Introducción a la estadística

Adelfa Patricia Colón García

Contenido

Introducción a la estadística

La palabra "estadística" se utiliza comúnmente para referirse a una colección de datos numéricos. Este es el significado más común de la palabra estadística. Sin embargo, la información numérica por si sola puede no constituir una estadística, para merecer este nombre, los datos deben constituir un conjunto coherente, establecido de forma sistemática y siguiendo criterios para su recolección, ordenación e interpretación. Esta clase les brindara las herramientas necesarias para poder desarrollar su pensamiento analítico y adentrarse dentro del mundo de la estadística, desde cómo recolectar, organizar, representar la información para cierto estudio hasta la toma de decisiones en base a las conclusiones respecto a los datos obtenidos de dicha información, es por esto que como verán a lo largo del mismo que las aplicaciones de la estadística son muy amplias, razón por la cual les brindará muchas herramientas para poder aplicar tanto en las clases siguientes como en proyectos de investigación y más aún es su quehacer como profesionales. Espero que las capacidades que están por adquirir en esta clase les permitan ser muy buenos profesionales y por sobre todo excelente personas, los animo a abordarlo con mucho entusiasmo y con una actitud positiva en el aprendizaje. ¿Según tu propia opinión porque consideras importante aprender sobre estadística? ¿Cómo consideras que la estadística podría ser de utilidad en tu futuro profesional?

Historia de la estadística

De acuerdo con Levin (2004,) el vocablo statistik proviene de la palabra italiana statista (que significa “estadista”). Fue utilizada por primera vez por Gottfried Achenwall (1719-1772), un profesor de Marlborough y de Göttingen. El Dr. E. A. W. Zimmerman introdujo el término statistics (estadística) a Inglaterra. Su uso fue popularizado por sir John Sinclair en su obra Statistical Account of Scotland 1791-1799 (“Informe estadístico sobre Escocia 1791-1799”). Sin embargo, mucho antes del siglo XVIII, la gente ya utilizaba y registraba datos, como puedes observar en el dibujo (antes de la era cristiana).

El Viejo Testamento contiene varios informes sobre levantamiento de censos. Los gobiernos de los antiguos Babilonia, Egipto y Roma reunieron registros detallados sobre población y recursos. En la Edad Media, los gobernantes empezaron a registrar la propiedad de la tierra.

En el año 762 de nuestra era, Carlomagno pidió una descripción detallada de las propiedades de la Iglesia. A principios del siglo IX terminó la enumeración estadística de los siervos que habitaban los feudos. Por el año 1806, Guillermo el Conquistador ordenó que se escribiera el Domesday Book, un registro de la propiedad, extensión y valor de las tierras de Inglaterra. Este trabajo fue el primer resumen estadístico de Inglaterra. (Levin 2004). Con estos datos lo único que se pretende es dar a conocer que la estadística es tan antigua como la vida misma, pues siempre hubo necesidad de contar, comenzando por las posesiones que la gente o los pueblos tenían.

Para que nos sirve la estadística

Introducción a la estadística

Quien tiene la información tiene el poder. Por esta razón los países invierten en generar conocimientos a través de investigaciones de las más diferentes disciplinas. Hoy en día la generación de información y su recopilación ha adquirido gran volumen y se requiere de instrumentos que sean capaces de procesarla en volumen y rapidez. La información siempre, y con mayor razón hoy en día, es importante para la toma de decisiones las que deben ser oportunas y óptimas. Con mala o insuficiente información posiblemente la decisión sea mala, por muy bueno que sea el procesamiento de ésta. Por el contrario, por muy buena que sea la información si el procesamiento es malo seguramente también la decisión sea equivocada. En consecuencia, un sólido respaldo para una acertada toma de decisiones contempla ambos aspectos: información buena y suficiente, procesamiento correcto.

Como estudiante de economía agrícola, requerirás conocimientos básicos y habilidad para organizar, analizar y transformar datos, así como para presentar la información. En esta clase aprenderás las técnicas y métodos estadísticos básicos que mejorarán su destreza para tomar buenas decisiones a nivel personal, en la investigación y en el campo ecónomo-administrativo. Esta primera parte de estadística es básica para avanzar a Estadística II que toca la parte inferencial de la estadística y para Econometría que sirve para realizar modelos económicos que te permitirán hacer predicciones con cierto grado de precisión. Las tres asignaturas son básicas para la investigación en el campo de la economía agrícola.

Subdivisiones de la estadística

Rustom, J., A. (2012). ESTADISTICA DESCRIPTIVA, PROBABILIDAD E INFERENCIA. Una visión conceptual y aplicada.

Muchos autores han dividido la estadística en dos ramas:

ESTADÍSTICA DESCRIPTIVA: Incluye las técnicas que tratan de la recolección, organización, representación y análisis de la información; con el propósito de describir los fenómenos mediante herramientas como la tabulación, representación en forma gráfica y el cálculo de medidas descriptivas. Esta comprende cualquier actividad para resumir o describir las variables en estudio, sin intentar nada que vaya más allá de los datos. En este curso de métodos estadísticos I se aborda solamente este aspecto de la estadística.

ESTADISTICA INFERENCIAL: Se deriva de las observaciones hechas solo a una parte de un conjunto numeroso de elementos; implicando así que su análisis requiera de generalizaciones que van más allá de los datos, o sea que la estadística inferencial investiga, analiza y saca conclusiones sobre una población partiendo de una muestra sacada de la misma. Algunas de sus herramientas son estimaciones, intervalos y pruebas de hipótesis.

Los resultados obtenidos pueden ser utilizados para otra rama de la estadística conocida como **Teoría de decisiones**. (Levin&Rubin, 2004)

