de una pequeña parte de la misma.

Con el estudio de esta sección se podrá responder a interrogantes que incluyan la **variabilidad numérica** al plantear medidas numéricas que describan con mayor precisión el valor numérico representativo de la información. También se podrán generar medidas "resumen" que cuantifiquen el grado de **dispersión** en un conjunto de datos

Variabilidad numérica: término utilizado para indicar la heterogeneidad presente en un conjunto de datos numéricos.
Dispersión: grado de distanciamiento de un conjunto de valores respecto a un valor representativo.

Medidas de tendencia central: media y mediana

Se denomina **medida de tendencia central** al número que describe de manera aproximada donde están localizados o centrados los datos con un valor representativo de las observaciones. Las más utilizadas con la media y la mediana.

Revisa más de este tema en:
Media y Mediana es un tema que podrás revisar también en el **Módulo 20** página 157.

La media

La **media** o **promedio muestral** de un conjunto de datos numéricos es el resultado de la suma de las observaciones dividido entre el número de observaciones mismas.

Es útil contar con una notación concisa para la variable sobre la cual fueron hechas las observaciones, para el número de observaciones en el conjunto de datos y para cada una de las observaciones individuales en la forma siguiente:

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

El subíndice de la x no tiene relación con la **magnitud** de la observación, sólo indica el número de observación es el conjunto de datos y no necesariamente es la más pequeña (x_1) o la más grande (x_n) . La suma de $x_1, x_2, x_3, \ldots, x_n$ se puede representar con la letra griega ∑ pues representa la suma de todos los valores de x en los datos bajo consideración.

Magnitud: propiedad física que puede ser medida; por ejemplo, la temperatura, el peso, etc.

La **media muestral** de un conjunto de observaciones $x_1, x_2 \ldots x_n$ representada por $\bar{x}$ se define como:

$$\bar{x} = \frac{\text{suma total de observaciones en la muestra}}{\text{número de observaciones en la muestra}} = \frac{x_1 + x_2 + \ldots, x_n}{n} = \frac{\sum x}{n}$$

EJEMPLO:

Los ortopedistas de un hospital comunitario que atienden a personas de la tercera edad con pérdida de movilidad en la rodilla, realizaron un estudio para investigar el efecto que una nueva terapia tiene sobre la movilidad de la rodilla. Para ello, midieron la capacidad de extender la rodilla. Loa valores para extensión de la rodilla (en grados) para cada uno de los 10 participantes en el estudio se da a continuación:

$x_1 = 59$, $x_2 = 46$, $x_3 = 64$, $x_4 = 49$, $x_5 = 56$, $x_6 = 70$, $x_7 = 45$, $x_8 = 52$, $x_9 = 63$, $x_{10} = 52$

La suma de estos valores muestrales es 59 + 46 +64 + …52=556 y la media muestra es la extensión de la rodilla es:

$$\bar{x} = \frac{\sum x}{n} = \frac{556}{10} = 55.6$$

Existen otros conjuntos de mediciones cuyo valor promedio calculado no es parte del conjunto de valores observados, más aún, por la naturaleza de la variable su valor ni siquiera tiene alguna posibilidad de ser observado, sin embargo, puede considerarse representativo porque la mayoría de las cantidades observadas se encuentran concentradas alrededor de su valor.

La media de muestra $\bar{x}$ se calcula a partir de observaciones de la muestra, lo que convierte en una característica de esa muestra. Las características de la población se suelen designar por letras griegas. Una de las más importantes de estas características es la media poblacional.

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

Para el caso de una distribución de probabilidad normal, el valor de la **media** μ es el que divide en dos partes iguales a esta distribución de probabilidad. El valor $\bar{x}$ varía de una muestra a otra, mientras que sólo hay un valor de μ.

Una *posible desventaja potencial de la media*, como medida de tendencia central de un conjunto de datos, es que su valor puede verse afectado por la presencia de valores extremos o atípicos en el conjunto de datos.

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 menor a mayor, la **mediana** es el valor medio de la lista y es el que divide en dos partes iguales.

La **mediana** de la muestra se obtiene al ordenar un conjunto de n observaciones tomando como criterio de orden iniciar con el valor más pequeño y terminando 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 medio si n es par.

EJEMPLO:
El tamaño de la muestra para los datos del acceso al sitio web es de 40 estudiantes, se inscribieron a un curso de educación a distancia. El programa lleva la cuenta de la frecuencia con que cada estudiante accede a las páginas web de la clase. El instructor solicitó un informe del acceso al curso:

Número de veces que cada estudiante accede al sitio de la clase											
20	37	4	20	0	84	14	36	5	331	19	0
0	22	3	13	14	36	4	0	18	8	0	26
4	0	5	23	19	7	12	8	13	16	21	7
13	12	8	42								

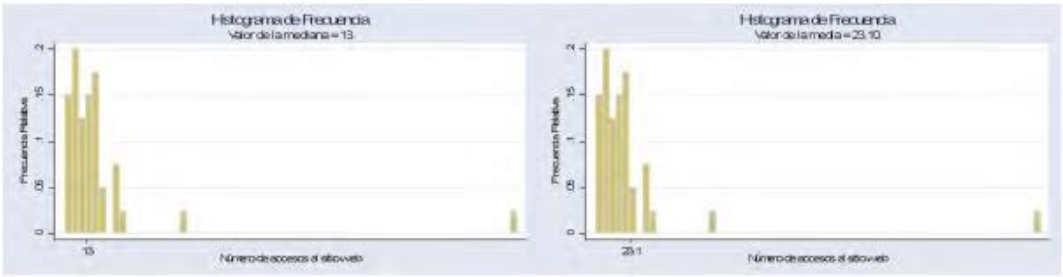
La media muestral para este conjunto de datos es de $\bar{x} = 23.10$. El tamaño de la muestra es $n = 40$. La **mediana** es el promedio de los valores para las observaciones 20 y 21 (el centro de los datos o la mitad) en la lista ordenada. La organización de menor a mayor genera la siguiente lista (con los dos valores centrales resaltados):

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 determinó como:

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

Al mirar los siguientes histogramas para este conjunto de datos —que muestran los valores marcados sobre el eje x para la mediana y la media respectivamente— el valor de la mediana parece ser más típico que el valor de la media $\bar{x} = 23.10$. Pues se encuentra más hacia el centro del conjunto de los datos.

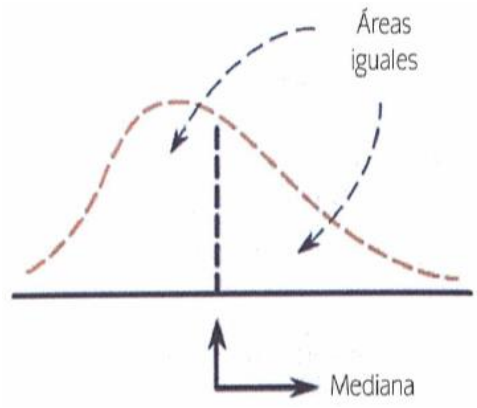


La media de la muestra puede ser sensible incluso a un único valor que se encuentra muy por encima o por debajo el resto de los datos. La mediana, por el contrario, es bastante insensible a los valores atípicos. La estabilidad de la mediana es lo que a veces justifica su uso como una medida de tendencia central.

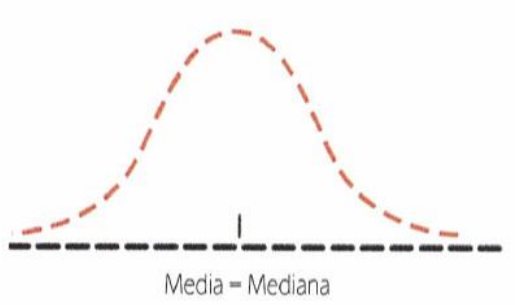
La **moda muestral** de un conjunto de números es otra medida alternativa de tendencia central. Se define como el valor que ocurre con mayor frecuencia en un conjunto de datos, es decir, el valor más frecuente. Por ejemplo, el conjunto de observaciones 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12 y 18 tiene moda 9.

Comparación de la media y la mediana

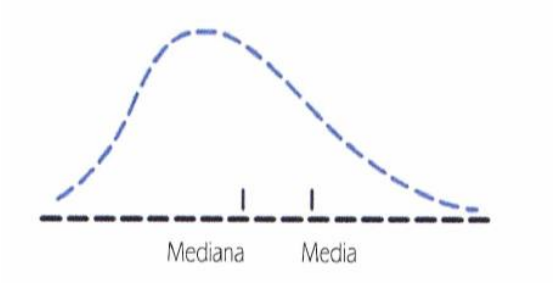
La siguiente figura muestra un histograma en forma de una curva suave que podría representar la distribución de valores de una muestra, o bien la de una población. Gráficamente, la mediana es el valor en el eje de medición que separa a la curva en dos partes, con 0.5 (50%) del área bajo la curva en cada parte.



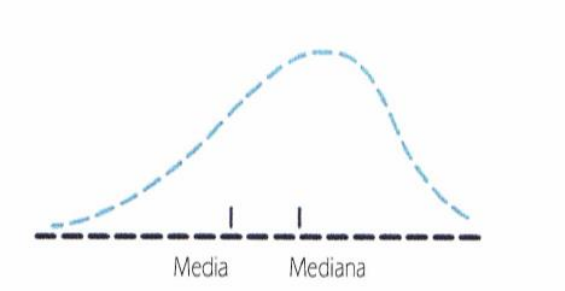
Cuando el histograma es simétrico, como es el caso de la curva de una distribución normal, como la que se muestra en la siguiente figura, el punto de simetría es el punto de división de la curva en dos áreas iguales y además es el punto de equilibro. En este caso la media y la mediana son iguales.