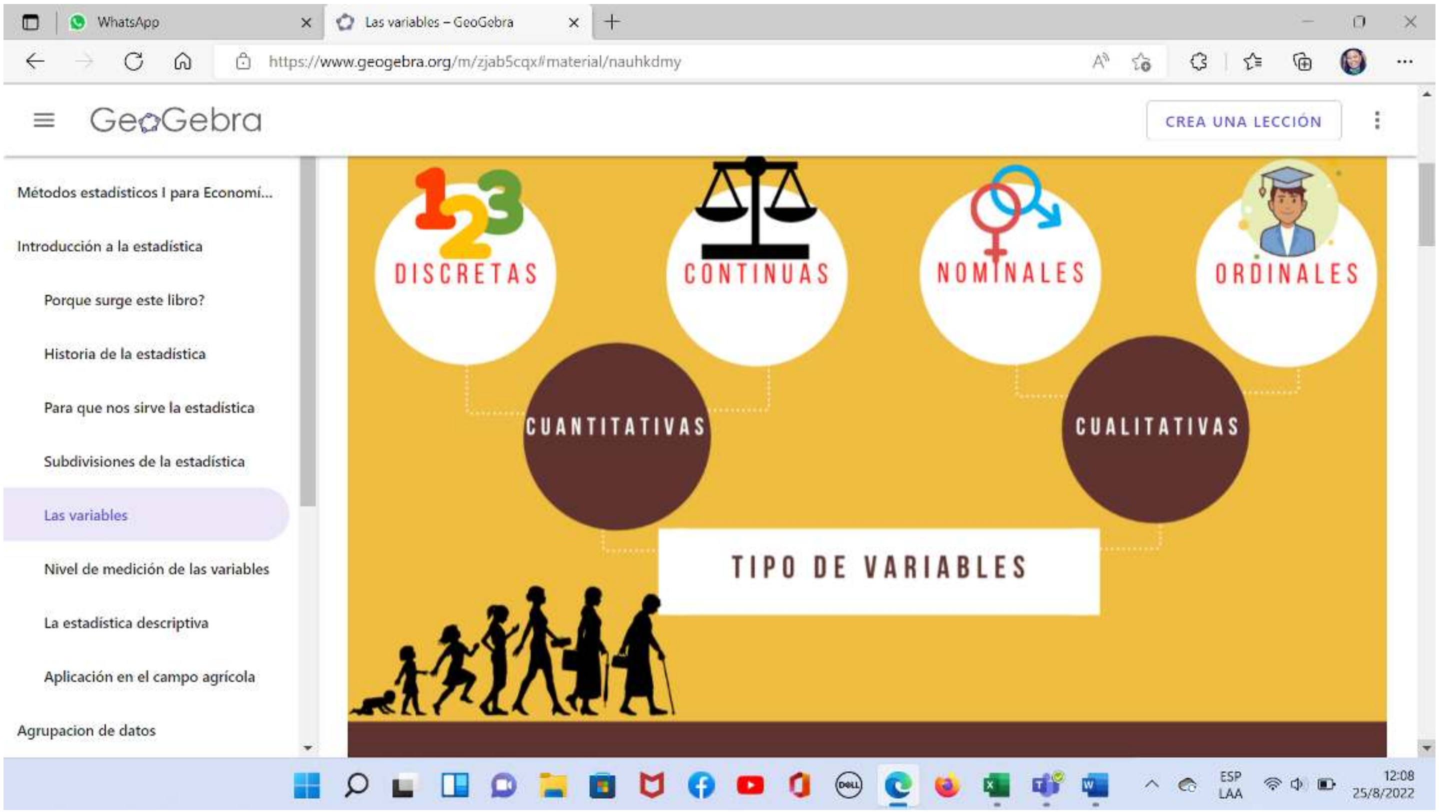
En muchos de los ejercicios de la clase te tocará tomar una decisión en función de los resultados obtenidos. Es importante que desde el inicio vayas practicando la toma de decisiones.

CAPITULO I

Las variables

Las observaciones o mediciones (x1, x2, x3, x4, ... etc.) sobre los elementos de una población o de una muestra constituyen la materia prima con la cual se trabaja en Estadística. Para que dichas observaciones puedan ser tratadas estadísticamente deben estar expresadas o poder ser expresadas en términos numéricos.

Aunque sea obvio, se destaca que la característica de interés a observar o medir en cada elemento de la población debe ser la misma, en tanto que se espera que no asuma el mismo valor en cada uno de los elementos que la conforman. Por ejemplo, una característica de interés cada vez que se realizan investigaciones es la edad de las personas que responden a nuestro cuestionario. Aquellas características como la edad que van cambiando en su estado o expresión entre los elementos de la población se denominan "variables", mientras que aquellas que no cumplen esta condición son llamadas "constantes".



Por lo tanto, las variables son los elementos que cambian de un individuo a otro. Haz el ejercicio en tu salón de clases. Compara tu edad con la de tus compañeros. La edad es una variable porque va cambiando de un compañero a otro. A continuación, conoceremos los diferentes tipos de variables y sus ejemplos:

Variables cualitativas o categóricas:

Un ejemplo de variables cualitativas o categóricas son las preguntas sobre el sexo (masculino o femenino), el estado civil (casado, soltero, viudo, divorciado o en unión libre) y la carrera universitaria (economía agrícola, agronomía, forestal, etc). Se les llama también variables **nominales** y sus categorías no implican ningún orden.

Otros ejemplos:

• Religión: católica, evangélica, ..., etc.

• Color de ojos: café, verdes, azules

También podemos encontrar variables cualitativas que son **ordinales** que Implican un orden en sus categorías, por ejemplo:

Nivel de escolaridad: primaria, secundaria, universitaria

Nivel de satisfacción: mal, regular, bueno, muy bueno, excelente

Variables cuantitativas:

Son variables que se pueden contar (discretas) o se pueden medir (continuas).

Las **discretas** son las que solo toman valores enteros que surgen de un proceso de conteo, por ejemplo, el número de estudiantes en cada clase, el número de asignaturas en cada plan de estudios, el número de gallinas en un corral, el número de vacas en un potrero. Mientras que las **continuas** son las que pueden tomar cualquier valor real que surgen de un proceso de medición por ejemplo la estatura, el peso y la talla de los estudiantes requieren de un aparato de medición (balanza, metro, etc.)

En general, cualquiera sea el tipo de la variable a resumir, existen tres formas de representarlas:

1° Por medio de **tablas de frecuencias**, que corresponde a una tabla formada por columnas, donde en la primera columna se anotan los diferentes valores de la variable (clases o categorías) y en las siguientes columnas los diversos tipos de frecuencia. Por frecuencia absoluta se entiende el número de individuos que pertenece a una misma clase.

2° Mediante **gráficos**, que son recursos pictóricos que permiten ilustrar mediante un dibujo lo que aparece en la tabla de frecuencias. Existen diversos tipos de gráficos y el uso de cada uno depende del tipo de variable a representar.

3° Con **estadígrafos o parámetros,** según las medidas resúmenes se traten de una población o una muestra, y que sirven para mostrar, medidas de posición, o medidas de dispersión el grado de concentración de estos.

En el siguiente tema veremos cada una de estas formas de representación.

Clasificar las siguientes variables en continuas o discretas:

a) Número de semillas de maíz por metro de surco sembrado.

b) Temperaturas registradas en un laboratorio, durante una semana.

c) Período de tiempo desde el almacenamiento y hasta que se produce el deterioro del 50% de los frutos almacenados.

d) Milímetros de precipitación de una localidad durante un año.

e) Número de semillas defectuosas en cajas de 50 semillas.

f) Número de materias aprobadas con más de 90% por estudiantes de Economía Agrícola en el 2021

g) Cociente entre el largo y el ancho de los entrenudos de plantas de maíz.

Nivel de medición de las variables

Los datos se clasifican por niveles de medición. El nivel de medición de los datos es importante para saber qué tipo de cálculos se pueden realizar para resumir y presentar los datos. También determina que pruebas estadísticas se pueden realizar.

Datos de nivel nominal En el caso del nivel nominal de medición, las observaciones acerca de una variable cualitativa sólo se clasifican y se cuentan. No existe una forma particular para ordenar las etiquetas. El género representa un ejemplo del nivel nominal de medición. Suponga que hace un conteo de los estudiantes que cursan esta asignatura e informa cuántos son hombres y cuántas mujeres. Podría presentar primero a los hombres o a las mujeres. Para el nivel nominal, la medición consiste en contar. A veces, para una mejor comprensión de lectura, estos conteos se convierten en porcentajes.

Ejemplo: en tu salón de clases anota:

Sexo	Frecuencia (cantidad)	Porcentaje (Cantidad/total) *100
Femenino		
Masculino		

Datos de nivel ordinal El nivel inmediato superior de datos es el nivel ordinal. Cada estudiante de la clase respondió la pregunta: "En términos generales, ¿cómo calificas al profesor del curso?" La calificación variable ilustra el uso de la escala ordinal de medición. Una calificación es más alta o mejor, que la siguiente: superior es mejor que bueno, bueno es mejor que promedio, etc.

Sin embargo, no es posible distinguir la magnitud de las diferencias entre los grupos. ¿La diferencia entre superior y bueno es la misma que entre malo e inferior? No es posible afirmarlo. Si sustituye 5 por superior y 4 por bueno, concluirá que la calificación superior es mejor que la calificación bueno, pero si suma una calificación de superior y una de bueno no espere que el resultado tenga significado.