Por otra parte, cuando el histograma tiene un solo pico (es unimodal), con una cola superior más larga, como se muestra en la siguiente gráfica (a esta característica de la gráfica 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. Por ejemplo, en un examen, una calificación inusualmente alta, eleva la media pero no afecta a 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.



Puede darse el casi de que la moda no exista: por ejemplo el conjunto de datos 3, 5, 8, 10, 12, 15 y 16 pero también que esta no sea la única; por ejemplo el conjunto de datos 2, 3, 4, 4, 4, 5, 5, 7, 7, 7 y 9 tiene dos modas, 4 y 7, y se llama **bimodal**.

La media y la mediana se utilizan comúnmente como parámetros para comparar dos distribuciones de población.

2.2 es una cantidad numérica que puede servir como medida de tendencia central para un conjunto de datos categóricos. Es la frecuencia relativa o proporción muestral de una categoría. La proporción muestral de una categoría es la proporción (fracción) de respuestas que están en la correspondiente categoría. Por ejemplo, si una variable categórica tiene sólo dos posibles respuestas (hombre o mujer, enfermo o sano) y etiquetamos a cada una de estas dos posibles respuestas como E (éxito) y la otra F (fracaso), la proporción de éxitos (E) es llamada la proporción muestral de éxitos.

La proporción muestral de éxitos, representada por $\hat{p}$, es:

$$\hat{p} = \frac{\text{número de éxitos en la muestra}}{n}$$

Medidas de dispersión

Una medida de tendencia central sólo proporciona información parcial sobre un conjunto de datos, ya que no permite cuantificar que tanto las observaciones difieren unas de otras. Aunque varias muestran tengan una media y mediana iguales, la dispersión de los datos puede ser completamente distinta.

El **Rango muestral** de un conjunto de datos es definido como:

Rango = observación más grande – la observación más pequeña.

En general, una mayor variabilidad reflejará un rango muestral más amplio. Sin embargo, la variabilidad es una característica de cualquier conjunto de datos. Si no se consideran todos los datos para el cálculo del rango, la información sobre la verdadera naturaleza de la variación de un conjunto de datos puede ser poco confiable.

La limitación del rango hace necesario contar con otras medidas de dispersión con capacidad para incorporar la contribución que tiene cada observación a la variabilidad total del conjunto de datos.

Desviación de la media

La 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.

Dada una muestra de tamaño n , las n desviaciones de la media muestral son las diferencias

$$\left(x_1 - \bar{x}\right), \left(x_2 - \bar{x}\right), ..., \left(x_n - \bar{x}\right)$$

Una desviación en particular será positiva si el valor de x correspondiente es mayor de $\bar{x}$ y negativo si el valor de x es menor que $\bar{x}$. Algunas de las desviaciones son muy grandes en **magnitud absoluta**, indicando con esto las observaciones que están más lejos de la media muestral.

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.

A mayor variabilidad en la muestra, mayores serán las magnitudes de las desviaciones. Para combinar las desviaciones en una sola medida de variabilidad se podría calcular la desviación promedio al sumar todas las desviaciones medias (representada mediante $\sum(x - \bar{x})$) y luego dividirla por n .

La limitación en la aplicabilidad de la desviación media como una medida general de la variabilidad de un conjunto de datos muestrales, consiste en que las **contribuciones** 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.

Contribución: se define como tal al valor numérico asociado con una desviación media. 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.

Varianza y desviación estándar

Una manera de evitar que se eliminen las contribuciones negativas con las positivas es elevarlas al cuadrado antes de sumarlas. La hacer esto, las contribuciones con diferente signo pero con la misma magnitud o se anularán al sumarse porque bajo esta medida tendrán idéntica contribución a la variabilidad. Elevar las contribuciones al cuadrado equivale a elevar al cuadrado cada una de las desviaciones de la media. El cuadrado de estas desviaciones para muestras de tamaño n se representa como $(x_1 - \bar{x})^2$, $(x_2 - \bar{x})^2$,…, $(x_n - \bar{x})^2$, así que su respectiva suma es:

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

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 $\sum\left(x - \bar{x}\right)^2$ 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 muestra y se representa como S.

Una ***mayor variabilidad*** en la muestra da como resultado un valor *relativamente grande* de S^2 o S, mientras que un valor de S^2 o S cercano a *cero* indica una ***menor variabilidad***. Se debe tomar en cuenta que cualquier unidad utilizada para medir la variable x y el cuadrado de sus desviaciones se encuentran en unidades al cuadrado, por lo que, al tomar la raíz cuadrada, esta tendrá las mismas unidades que la variable x.

EJEMPLO:

Un laboratorio multinacional produce vacunas contra el virus de la influenza y las distribuye en diversos países, cuyo costo es variable según el país. En la tabla siguiente se muestran los datos correspondientes al costo de una dosis de la vacuna (convertido a dólares estadounidenses) para algunos países de América Latina, donde para este conjunto de datos $\overline{x} = \$3.36$. Calcular la desviación estándar y la varianza muestral. En la siguiente tabla se muestra el cálculo de cada una de las desviaciones de la media y la suma:

	x	$(x - \overline{x})$	$(x - \overline{x})^2$
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
	$\overline{x} = 3.36$		$\sum(x - \overline{x})^2 = 2.4643$

A partir de este valor obtenemos los valores de la varianza y la desviación estándar de la siguiente forma:

$$S^2 = \frac{\sum(x-\overline{x})^2}{n-1} = \frac{2.4643}{7-1} = \frac{2.4643}{6} = 0.4107$$

$$S = \sqrt{0.4107} = 0.641$$

Hay medidas de variabilidad para toda la población, que 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 σ^2 y σ, respectivamente.

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

El valor de σ suele desconocerse, en su lugar usamos un valor muestral calculado que esperamos sea cercano a σ; es decir, un valor que sea un buen estimador de σ. Por esta razón se debe calcular el valor de S^2 a partir del cociente cuyo denominador es $n-1$; ya que en un muestreo repetitivo, en lugar de n, el valor S^2 resultante tiende a estar más cerca de σ^2 en promedio.

Las medidas de dispersión o variabilidad informan sobre cuánto se alejan del centro los valores de dispersión. La medición de la variabilidad es un parámetro para comparar dos distribuciones de población.

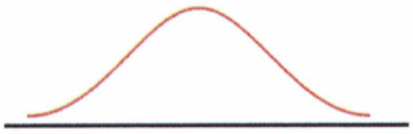
ACTIVIDAD:

- ✓ Realiza la actividad 1 sobre **medidas de tendencia central y dispersión** de la página 104 de la versión electrónica correspondiente a este módulo, corrobora tus respuestas en la pág., 180 del Apéndice 1.

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

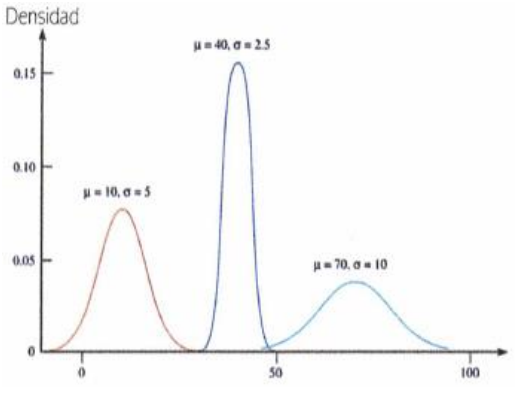
Distribución de probabilidad normal

Con anterioridad definimos a las **distribuciones normales** como distribuciones de probabilidad continuas o de curvas normales que tienen forma de campana y son simétricas, como se muestra en la siguiente figura.



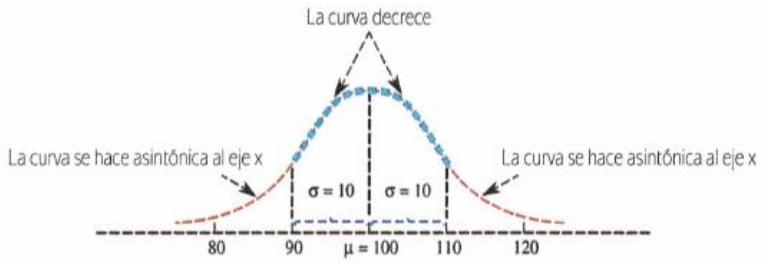
Los tipos de distribuciones normales se distinguen entre sí por los valores de su media μ y su desviación estándar σ. Mientras que la media μ de una distribución normal describe dónde está centrada la curva correspondiente, la desviación estándar define la dimensión de la curva cuando se extiende alrededor de ese centro. Al igual que todas las distribuciones de probabilidad continua, el área total bajo cualquier curva normal es igual a 1.

La siguiente figura muestra tres distribuciones normales en las que se puede observar cuánto es menor la desviación estándar o 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 concentran a una distribución menor de su media.



El valor de la media es el número que en el eje de medición (eje x) se extiende debajo de la parte superior de la campana. El valor de la desviación estándar σ también puede determinarse

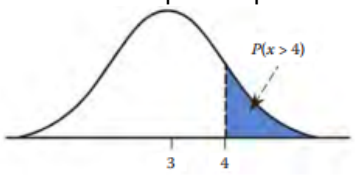
a partir de la gráfica de una curva normal, tal y como se observa en la curva normal de la siguiente figura.



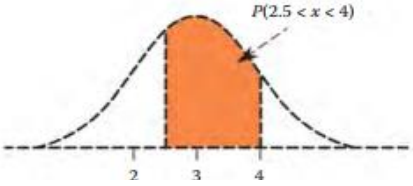
En la parte superior de la campana y hacia la derecha, la curva tiende a bajar hasta que está alrededor del valor 110. Después de ese punto, continúa decreciendo, de manera menos notable, haciéndose **asintótica** al eje x. En general, σ es la máxima distancia a cada lado de μ, a partir de la cual una curva normal disminuye su decrecimiento. Si una distribución normal se utiliza como un modelo poblacional, su media y su desviación estándar deben especificarse.

Asintótico: término utilizado en matemáticas para describir el comportamiento de la gráfica de una función que consiste en la aproximación de la curva de la gráfica de una función a una línea recta horizontal (por arriba o por abajo) pero sin llegar jamás a cortarla.

EJEMPLO:
Una distribución normal con media 3 y desviación estándar 1 podría usarse como modelo para la distribución del peso de niños recién nacidos. Si este modelo es una descripción razonable de esta distribución de probabilidad podríamos usar áreas bajo la curva normal con μ = 3 y σ = 1 para aproximar probabilidades de diversos eventos relacionados con el peso de un recién nacido. La probabilidad de que el peso de un recién nacido sea mayor a 4 kg —expresado simbólicamente como $P(x > 4)$— corresponde al área sombreada en la siguiente gráfica "Distribución normal para el peso de un recién nacido".



Por otra parte, la probabilidad $P\ (2.5 < x < 4)$ de que un recién nacido pese entre 2.5 y 4 kg corresponde al área sombreada en la siguiente gráfica.



El cálculo directo de estas probabilidades (áreas bajo la curva normal) es complejo por lo que se utiliza una tabla de áreas para una distribución normal de referencia, llamada distribución normal estándar.

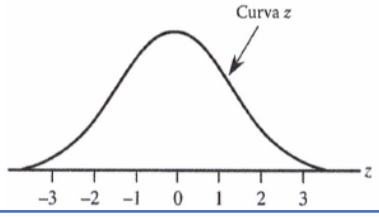
La **distribución normal estándar** es una **distribución normal** con $\mu = 0$ y $\sigma = 1$. La curva de densidad correspondiente también es denominada curva Z o curva normal estándar. Se acostumbra utilizar la letra Z como la variable para representar los posibles valores que puede tomar una distribución normal.

Algunas variables pueden describirse mediante la distribución normal estándar, que también se utiliza en los cálculos de probabilidad para otras distribuciones normales.

La distribución normal estándar

Para trabajar con la **distribución normal estándar** es necesario saber calcular probabilidades o áreas bajo una curva normal estándar por encima de los intervalos dados de posibles valores para Z.

La curva normal estándar (Z) que se muestra en la siguiente figura está centrada en $\mu = 0$ y tiene desviación estándar $\sigma = 1$; 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).



Uso de la tabla normal estándar para el cálculo de probabilidades

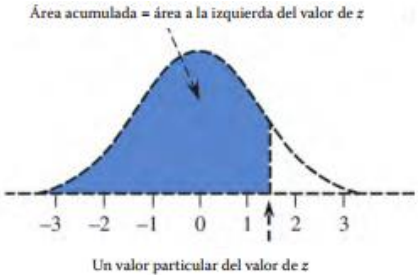
Para cualquier número z^* entre -3.89 y 0, redondeado a dos cifras decimales, la tabla del Apéndice 4 calcula los valores $P(z<z^*)$, cuya expresión representa el área de la curva acumulada a la izquierda de z^*.

Para calcular el valor de esta probabilidad con la tabla se deben llevar a cabo dos pasos:

1. Localizar el renglón correspondiente al dígito a cada lado del punto decimal del valor observado de z^*. Por ejemplo si el valor observado es $z^*=-1.75$, en la tabla se debe localizar el renglón correspondiente a z^* de -1.7.
2. Ubicar la columna identificada con el segundo dígito a la derecha del punto decimal en z^*. Por ejemplo, con el valor observado de $z^*=-1.75$ en la tabla se debe localizar la columna correspondiente al valor de .05.

La probabilidad deseada $P(z<z^*)$ es el número en la intersección de este renglón con esta columna. Por ejemplo, para $z^*=-1.75$ la probabilidad deseada es:
$P(z<-1.75) = 0.0401$.

EJEMPLO 1:
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).
Ahora bien, 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 y del tipo mostrado en la siguiente figura:



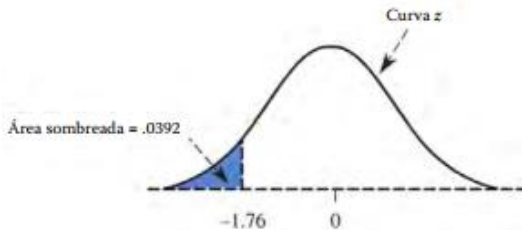
En esta tabla, el valor más pequeño de Z * es -3.89, el siguiente es -3.88, y así sucesivamente en incrementos de 0.01 terminando con el área acumulada total a la izquierda de 3.89 de 1 o 100%.
En la siguiente sección de la tabla de la curva normal estándar se puede ubicar el área bajo la curva z a la izquierda de 1.42. Para hacerlo hay que buscar la fila con la etiqueta 1.40 y la columna etiquetada con 0.02 (la fila y la columna resaltadas en azul). El valor correspondiente a la intersección de esta fila y esta columna es 0.9222. Por lo tanto, el área de la curva z a la izquierda de 1.42 = 0.9222.
También se puede utilizar en la tabla el área a la derecha de un valor particular. Debido a que el área total bajo la curva de z es 1, se deduce que el área de la curva z a la derecha de 1.42 es 1-el área de la curva z a la izquierda de: 1.42 = 1-0.9222 = 0.0778. Estas probabilidades se pueden interpretar en el sentido de que en una larga secuencia de observaciones, aproximadamente 92.22% de los valores observados z será menor que 1.42 y 7.78% será mayor de 1.42.

z*	0.00	0.01	0.02	0.03	0.04	0.05
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678

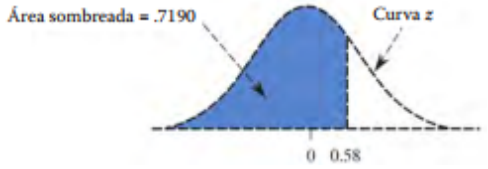
EJEMPLO 2:

La probabilidad $P(z<-1.76)$ se encuentra en la intersección de la fila -1.7 y de la columna de 0.06. El resultado es $P(z<-1.76)=0.0392$ como se muestra en la siguiente figura:

Tabla de probabilidades acumuladas de una distribución normal estándar										
z*	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367



En otras palabras, en una secuencia larga de observaciones 3.9% de los valores observados z, aproximadamente, será menor que -1.76. Del mismo modo, el valor corresponde P(z ≤ 0.58) al valor 0.7190 identificado en la intersección de la fila 0.5 y la columna 0.08 de la tabla. El área bajo la curva correspondiente a esta probabilidad se muestra en la siguiente gráfica:



Ahora considérese $P(z < -4.12)$. Esta probabilidad no aparece en la tabla del Apéndice 4, pues no hay un renglón con el valor de -4.1. Sin embargo, debe ser menor que $P(z < -3.89)$, el valor más pequeño de z* en la tabla, ya que -4.1 está más lejos en la cola izquierda de la curva z. Como $P(z < -3.89) = 0.0000$ (esto es cero, redondeado a cuatro cifras decimales), se sigue que:

$$P(z < -4.12) \approx \text{(se lee aproximadamente)}\ 0.0000$$

De manera similar,

$$P(z < 4.12) > P(z < 3.89) = 1.0000$$

de donde concluimos que

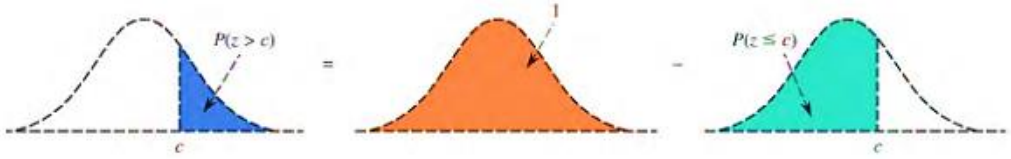
$$P(z < -4.12) \approx 1.0000$$

Podemos usar las áreas acumuladas tabuladas en la tabla del Apéndice 4 para calcular otras probabilidades que involucran a z. La probabilidad de que z sea mayor que un valor c es:

$$P(z > c) = \text{área bajo la curva z al lado derecho de c.}$$

O en forma equivalente $P(z > c) = 1 - P(z \le c)$

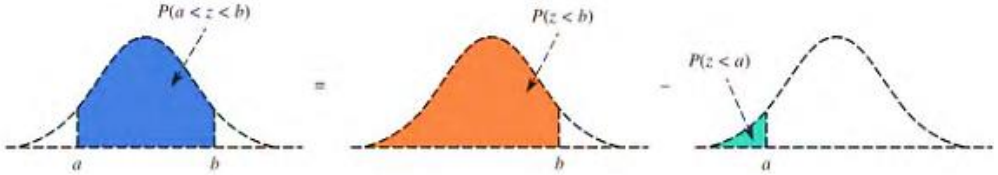
En otras palabras, el área a la derecha de un valor (un área de cola derecha) es 1 menos su correspondiente área acumulada, como se ilustra en la siguiente gráfica.