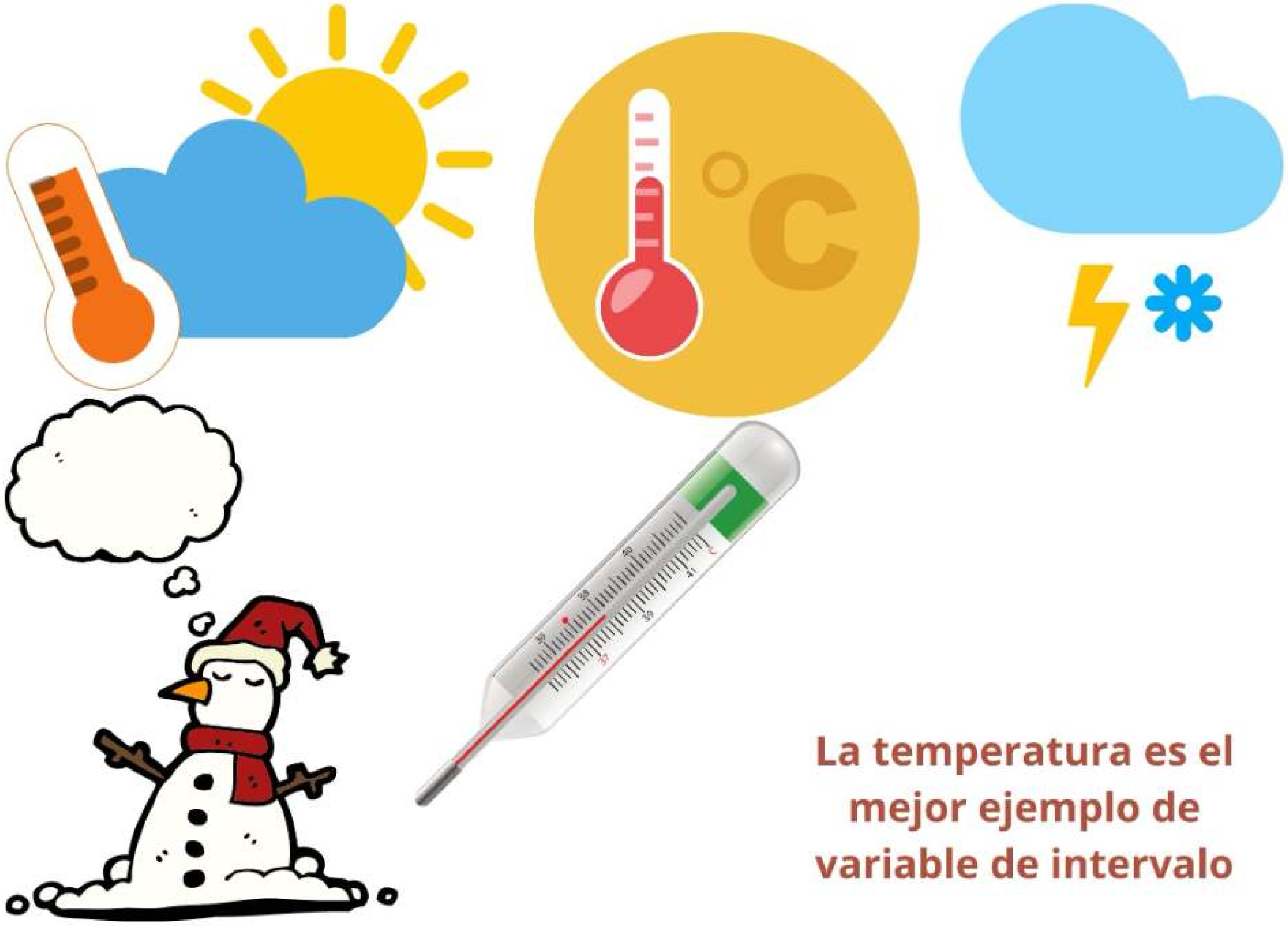
Además, no debe concluir que la calificación de bueno (calificación de 4) sea necesariamente dos veces más alta que malo (calificación de 2). Sólo tendrá claro que la calificación bueno es mejor que la calificación mala, no en qué grado es mejor.

Datos de nivel de intervalo

El nivel de intervalo de medición es el nivel inmediato superior. Incluye todas las características del nivel ordinal, pero, además, la diferencia entre valores constituye una magnitud constante. Un ejemplo de nivel de intervalo de medición es la temperatura. Suponga que las temperaturas altas durante tres días consecutivos de invierno en Boston son de 28, 31 y 20 grados Fahrenheit.

Estas temperaturas se clasifican fácilmente, aunque, además, es posible determinar la diferencia entre ellas, gracias a que un grado Fahrenheit representa una unidad de medición constante. Diferencias iguales entre dos temperaturas son las mismas, sin importar su posición en la escala. Es decir, la diferencia entre 10 y 15 grados Fahrenheit es de 5; la diferencia entre 50 y 55 grados también es de 5.

Es importante destacar que 0 es un punto más en la escala. No representa la ausencia de estado. Cero grados Fahrenheit no representa la ausencia de calor, sino sencillamente el hecho de que hace frío. De hecho, 0 grados Fahrenheit equivale aproximadamente a −18 grados en la escala Celsius. Por tanto, el cero es arbitrario.



Datos de nivel de razón

El nivel de razón es el más alto. Posee todas las características del nivel de intervalo, aunque, además, el punto 0 tiene sentido y la razón entre dos números es significativa. Ejemplos de la escala de razón de medición incluyen salarios, unidades de producción, peso, cambios en los precios de las acciones, la distancia entre sucursales y la altura. El dinero ilustra bien el caso. Si tiene cero dólares, entonces no tiene dinero. El peso constituye otro ejemplo. Si el cuadrante de la escala de un dispositivo correctamente calibrado se ubica en 0, entonces hay una ausencia total de peso. La razón entre dos números también resulta significativa. Si Jorge gana L400 000 anuales vendiendo seguros y María gana L800 000 al año en el negocio de los automóviles, entonces María gana el doble de lo que gana Jorge.

Introducción a la estadística

Fuente:

https://cape.fcfm.buap.mx/jdzf/cursos/est1/libros/book1e1.pdf

La estadística descriptiva

El registro de observaciones es una práctica común cuando se realizan investigaciones. Estas observaciones surgen durante un proceso de observación bajo condiciones dadas o de un proceso experimental. Generalmente la información registrada en un proceso de observación es tratada, en un primer momento, con el objetivo de describir y resumir sus características más sobresalientes. Esto se conoce como **estadística descriptiva** y generalmente se basa en el uso de tablas y gráficos, y en la obtención de medidas de resumen. El objetivo de este capítulo es reconocer la población y las variables relevantes en un proceso de investigación, caracterizar y describir muestras de las poblaciones mediante **medidas resumen**, **tablas de frecuencias** y **representaciones gráficas** y conocer algunas metodologías de extracción de muestras. Antes de abordar el problema de describir un conjunto de observaciones se verán algunos conceptos básicos que permiten la introducción de los procedimientos estadísticos.

Aplicación en el campo agrícola

Brasil es considerado líder mundial en producción de granos. Actualmente es el segundo mayor productor de soja del mundo, solo superado por Estados Unidos, y el tercer mayor productor de maíz. Una base de datos recopila información sobre la producción de maíz, soja, sorgo, trigo, habas, frijoles, avena, centeno, cebada y guisantes de cada condado de Brasil desde 1974. La información del área se expresa en hectáreas, la producción en toneladas métricas y el rendimiento en toneladas / ha. ¿De acuerdo con la situación planteada, cuantas variables identifica en este enunciado y de qué tipo son estas variables?