De igual manera, la probabilidad de que z caiga en un intervalo con límite inferior y límite superior b es

$P(a < z < b) =$ área bajo la curva z en el intervalo de a a $b = P(a < b) - P(z < a)$.

Esto es, $P(a < z < b)$ es la diferencia entre dos áreas acumuladas, como se muestra en la siguiente figura.

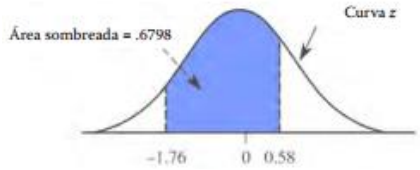


EJEMPLO 3:

La probabilidad de que z esté entre -1.76 y 0.58 es

$$P(-1.76 < z < 0.58) = P(z < 0.58) - P(z < -1.76)$$
$$= 0.7190 - 0.0392$$
$$= 0.6798$$

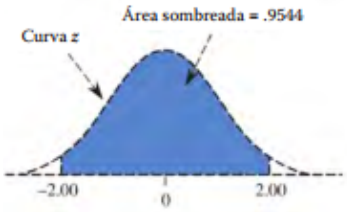
Como se muestra en la siguiente figura:



La probabilidad de que z esté entre 2 y -2 (dentro de dos desviaciones estándar de la media, ya que μ = 0 y σ = 1) es

$$P(-2 < z < 2) = P(z < 2) - P(z < -2)$$
$$= 0.9544$$
$$= 0.95$$

como se muestra en la siguiente gráfica:

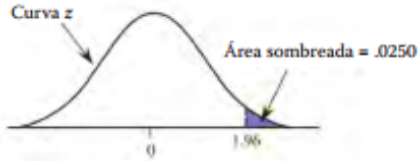


Esta última probabilidad es la base para una parte de la regla empírica que establece que cuando un histograma está bien aproximado por una curva normal, aproximadamente 95% de los valores están dentro de dos desviaciones estándar de la media.

La probabilidad de que el valor de z exceda 1.96 es

$$P(z > 1.96) = 1 - P(z \leq 1.96)$$
$$= 1 - 0.9750$$
$$= 0.0250$$

como se muestra en la siguiente figura:

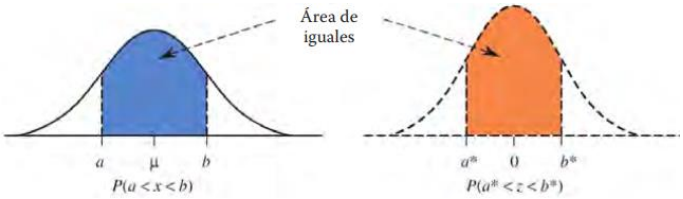


Es decir, 2.5% del área bajo la curva z se encuentra a la derecha de 1.96, en el extremo superior. De igual forma:

$$P(z > -1.28) = \text{área a la derecha de } -1.28$$
$$P(z > -1.28) = 1 - P(z \leq -1.28)$$
$$= 1 - .1003$$
$$= .8997$$
$$= .90$$

Otras distribuciones normales

Las áreas de la curva z también pueden emplearse para el cálculo de probabilidades de cualquier distribución normal. Usaremos la letra z sólo para aquellas variables que tiene una distribución normal estándar y la letra x para aquellas variables que tienen una distribución normal estándar y la letra x para cualquier variable cuya distribución se describe mediante una curva normal con media μ y desviación estándar σ. Si queremos calcular P(a<z<b), la probabilidad de que la variable x se encuentre en un rango determinado.



Para obtener esta probabilidad se debe encontrar un problema equivalente, relacionado con la distribución normal estándar. La determinación de un intervalo (a*, b*), tiene la misma probabilidad para z, al igual que el intervalo (a, b) en nuestra distribución normal original, como se muestra en la siguiente figura en la que se representa la igualdad entre las áreas de una curva normal estándar y una curva normal y uno no estándar.

La notación asterisco (*) se utiliza para distinguir a y b (los valores de la distribución normal original con media μ y desviación estándar σ), de a* y b *, los valores correspondientes a la curva de z. Para encontrar a* y b* se deben calcular las puntuaciones z para los puntos finales del intervalo para el que se desea calcular una probabilidad. Este proceso se llama normalización de los criterios de valoración.

La puntuación z es la puntuación típica o resultado estándar individual más utilizada. Gracias a la puntuación z podemos comparar los resultados de un individuo (o de varios) en distintas pruebas en las que las distribuciones de los resultados de los otros miembros de la población sean distintas.

La puntuación z correspondiente a un valor particular es

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

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

EJEMPLO

Si la variable x tiene una distribución normal con media μ = 100 y desviación estándar σ =5, para encontrar: $P(98 < z < 107)$

lo primero que se debe hacer es convertir este problema en uno equivalente a la distribución normal estándar. Para ello, comenzaremos 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 puntaje estandarizado dado por

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

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.

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

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

Donde z es una variable cuya distribución es normal estándar y

$$a^*=\frac{a-\mu}{\sigma} \quad b^*=\frac{b-\mu}{\sigma}$$

EJEMPLO

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

Para responder esta pregunta, debemos encontrar: $P(2900<x<4700)$

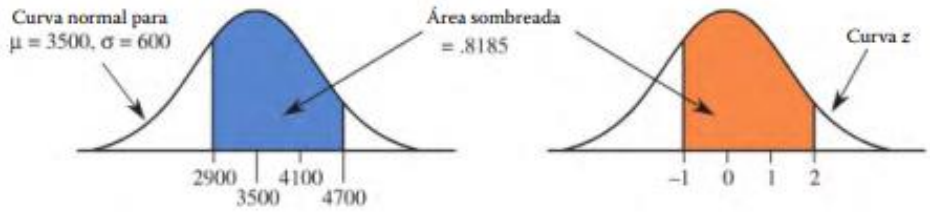
En primer lugar, se transforman los extremos $a=2900$ y $b=4700$ del intervalo correspondiente a la distribución de la variable x, a los extremos a^* y b^* del intervalo equivalente a la distribución normal estándar:

$$a^*=\frac{a-\mu}{\sigma}=\frac{2900-3500}{600}=-1.00$$
$$b^*=\frac{b-\mu}{\sigma}=\frac{4700-3500}{600}=2.00$$

Entonces:

$$P(2900<x<4700)=P(-1.00<z<2.00)$$
$$=0.9772-0.1587$$
$$=0.8185$$

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



Este resultado significa que si se midieran los pesos al nacer para muchos bebés de esta población, el peso del 82% de ellos se situaría entre 2900 y 4700 gramos. Ahora bien, ¿cuál es la probabilidad de que un niño elegido al azar tenga un peso al nacer mayor de 4500 gramos? Para calcular esta probabilidad debemos evaluar $P(x > 4500)$.

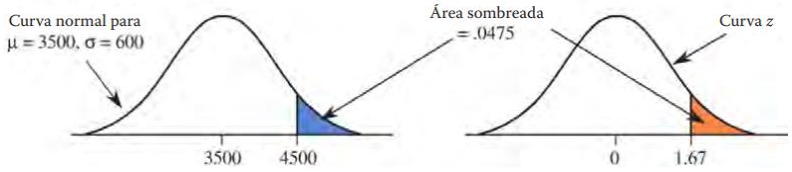
Así que primero calculamos: $a^* = \frac{a-\mu}{\sigma} = \frac{4500-3500}{600} = 1.67$

Entonces:

$$P(x > 4500) = P(z > 1.67)$$

= área de la curva z a la derecha de 1.67
= 1 – (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

Cuando las características de un experimento son de un carácter **binomial** la fórmula para calcular las probabilidades de sus valores es:

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

$$P(x) = P(x \text{ éxitos en } n \text{ ensayos}) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$
$$x = 0, 1, \ldots, n$$

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 x elementos de un conjunto de n elementos.
El símbolo de $n!$ (Léase " n factorial ") se define como

$$n! = n(n-1)(n-2)\ldots(2)(1)$$

y 0! = 1

La fórmula de la probabilidad binomial puede entonces expresarse como

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

La fórmula de la distribución binomial se puede deducir a partir de la aplicación de conceptos de probabilidad básica. Si tienes interés en conocer como se deduce esta fórmula puedes consultar el Apéndice 5.

EJEMPLO 1:

Si quisiéramos registrar el sexo de cada uno de los próximos cinco niños recién nacidos en un hospital comunitario, podríamos cuestionarnos: ¿Cuál es la probabilidad de que no de los cinco sea mujer?, ¿cuál de que entre dos y cuatro sean mujeres? Antes de responder las dos preguntas centremos nuestra atención en la naturaleza del experimento. Si lo analizamos, podemos percatarnos que éste posee las siguientes características:

1. El experimento consiste de $n = 5$ intentos (ensayos) idénticos, pues cada intento consiste en registrar el sexo de un recién nacido.
2. Cada intento da lugar exactamente a dos resultados, ya que cada recién nacido sólo puede ser registrado como hombre o mujer.
3. Los cinco intentos son independientes pues el resultado obtenido de cada registro es independiente de los resultados obtenidos de 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).

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

$$P(1) = P(x=1)$$
$$= \binom{5}{1}(0.4)^1(0.6)^4$$
$$= 0.26$$

Y si ahora queremos calcular la probabilidad de que entre dos y cuatro recién nacidos sean mujeres obtenemos:

$$P(2 \le x \le 4) = P(x=2 \text{ o } x=3 \text{ o } x=4)$$

Dado que estos resultados son disjuntos, esto es igual a:

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

EJEMPLO 2:

Un botánico realiza la fertilización cruzada de especies afines de plantas, con flores blancas y azules, produciendo vástagos de los cuales 20% tiene flores blancas. Este experimento es una aplicación de la teoría de Mendel sobre caracteres hereditarios. El ensayo consiste en cruzar una planta de flores azules con una de flores blancas; un éxito es obtener de esta cruza un vástago de flores blancas y $p = 0.20$. De acuerdo con la teoría de Mendel, los ensayos son independientes; en consecuencia el experimento es binomial con $n = 6$ y $p = 0.2$. Si deseamos determinar la probabilidad de que a lo más haya dos flores blancas entre sus vástagos, se define la variable.

$x =$ el número de vástagos con flores blancas

Entonces las probabilidades para los posibles valores de x pueden calcularse a partir de la fórmula de la probabilidad binomial.

$$P(x) = \binom{6}{x}(0.2)^x (0.8)^{6-x}$$

$$x = 0,1,2,3,4,5,6.$$

La probabilidad de que haya a lo más dos flores blancas entre sus vástagos es:

$$\begin{aligned} P(x \le 2) &= P(x = 0 \text{ o } x = 1 \text{ o } x = 2) \\ &= P(x = 0) + P(x = 1) + P(x = 2) \\ &= \binom{6}{0}(0.2)^0 (0.8)^6 + \binom{6}{1}(0.2)^1 (0.8)^5 + \binom{6}{2}(0.2)^2 (0.8)^4 \\ &= \frac{6!}{0!6!}(0.2)^0 (0.8)^6 + \frac{6!}{1!5!}(0.2)^1 (0.8)^5 + \frac{6!}{2!4!}(0.2)^2 (0.8)^4 \\ &= (0.2)^0 (0.8)^6 + (6)(0.2)^1 (0.8)^5 + (15)(0.2)^2 (0.8)^4 \\ &= 0.2621 + 0.3932 + 0.2458 \\ &= 0.90 \end{aligned}$$

Distribución de probabilidad de Poison

Si el evento cumple con ciertas hipótesis entonces es posible suponer que se trata de un **experimento de Poisson**.

Experimento de Poisson: experimento en el que se analiza el número de veces que ocurre un evento por unidad de tiempo o unidad de espacio.

Si expresamos con x la variable que representa el número de sucesos resultantes de este experimento, entonces esta representa una variable discreta que puede tomar valores enteros. La fórmula que nos permitirá calcular las probabilidades de los valores asociados con x es:

Fórmula de la distribución de Poisson

Sea

λ = número medio de observaciones por unidad de tiempo
e = base de los logaritmos neperianos. Su valor aproximado es de 2.71828

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

Debe tomarse en cuenta que su una distribución de probabilidad se especifica mediante una fórmula, esta permite el cálculo de distintas probabilidades sin tener que recurrir al usa de una tabla o un histograma de frecuencias.

EJEMPLO

Ahora podemos calcular la probabilidad de que en una hora lleguen exactamente dos pacientes al tomar $x = 2$ y $\lambda = 3$ (número promedio de pacientes que arriban al servicio de emergencias por unidad de tiempo) y sustituir en la fórmula de la distribución Poisson:

$$p(2)=\frac{(3)^2 e^{-3}}{2!}$$
$$=\frac{(9)(0.04979)}{2}$$
$$=0.22$$

Asimismo, la probabilidad de que en una hora lleguen al menos dos pacientes es:

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

Dado que este resultado es difícil de evaluar pues considera una infinidad de eventos tantos como valores puede tomar la x). Entonces recurrimos al cálculo de esta probabilidad a partir de su complemento, de donde resulta que:

$$P(x\geq 2)=1-P(x<2)$$
$$=1-P(x=0 \text{ o } x=1)$$
$$=1-P(x=0)-P(x=1)$$
$$=1-P(0)-P(1)$$
$$=1-\frac{(3)^0 e^{-3}}{0!}-\frac{(3)^1 e^{-3}}{1!}$$
$$=1-e^{-3}-3e^{-3}$$
$$=.80$$

ACTIVIDAD:

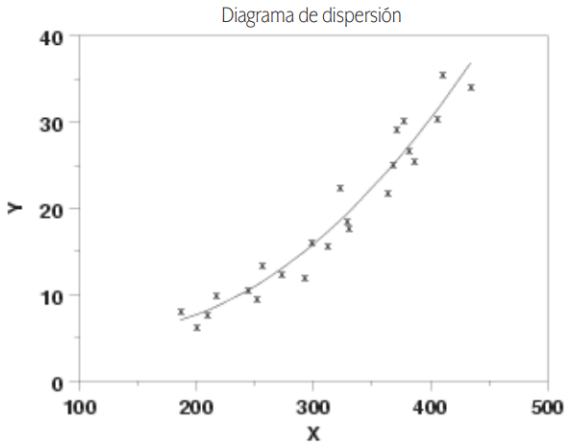
- ✓ Realiza la actividad 2 sobre **cálculo de probabilidades** de la página 121 de la versión electrónica correspondiente a este módulo, corrobora tus respuestas en las págs., 181 y 182 del Apéndice 1.

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

Los métodos que complementan las gráficas de dispersión permiten describir las relaciones entre dos variables numéricas y, también, evaluar la fortaleza de esta relación.

Correlación

Además de las gráficas de dispersión, para argumentar de forma precisa la tendencia con la que los valores de las observaciones pareadas se relacionan, hay que usar el **coeficiente de correlación muestral**, cuyo valor nos da una evaluación numérica de la tendencia y de la fuerza de la relación entre los valores x y y en un conjunto de datos con observaciones pareadas de la forma (x, y). El más utilizado es el de Pearson.

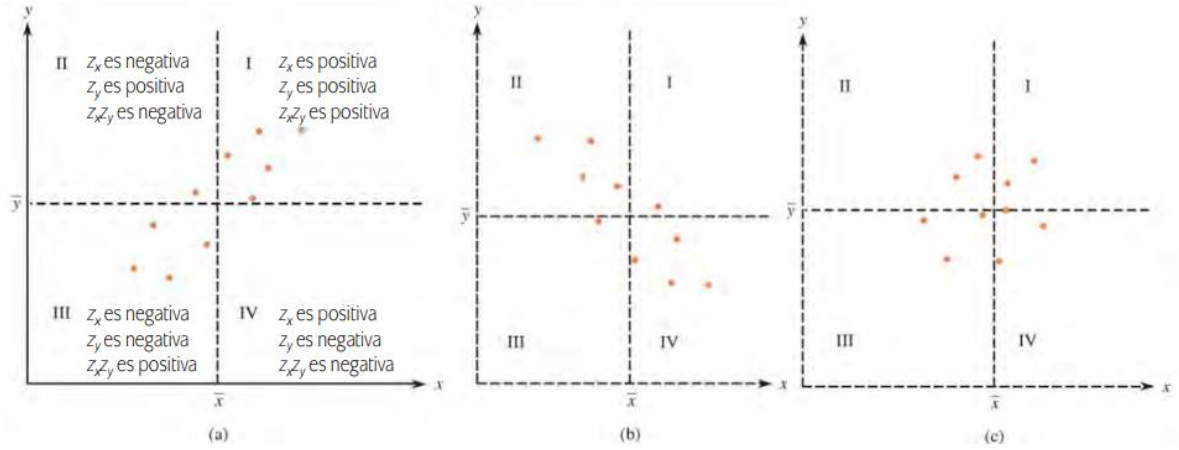


Coeficiente de correlación muestral de Pearson

El coeficiente de correlación muestral de Pearson se basa en la suma de los productos de z_x y z_y para cada observación en el conjunto de observaciones pareadas. Su notación algebraica es $\sum z_x z_y$.

Veamos algunas gráficas de dispersión para comprender con mayor profundidad el coeficiente descrito.

La gráfica (a) presenta una marcada tendencia creciente y por lo tanto una fuerte relación positiva. Si se traza una línea vertical que pase por x y una horizontal a través de la media y se divide el primer cuadrante del plano cartesiano en cuatro regiones.



En la región I, los valores graficados de x y y superan sus valores medios, por lo que la puntuación z_x de x y el puntaje z_y de y son números positivos. De ello resulta que $z_x z_y$ son positivos. El producto de las puntuaciones z también es positivo para cualquier punto de la región III ya que tanto la puntuación z_x como de z_y es negativa en la Región III y la multiplicación de dos números negativos da uno positivo. En cada una de las otras dos regiones, una puntuación z es positiva y la otra negativa, por lo que $z_x z_y$ es negativo. Pero

como en este caso, por lo general, los puntos caen en las regiones I y III, los productos de las puntuaciones z tienden a ser positivos. Por lo tanto, la suma de los productos será un número positivo relativamente grande. Aplicando un razonamiento similar para los datos mostrados en la gráfica (b), donde se exhibe una fuerte relación negativa, implica que $\sum z_x z_y$ será un número negativo relativamente grande (en magnitud). Cuando no hay una relación fuerte, como se muestra en la gráfica (c), los productos positivos y negativos tienden a contrarrestarse el uno con el otro, generando un valor de $\sum z_x z_y$ que es cercana a cero.