Capitulo II

Agrupación de datos

Datos sin procesar

Cuando se realizan investigaciones obtenemos los datos de manera desordenada, así como son recolectados en el campo de estudio. Estos datos revelan muy poco y es muy difícil determinar el significado de estos datos. Por esta razón es necesario ordenarlos de manera que nos permitan obtener información.

La información obtenida, antes de ser organizada y analizada se conoce como datos sin procesar. Aun no se les ha aplicado ningún tratamiento estadístico. Los datos sin procesar son datos informáticos o información resultado de encuestas no procesados. Esta información puede almacenarse en un archivo, o simplemente puede ser una colección de números y caracteres almacenados en papel o en algún lugar del disco duro de la computadora. Por ejemplo, la información ingresada en una base de datos a través de los cuestionarios de Google forms y que se reflejan en Excel a menudo se denomina datos sin procesar. **Los datos pueden ser ingresados por un usuario o generados por la propia computadora.** Debido a que no ha sido procesado por la computadora de ninguna manera, se considera "datos en bruto". Los siguientes datos corresponden a resultados parciales sin procesar de una encuesta en línea aplicada a estudiantes de la carrera de Economía Agrícola en julio del 2022, con relación a su índice académico. (Se tomaron solamente 11 al azar)

69 70 70 74 79 76 82 75 81 80 69

Ordenamiento de los datos

Una serie de datos ordenada nos presentara las observaciones de mayor a menor o viceversa. La serie ordenada nos da alguna idea de la tendencia de los datos, comenzando por conocer sus valores extremos y la distancia entre ellos.

Para que los datos sean útiles, necesitamos organizar nuestras observaciones de modo que podamos distinguir patrones y llegar a conclusiones lógicas.

Continuando con el último ejemplo, el primer tratamiento estadístico que le haremos a los datos será colocarlos en orden.

Notas de los 11 estudiantes de presentados en orden ascendente.

69 69 70 70 74 75 76 79 80 81 82

¿Qué podemos observar en estos datos que no era tan evidente en el listado anterior?

1. Podemos ver el valor mínimo: (Nota menor es 69)

2. Podemos ver el valor máximo: (Nota mayor es 82)

3. Es fácil dividir los datos en secciones: Vemos que la mitad de los números están por debajo de 75 y la otra mitad por encima.

4. Podemos ver la distancia entre número sucesivos de valores.

5. Podemos notar fácilmente que hay valores que se repiten (69 y 70)

Distribución de frecuencias

Una distribución de frecuencias o tabla de frecuencias ordenará los datos si estos se dividen en clases y se registra el número de observaciones en cada clase. Cada clase tiene un límite inferior y un límite superior. El número de clases es arbitrario, pero se recomiendan entre 5 y 20 clases.

Representación gráfica de distribuciones de frecuencia

La organización de los datos constituye la primera etapa de su tratamiento, puesto que facilita los cálculos posteriores.

Para ir analizando los datos podemos organizarlos en una tabla de frecuencias, o sea una tabla que divide los valores en grupos similares llamados clases llamada distribución de frecuencias. Es una agrupación de datos en clases mutuamente excluyentes que indican el número de observaciones en cada clase(frecuencias).

Distribución de frecuencias para variable cualitativa

La siguiente tabla resume la información obtenida acerca del color de ojos de 20 personas

Color de ojos	F
Negros	6
Café	5
Azules	2
Verdes	7

Distribución de frecuencias para datos no agrupados Se le pregunto a un grupo de 10 individuos el número de personas que dependen económicamente de ellos:

Distribución de frecuencias datos agrupados La siguiente tabla resume las calificaciones obtenidas por un grupo de 36 alumnos.

calificaciones	# de alumnos
54-61	3
62-69	3
70-77	4
78-85	10
86-93	8
94-100	8
Total	36

Elementos de una distribución de frecuencias: Clases, frecuencias absolutas (fi), frecuencia relativa (fr), frecuencia relativa porcentual (fr%), frecuencia acumulada (Fa), frecuencia des acumulada (Fd) Además, siempre que la distribución de frecuencias sea para datos agrupados se debe incluir Limites reales y Marcas de clase.

Clases o intervalos de clase: Constituyen categorías que permiten agrupar los valores de las variables, se suele denotar esta columna por medio del nombre de la variable estudiada. Para cada clase se debe cumplir que

1. Deben ser del mismo tamaño
2. No deben traslaparse

Para escoger la cantidad de clases se siguen diversos criterios. Uno de ellos es la cantidad de datos y el rango o sea la diferencia entre el valor máximo y el valor mínimo de un conjunto de datos.

Algunas formulas

Número de intervalos. Puede establecerse mediante la fórmula: $k=1+3.322*\log(n)$

Donde n es el total de datos Ancho de clase: c donde c= rango/k •Con valores enteros LRI = LI - 0.5 LSR= LS+0.5

Marca de clase $X_i= (LI + LS)/2$

Explicación de los elementos de una distribución de frecuencias

Frecuencia absoluta (fi): O solamente frecuencia especifica cuantos datos del conjunto se encuentran en una clase determinada.

Frecuencia relativa (fr): Muestra la frecuencia de datos de cierta clase proporcionalmente a todo el conjunto, resulta de comparar la frecuencia de cada clase con el total de frecuencias. La suma de las frecuencias relativas deber ser igual a 1 salvo errores de aproximación por el redondeo de cifras. fr= fi/total de observaciones

Frecuencia relativa porcentual (fr%): Indica en porcentajes como es la frecuencia en cada clase $fr\%=fr\times100$

Frecuencia acumulada (Fa): Acumula las frecuencias para llevar un conteo con respecto a un límite superior. Se obtienen sumando cada frecuencia con las frecuencias de las categorías anteriores En la tabla de distribución de frecuencias la columna de Fa siempre inicia con la primera frecuencia absoluta y siempre termina con el total de datos.

Frecuencia des acumulada (Fd): Des acumula o resta del total de datos las frecuencias para llevar un conteo con respecto a un límite inferior. Se obtienen restando del total de datos las frecuencias

En la tabla de distribución de frecuencias la columna de FD siempre debe iniciar con el total de datos y siempre debe terminar con la última frecuencia absoluta.

¿Cómo crear una distribución de frecuencias? El primer paso consiste en acomodar los datos en una tabla que muestre las clases y el número de observaciones que hay en cada clase. Los pasos para construir una distribución de frecuencias se entienden mejor con un ejemplo.

Paso 1: Defina el número de clases. El objetivo consiste en emplear suficientes agrupamientos o clases, de manera tal que se perciba la forma de la distribución. Aquí se necesita criterio. Una gran cantidad de clases o muy pocas podrían no permitir ver la conformación fundamental del conjunto de datos.

Paso 2: Determine el intervalo o ancho de clase. El intervalo o ancho de clase debería ser el mismo para todas las clases. Todas las clases juntas deben cubrir por lo menos la distancia del valor más bajo al más alto de los datos. Investigue la formula.

Paso 3: Establezca los límites de cada clase. Este paso es importante para que sea posible incluir cada observación en una sola categoría. Esto significa que debe evitar la superposición de límites de clase confusos.

Paso 4: Anote los elementos que pertenecen a cada clase.

Paso 5: Cuente el número de elementos de cada clase. El número de elementos que hay en cada clase recibe el nombre de frecuencia de clase.

Ejercicio: continuando con los datos del ejercicio de los 100 estudiantes

$k=1+3.322*\log(n) = 1+3.322(2) = 1+ 6.644 = 7.644$ redondeamos 8 clases

Rango es 96 – 49 = 47, por tanto 47/8= 5.875 redondeamos a 6. Cada intervalo será de 6 (Esto aplica solo si trabajamos con números enteros)

Completa el cuadro

Clase	Frecuencia	Fr	Fr%	Fa	Fd
47 a 53	2				
54 a 60	15				
61 a 67	22				
68 a 72	22				
73 a 79	24				
80 a 86	10				
87 a 93	3				
94 a 100	2				

La anterior es una tabla compuesta por 5 clases. La frecuencia en este caso es llamada frecuencia absoluta y nos indica cuantos valores hay en cada clase. Seguidamente agregaremos 2 indicadores más a nuestra tabla de frecuencias. La frecuencia relativa y la frecuencia acumulada. La frecuencia relativa es una medida estadística que se calcula como el cociente de la frecuencia absoluta de algún valor de la población/muestra (fi) entre el total de valores que componen la población/muestra (N). La frecuencia relativa se representa con las letras hi y su fórmula de cálculo es la siguiente:

Gráficos

Las gráficas proporcionan datos en un diagrama de dos dimensiones. El tipo de gráfico a utilizar dependerá del tipo de variable y de la información que queremos mostrar. en función del tipo de variable podemos utilizar los siguientes gráficos:

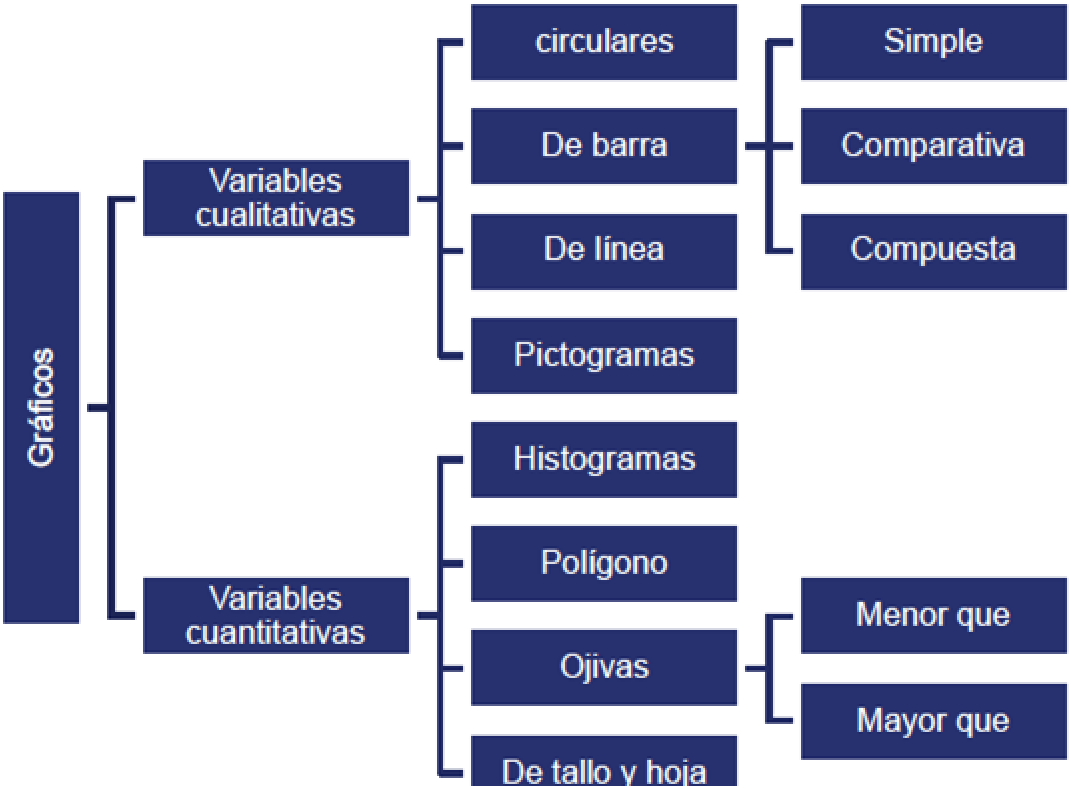
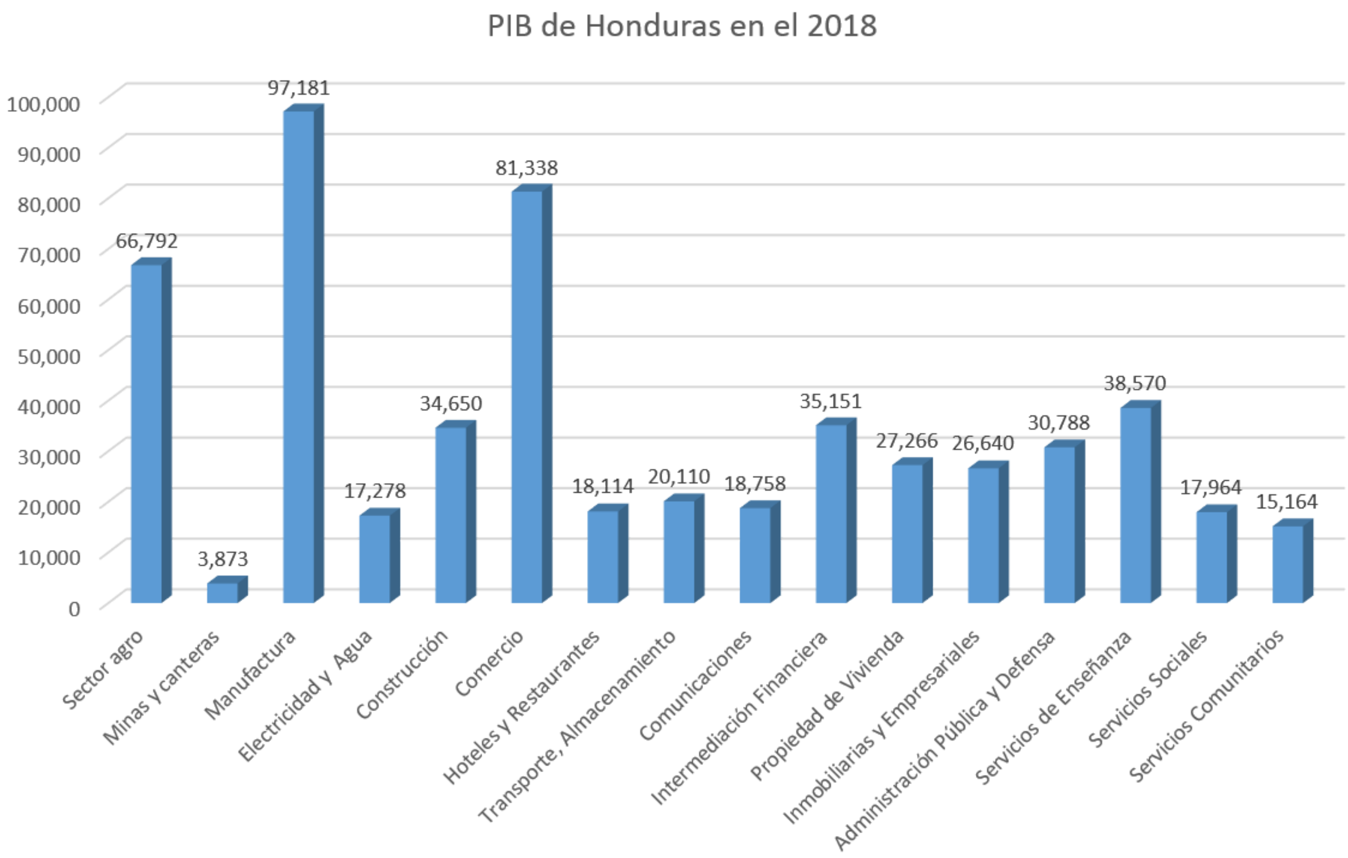
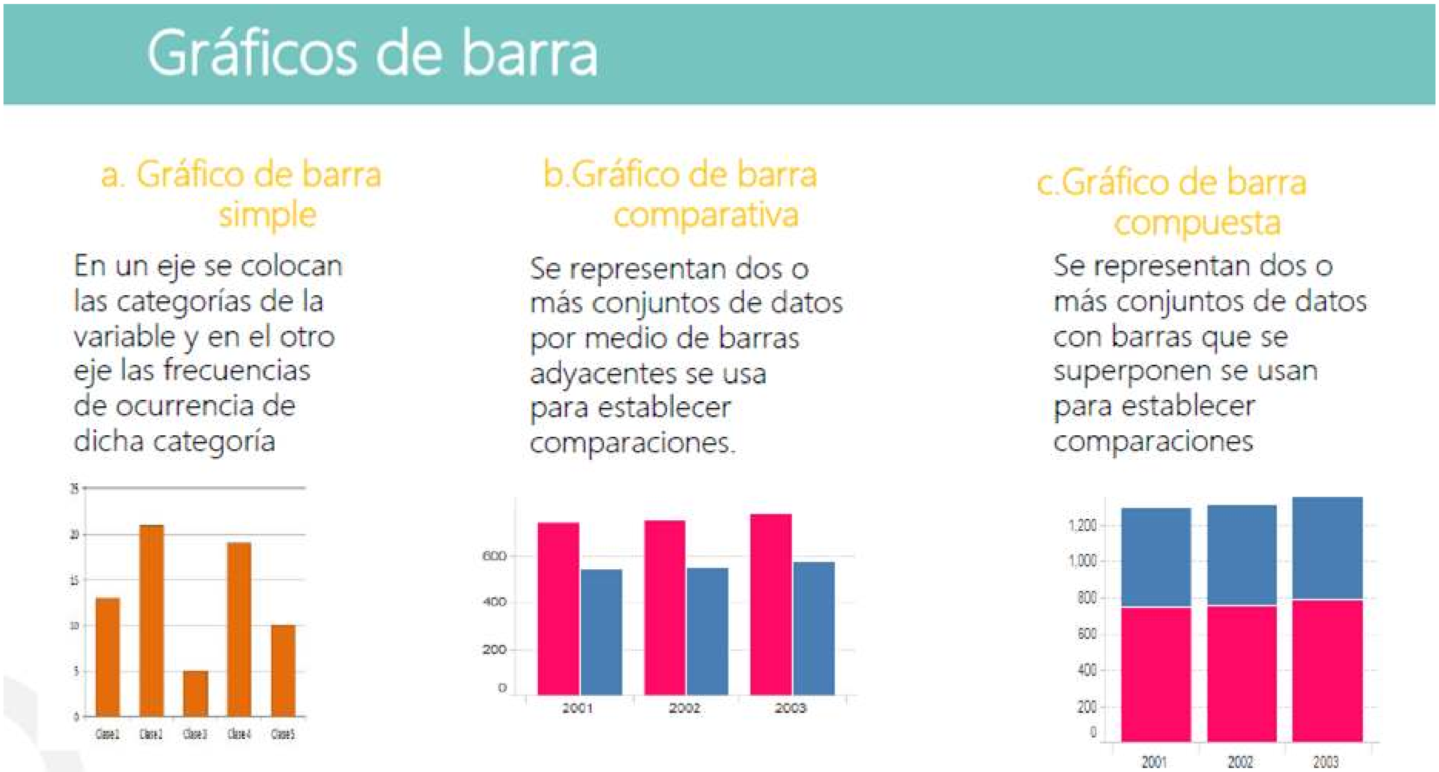