En resumen, $\sum z_x z_y$ parece ser una medida razonable para encontrar el grado de asociación entre x y y, el cual puede ser un número positivo grande, un número negativo grande, o un número cercano a 0, dependiendo de si la relación lineal positiva o negativa es fuerte, o no existe una fuerte relación lineal.

El coeficiente de correlación muestral de Pearson, se denota mediante r y se obtiene dividiendo $\sum z_x, z_y$, por (n – 1).

El coeficiente de correlación muestral de Pearson está dado por

$$\frac{\sum z_x z_y}{n-1}$$

Aunque hay varios coeficientes de correlación, el de Pearson es el más utilizado, por lo que el nombre de "Pearson" se omite con frecuencia y simplemente se denominada como el **coeficiente de correlación**.

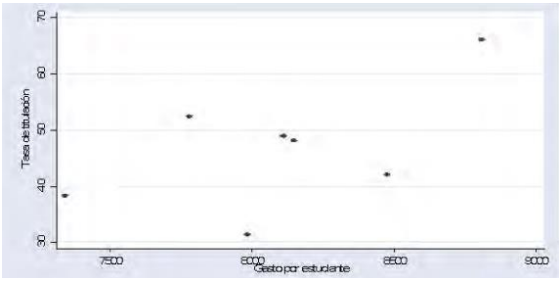
EJEMPLO:

Una preocupación permanente de las universidades públicas en países como México es la obtención de recursos para su financiamiento. En cierto estado del país existen siete universidades públicas con una matrícula registrada en conjunto de entre 10,000 y 20,000 estudiantes. Las tasas de titulación, así como el costo en pesos invertido por estudiante durante el último año en cada una de las siete universidades, se registraron de la siguiente manera:

Universidad	Tasa de titulación (%)	Gasto por estudiante ($)
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 siguiente es la gráfica de dispersión de estos datos.

Si x representa el gasto por estudiante y denota la tasa de graduación por año. Es fácil comprobar que: $\bar{x} = 8093.43 \;\; S_x = 472.39 \;\; \bar{y} = 46.73 \;\; S_y = 11.15$



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:

$$z_x = \frac{8810 - 8093.43}{472.39} = 1.52 \quad z_y = \frac{61.1 - 46.73}{11.15} = 1.74$$

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

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
				$\sum z_x z_y = 3.52$

Por lo que ahora con la información anterior es posible calcular el valor del coeficiente de correlación r:

$$\sum z_x z_y = 3.52$$

Entonces con n = 7

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

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.

Las propiedades del **coeficiente de correlación** (r) son:

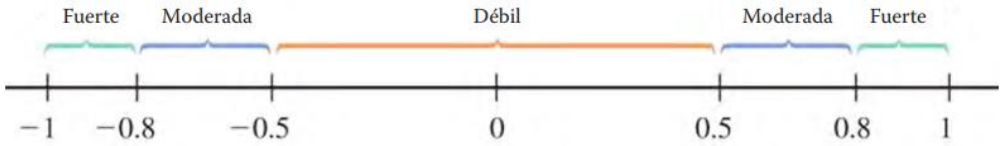
- El valor de r no depende de la unidad de medida de cada variable, así, si x en la altura, la correspondiente puntuación z es la misma; es decir, que si la altura se expresa en centímetros, metros o kilómetros el valor del coeficiente de correlación no se afecta.
- El valor de r no depende de cuál de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

- 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 e y en la muestra, r toma alguno de sus dos valores extremos posibles. Encontrar estos niveles de relación entre dos variables es muy raro.
- El valor de r es una medida del grado en que x e 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 e y, pues podría haber una fuerte relación que no es lineal entre x e y. De hecho la imposibilidad de detectar relaciones no lineales entre dos variables constituye una seria desventaja de esta medida de asociación.

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

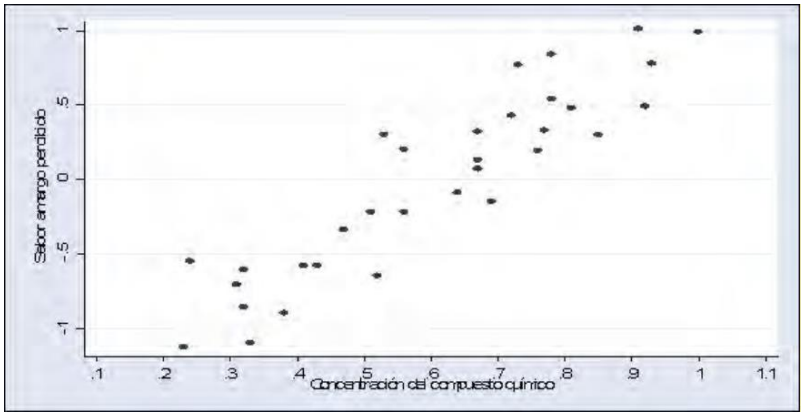
- Correlación débil si el coeficiente de correlación toma valores entre -0.5 y 0.5.
- Correlación moderada i sus valores están entre 0.5 y 0.8 o entre -0.5 y -0.8.
- Correlación fuerte si los valores del coeficiente se encuentran entre 0.8 y 0 o entre -0.8 y -1.



EJEMPLO:
Un laboratorio encargado de surtir medicamentos a un hospital realiza un estudio para determinar si existe una relación entre el sabor amargo de un jarabe pediátrico contra la tos y la concentración de un compuesto químico que se utiliza en su elaboración. A continuación se proporcionan los datos de x = concentración del compuesto químico y de y = el sabor amargo percibido, determinados por un panel de catadores del laboratorio.

x	y	x	y	x	y
0.72	0.43	0.76	0.19	0.52	-0.65
0.81	0.48	0.67	0.07	0.69	-0.15
0.92	0.49	0.56	-0.22	0.91	1.01
1.00	0.99	0.38	-0.90	0.64	-0.09
0.67	0.32	0.78	0.84	0.23	-1.13
0.53	0.30	0.67	0.13	0.78	0.54
0.51	-0.22	0.85	0.30	0.33	-1.10
0.56	0.20	0.41	-0.58	0.43	-0.58
0.77	0.33	0.93	0.78	0.32	-0.86
0.47	-0.34	0.31	-0.71	0.24	-0.55
0.73	0.77	0.32	-0.61		

El coeficiente de correlación r = 0.916 indica una fuerte relación positiva entre la concentración del compuesto químico y el sabor amargo percibido. Esto indica que conforme se aumenta la concentración del compuesto químico, la percepción del sabor amargo del jarabe tiende a aumentar. Se obtiene la siguiente gráfica de dispersión para este conjunto de datos:



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 e 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 e 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 (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 e y, ni de cuál variable que se denomine como x y cuál se denomine como y.

2. ρ = –1 o ρ = 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.

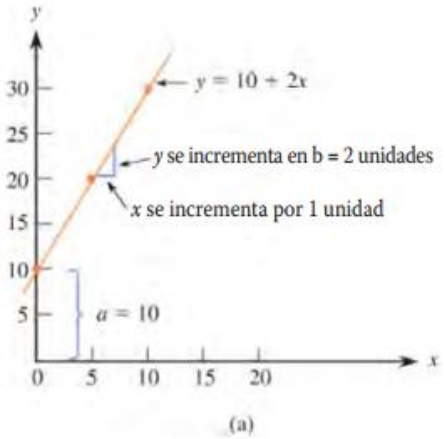
Es posible encontrar un coeficiente de correlación elevado entre dos variables que no tienen relación alguna, es decir, variables que no presentan relación justificada a través de alguna teoría específica (biología, astronomía, economía, entre otras). Cuando sucede esto, se dice que la correlación estadística existente entre estas variables es una correlación espúrea o sin sentido. De esta forma, es posible hablar de correlación espúrea entre variables relacionadas a la economía, a la biología, a la astronomía, entre otras.

Regresión lineal: ajuste de una recta a los datos bivariados

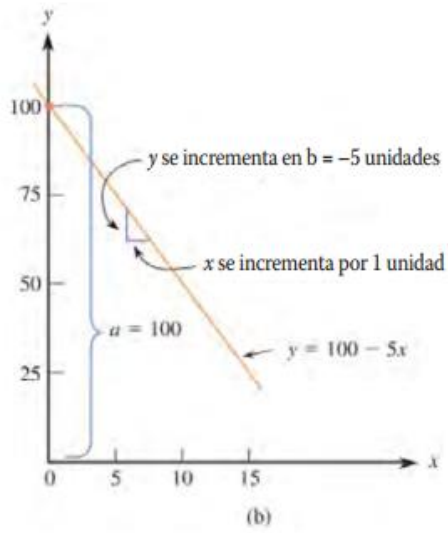
El objetivo del análisis de regresión es el uso de la información sobre una variable x, con la finalidad de obtener alguna conclusión acerca de una segunda variable, y.

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 independiente, predictora o explicativa**.

Características de las lineales y las relaciones lineales



(a)



(b)

La ecuación de una recta es $y = a + bx$. Una línea en particular se determina eligiendo los valores de a y b.