Diagrama de barras y grafica de pastel

PIB (Producto Interno Bruto) de Honduras en el 2018



Fuente: BCH



El grafico de barra reciben este nombre ya que la información de una tabla se ilustra utilizando barras que nos indican los valores correspondientes a las diferentes categorías; éstas pueden ser simples, comparativas y compuestas. El tipo de gráfico de barra a usar depende de cuál sea nuestro objetivo al analizar o comunicar la información.

Interprete la gráfica

¿Cuáles fueron los 3 sectores que más contribuyeron al PIB en el 2018?

Histograma

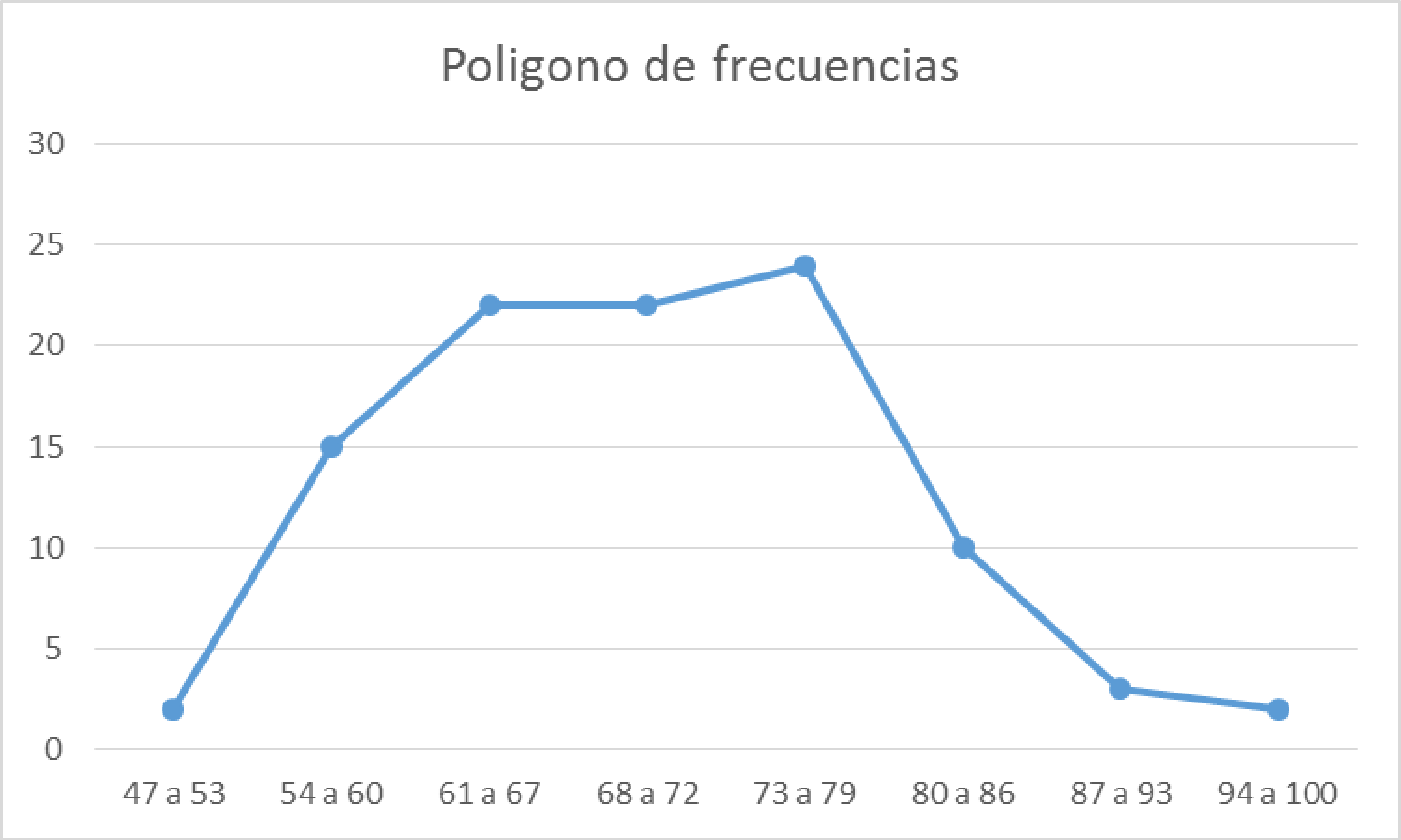
Un histograma consiste en una serie de rectángulos, cuyo ancho es proporcional al rango de los valores que se encuentran dentro de una clase, y cuya altura es proporcional al número de elementos que caen dentro de la clase. Si las clases empleadas en la distribución de frecuencias son del mismo ancho, entonces las barras verticales del histograma también tienen el mismo ancho. La altura de la barra correspondiente a cada clase representa el número de observaciones de la clase. Como consecuencia, el área contenida en cada rectángulo (base por altura) ocupa un porcentaje del área total de todos los rectángulos la cual es igual a la frecuencia absoluta de esa clase correspondiente respecto a todas las observaciones hechas.