En la ecuación de una línea recta

$$y = a + bx$$

A la cantidad *b*, se le llama la pendiente de 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 resultante se representa en la gráfica (b).

Es fácil trazar la línea que corresponde a la ecuación de una línea recta en particular. Seleccionamos dos valores de x, cualesquiera, y los sustituimos en la ecuación para obtener los valores correspondientes a 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.

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

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.

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.

Para avaluar el ajuste global de una línea (bondad de ajuste) es 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

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

es la suma de los cuadrados de las desviaciones alrededor de la línea

$$\sum [y - (a + bx)]^2 = [y_1 - (a + bx_1)]^2 + [y_2 - (a + bx_2)]^2 + \cdots + [y_n - (a + bx_n)]^2$$

La línea de mínimos cuadrados, también llamada la 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.

La pendiente de la línea de mínimos cuadrados es

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

y la intersección es

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

Escribimos la ecuación de mínimos cuadrados como la línea

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

donde la ^ 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.

Si la pendiente y la intersección se van a calcular a mano, la siguiente fórmula de cálculo sirve para reducir el tiempo para realizar los cálculos.

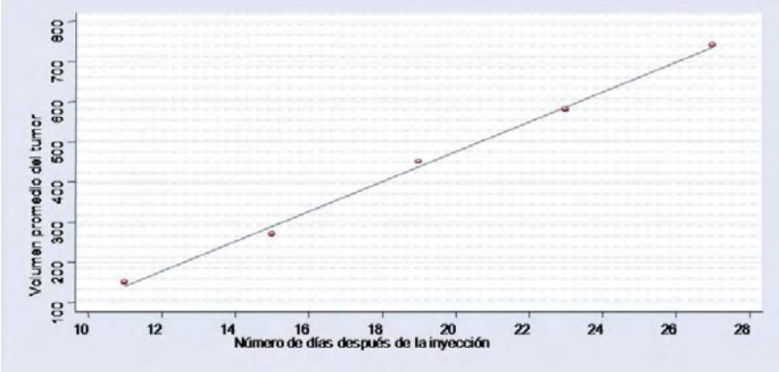
Fórmula para calcular la pendiente de la línea de mínimos cuadrados

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

EJEMPLO:
La granada, fruta originaria de Persia, se ha utilizado en la medicina popular de muchas culturas para el tratamiento de diversas dolencias. Los investigadores están ahora estudiando las propiedades antioxidantes de la granada para estimar su efecto benéfico en el tratamiento del cáncer. El tema de investigación de uno de los estudios es si el extracto de fruta de la granada (EFG) fue eficaz en la ralentización o el hacer más lento el crecimiento de los tumores del cáncer de próstata. En este estudio, 24 ratones fueron inyectados con células cancerígenas. Los ratones se asignaron aleatoriamente a uno de tres grupos de tratamiento. Un grupo de ocho ratones tomó agua potable normal, el segundo grupo de ocho ratones bebió agua potable complementada con 0.1% de EFG y el tercer grupo recibió agua potable complementada con 0.2% de EFG. El volumen tumoral promedio de los ratones en cada grupo se registró en varios puntos en el tiempo.
Los datos para las variables y = el volumen promedio de tumor (en mm^3) y x = número de días después de la inyección de las células cancerosas en los ratones que recibieron el agua potable normal se presentan a continuación:

x	11	15	19	23	27
y	150	270	450	580	740

La siguiente gráfica de dispersión de estos datos muestra que la relación entre el número de días después de la inyección de las células cancerosas y el volumen tumoral promedio podría, razonablemente, resumirse mediante una línea recta.



Las cantidades necesarias para calcular la ecuación de mínimos cuadrados correspondiente a este conjunto de datos son:

$$\sum x = 95 \quad \sum x^2 = 1965 \quad \sum xy = 47570$$

$$\sum y = 2190 \quad \sum y^2 = 1181900$$

A partir de estas ecuaciones calculamos:

$$\bar{x} = 19 \quad \bar{y} = 438$$

$$b = \frac{\sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}} = \frac{47570 - \frac{(95)(2190)}{5}}{1965 - \frac{(95)^2}{5}} = \frac{5960}{160} = 37.25$$

$$a = \bar{y} - b\bar{x} = 438 - (37.25)(19) = -269.75$$

La línea de mínimos cuadrados es entonces:

$$\hat{y} = -269.75 + 37.25x$$

El volumen promedio del tumor predicho para otro número de días después de la inyección de células cancerosas se calcula de manera similar. Debe tenerse cuidado al hacer predicciones con la línea de mínimos cuadrados, pues no se debe emplear para predecir el volumen promedio de un tumor para instantes del tiempo demasiado alejados del rango de 11 a 27 días (el rango de valores de x en el conjunto de datos), ya que no sabemos si el patrón lineal observado en la gráfica de dispersión continúa fuera de este rango. Esto a veces se conoce como el peligro de la extrapolación. En este ejemplo podemos ver que el uso de la línea de mínimos cuadrados para predecir el volumen tumoral promedio de menos de 10 días después de la inyección de las células cancerosas, lleva a predicciones sin sentido. Por ejemplo, si el número de días después de la inyección es de cinco, el volumen previsto del tumor promedio es negativo:

$$\hat{y} = -269.75 + (37.25)(5) = -83.5$$

Debido a que es imposible que el volumen del tumor promedio sea negativo, esto es una clara indicación de que el patrón observado para los valores de x en el rango de 11 a 27 no se continúa fuera de este rango. Sin embargo, la línea de mínimos cuadrados constituye una herramienta útil para hacer predicciones para valores de x dentro del rango de días 11 a 27.

ACTIVIDAD:

- ✓ Realiza la actividad 3 de la página 141 sobre cálculo del coeficiente de correlación muestral y el método de mínimos cuadrados, al conjunto de la versión electrónica correspondiente a este módulo, corrobora tus respuestas en el Apéndice 1.
- ✓ Realiza la actividad “Evalúa tu aprendizaje” de las páginas 146 a 148 del libro electrónico.

#Entrénate

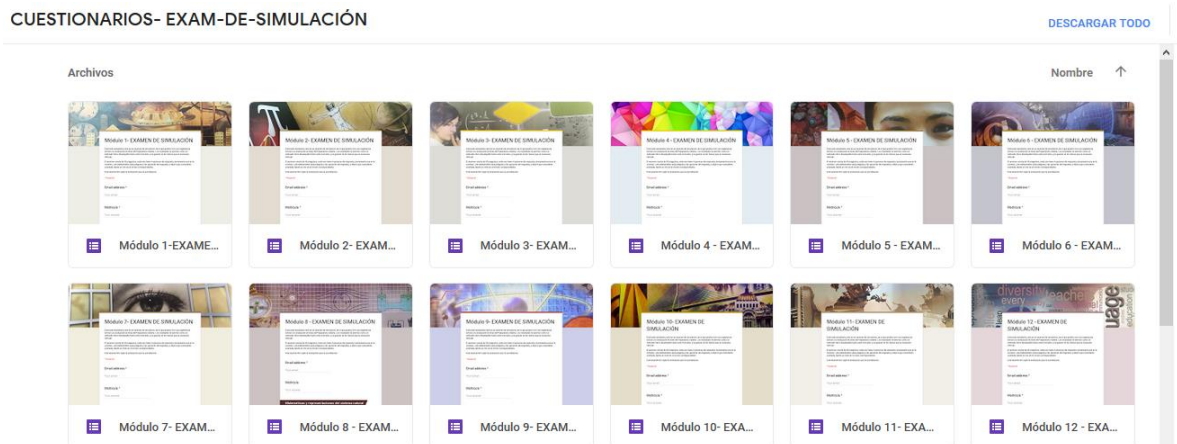
Hemos llegado al final de este módulo, ahora te toca revisar que tan preparado estás para el examen de acreditación.
Como pudiste observar en el Curso propedéutico de Comunicación, el examen para acreditar tu modulo tiene retos importantes en cuanto al manejo de Tecnologías de Información y Comunicación y el seguimiento de instrucciones.