Notas de los 100 estudiantes del CURLA

Un histograma que utiliza las frecuencias relativas de los datos puntuales de cada una de las clases, en lugar de usar el número real de puntos, se conoce como histograma de frecuencias relativas. Este tipo de histograma tiene la misma forma que un histograma de frecuencias absolutas construido a partir del mismo conjunto de datos. Esto es así debido a que, en ambos, el tamaño relativo de cada rectángulo es la frecuencia de esa clase comparada con el número total de observaciones. Recuerde que la frecuencia relativa de cualquier clase es el número de observaciones que entran en la clase, dividido entre el número total de observaciones hechas. La suma de todas las frecuencias relativas de cualquier conjunto de datos debe ser igual a 1. Poder presentar los datos en términos de la frecuencia relativa de las observaciones, más que en términos de la frecuencia absoluta, es de gran utilidad, ya que mientras los números absolutos pueden sufrir cambios, la relación entre las clases permanece estable. Resulta fácil comparar los datos de muestras de diferentes tamaños cuando utilizamos histogramas de frecuencias relativas.

Polígono de frecuencias

Aunque se utilizan menos, los polígonos de frecuencias son otra forma de representar gráficamente distribuciones tanto de frecuencias como de frecuencias relativas. Para construir un polígono de frecuencias señalamos éstas en el eje vertical y los valores de la variable que estamos midiendo en el eje horizontal, del mismo modo en que se hizo con el histograma. A continuación, graficamos cada frecuencia de clase trazando un punto sobre su punto medio y conectamos los puntos sucesivos resultantes con una línea recta para formar un polígono (una figura con muchos lados).



¿De qué manera podemos convertir un polígono de frecuencias en un histograma? Un polígono de frecuencias es sólo una línea que conecta los puntos medios de todas las barras de un histograma. Por consiguiente, podemos reproducir el histograma mediante el trazado de líneas verticales desde los límites de clase (señalados en el eje horizontal) y, luego, conectando esas líneas con rectas horizontales a la altura de los puntos medios del polígono.

Los histogramas y los polígonos de frecuencias son similares. ¿Por qué necesitamos ambos? Las ventajas de los histogramas son:

1. Los rectángulos muestran cada clase de la distribución por separado.

2. El área de cada rectángulo, en relación con el resto, muestra la proporción del número total de observaciones que se encuentran en esa clase.

Ojivas

Una distribución de frecuencias acumuladas nos permite ver cuántas observaciones están por encima de ciertos valores, en lugar de hacer un mero registro del número de elementos que hay dentro de los intervalos. La gráfica de una distribución de frecuencias acumuladas se conoce como ojiva.

La siguiente gráfica refleja el índice global de 98 estudiantes de economía agrícola utilizando la frecuencia absoluta acumulada.