Antes de realizar el examen real, entrénate con la evaluación **¿Ya estoy preparado(a)?** que se encuentra en la sección final del libro, donde al terminar podrás verificar tus respuestas en el **Apéndice 1**. También prepárate con nuestro **simulador de examen**, el cual te servirá como un indicador de tu desempeño y te guiará en los temas necesarios a reforzar, para ello sigue las siguientes indicaciones:

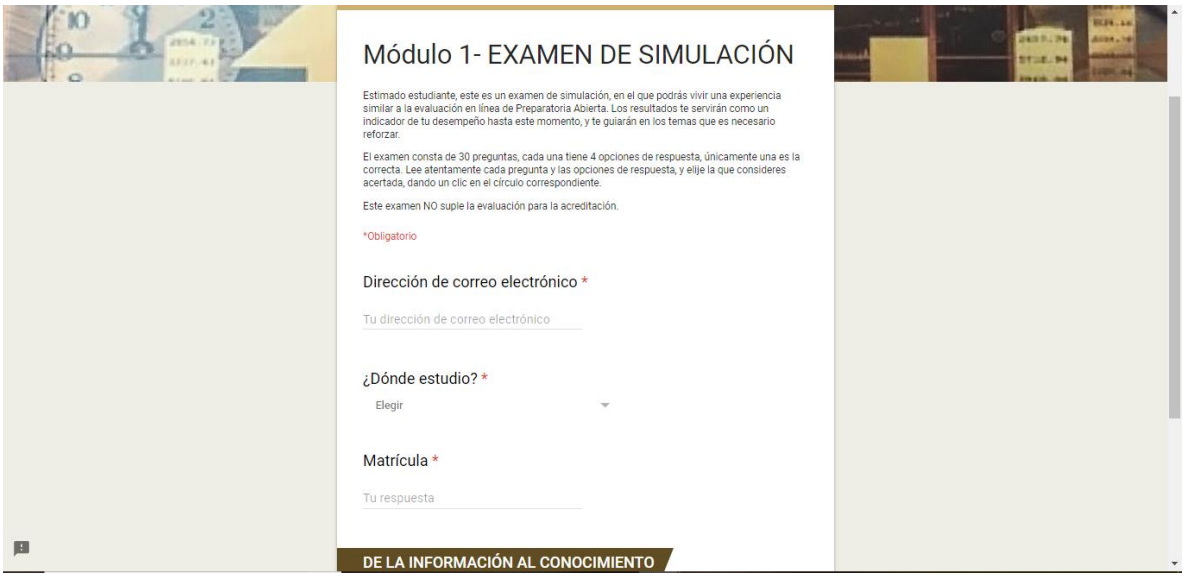
1. Ingresa a nuestra página de internet www.seg.guanajuato.gob.mx/prepaabierta.



2. Elige el examen de simulación correspondiente a este módulo.



3. Escribe tus datos de identificación: Nombre, matrícula, ¿Dónde estudio? en su caso grupo de estudio, Centro de Asesoría al que perteneces, o si eres estudiante independiente.



4. Comienza a contestar las preguntas, toma en cuenta los siguientes criterios:

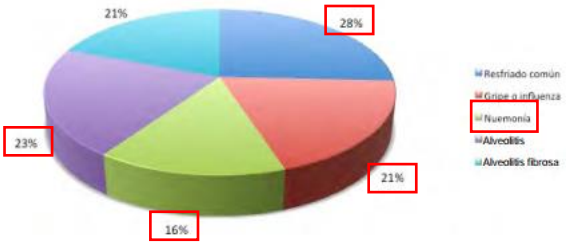
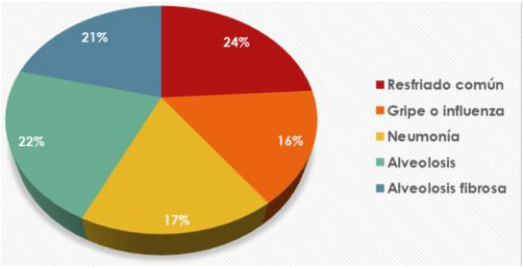
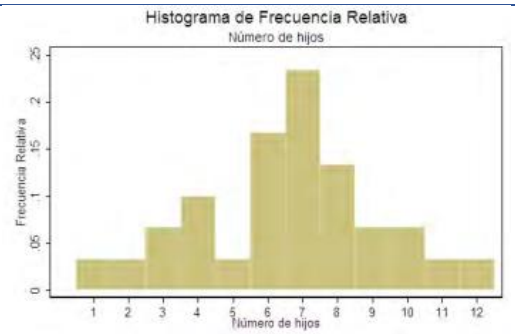
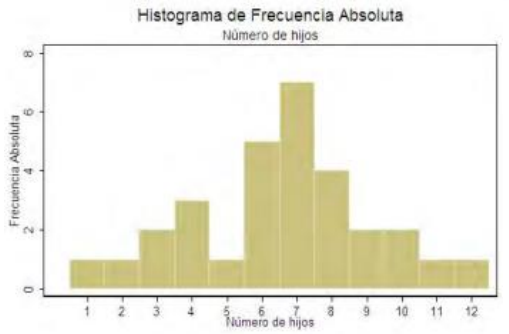
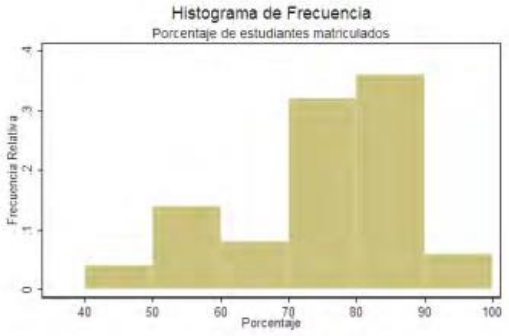
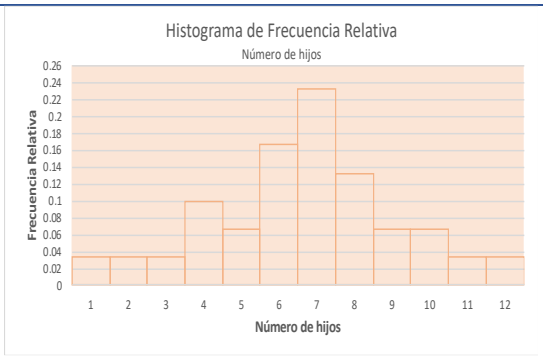
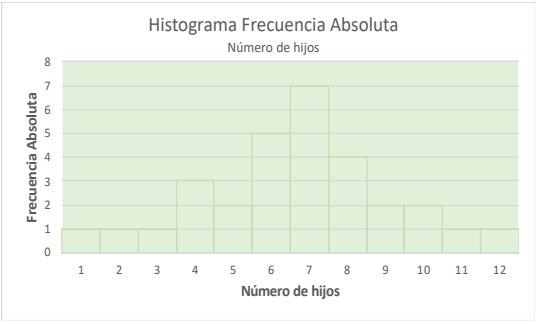
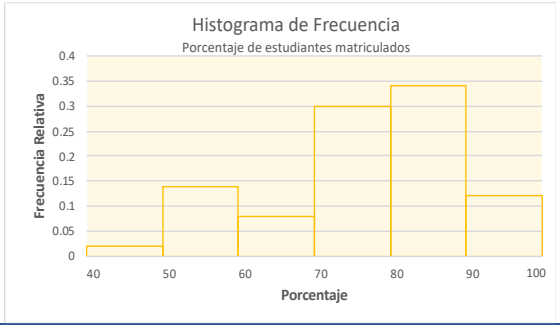
4.1. Lee con detenimiento cada una de las preguntas, sigue la guía ¿cómo son los exámenes? Que aparece en el Curso propedéutico de comunicación.

4.2 No dejes ninguna pregunta sin contestar

4.3. No hay límite de oportunidades para contestar el examen

5. Según tus resultados úbica los temas que necesitas revisar nuevamente antes de realizar el examen de acreditación.

ANEXO: CORRECCIONES DE EDICIÓN AL LIBRO

Referencia		Dice	Debe decir
Pág	Número/ inciso		
38	Tema: Asignación de la probabilidad a un evento aleatorio.	• Los porcentajes señalados NO son correctos.  • 36/100	•  • 36/150
62	Tema: Distribuciones de frecuencias para variables categóricas.	$P(x = \text{bachillerato}) = .110$ $P(x = \text{carrera técnica}) = .131$	Los resultados están cambiados $P(x = bachillerato) = 0.131$ $P(x = carrera\ técnica) = 0.110$
64 67		  	  

91	Tema: La estadística descriptiva y su aplicación en la explicación de los fenómenos naturales y los procesos sociales.	fecha límite de inscripción. Ante tal panorama, la Se cretaría de Salud decide revisar el funcionamiento de programa en los 49 municipios de la entidad. Lo datos sobre el porcentaje de registro por municipio s presentan en la siguiente tabla.	50 municipios
102	Cuadro	$\sum(x+\bar{x})^2 = 2.4643$	$\sum(x-\bar{x})^2 = 2.4643$
103	Cuadro	S^2 Varianza muestral σ^2 Varianza poblacional S^2 Desviación estándar muestral σ Desviación estándar poblacional	S Desviación estándar muestral
118	Tema: Distribución de probabilidad binomial.	$P(1) = P(x=1)$ $= \binom{5}{1}(.4)^1(.6)^4$ $(.6)^4 = .26$	$P(1) = P(x=1)$ $= \binom{5}{1}(0.4)^1(0.6)^4$ $= 0.26$
119	Ejemplo	$P(x) = \binom{6}{x}(.2)^x(.8)^{6-x}$	$P(x) = \binom{6}{x}(0.2)^x(0.8)^{6-x}$
120	Tema: Distribución de probabilidad de Poisson.	$p(2) = \frac{(3)^2 e^{-3}}{2!}$ $= \frac{(9)(0.4979)}{2}$ $= .22$	$p(2) = \frac{(3)^2 e^{-3}}{2!}$ $= \frac{(9)(0.04979)}{2}$ $= 0.22$

Fuentes consultadas

Cueva, F. (2012). Estadística en fenómenos naturales y proceso sociales. México: SEP.

Figura 1 http://horizontal.mx/la-tombola-y-su-logica-politica-2/

Figura 2 http://matepedia-estadistica.blogspot.mx/2016/05/grafica-de-barras-con-frecuencia_25.html

Figura 3 http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2002000400005

Figura 4 https://es.wikipedia.org/wiki/Distribuci%C3%B3n_normal

seg.guanajuato.gob.mx

GTO
Grandeza de México

Secretaría
de Educación

PREPARATORIA ABIERTA
GUANAJUATO