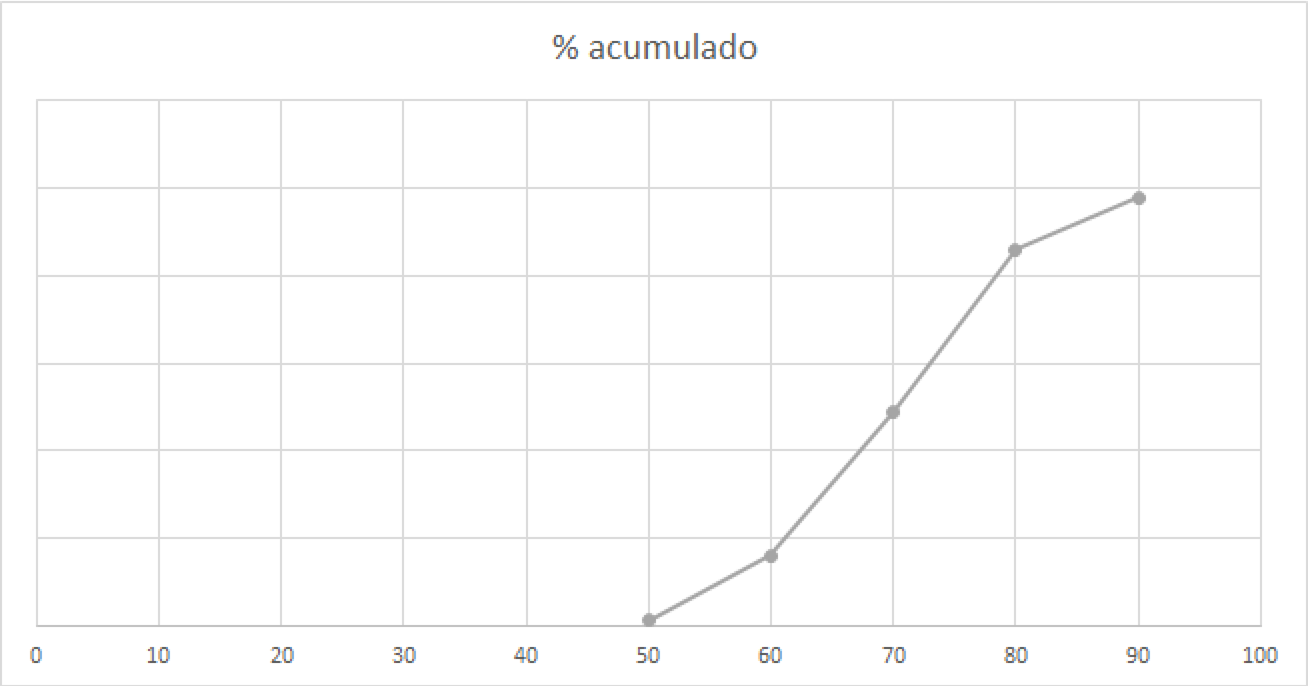


Diagrama de puntos

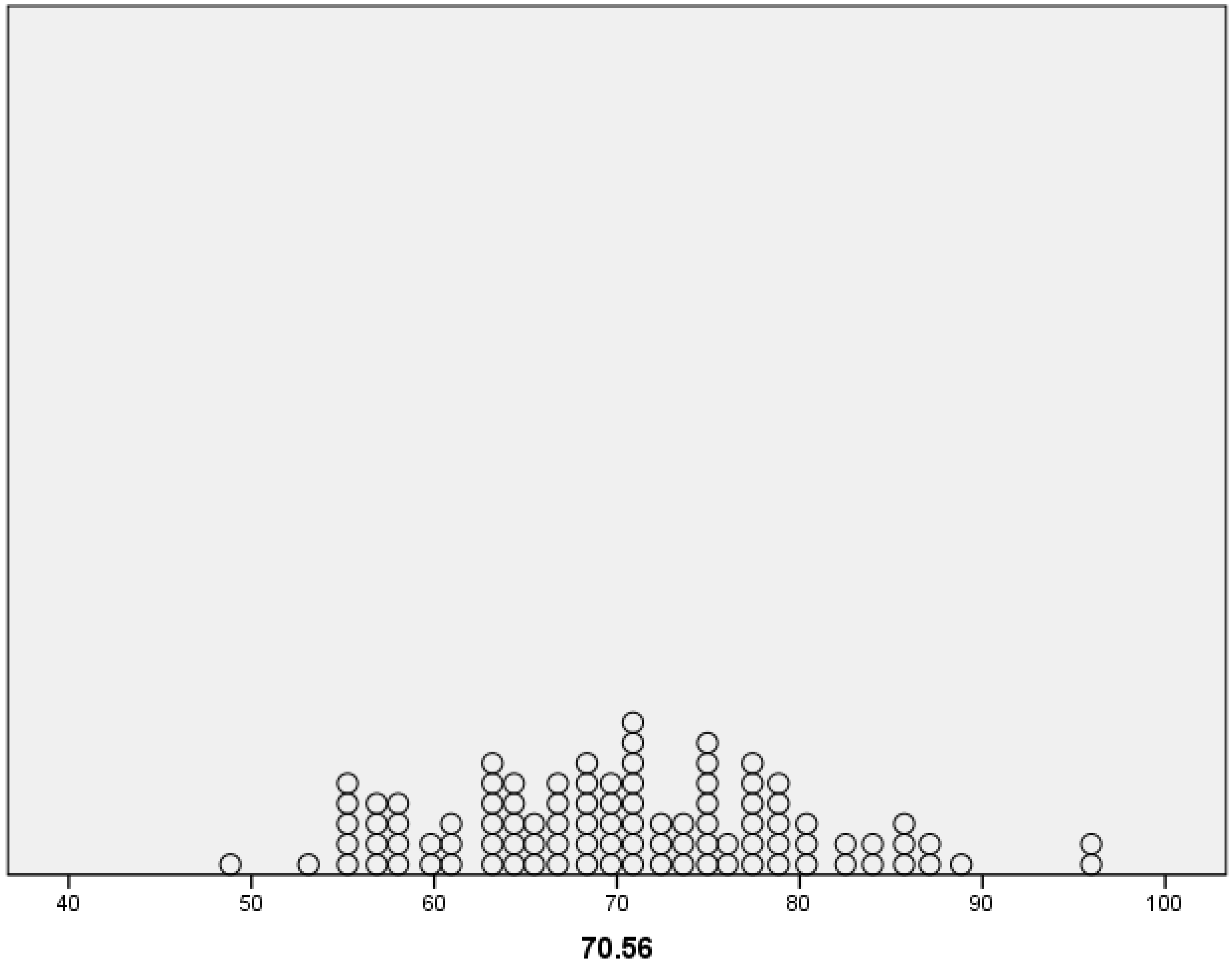
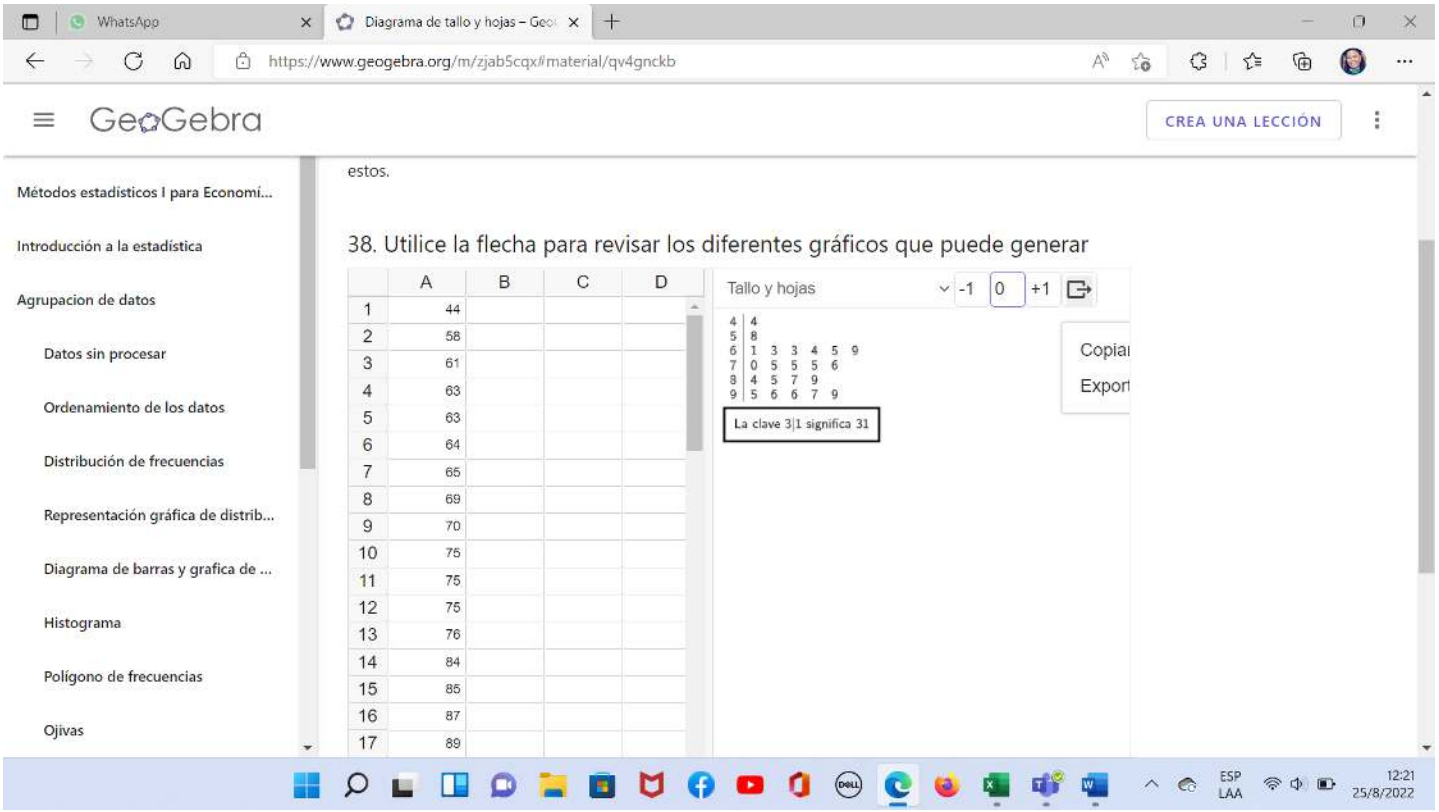


Diagrama de tallo y hojas

Una forma muy adecuada de organizar un número moderado de datos individuales consiste en dividir cada dato en dos partes, su tallo y su hoja. Si por ejemplo el conjunto de datos son números de dos dígitos, ya sea decenas y unidades o entero y decimal, entonces las decenas o el entero es el tallo y las unidades o el decimal es la hoja. El diagrama "tallo y hojas" permite obtener simultáneamente una distribución de frecuencias de la variable y su representación gráfica. Esta representación de los datos es semejante a la de un histograma, pero además de ser fáciles de elaborar, presentan más información que estos.


WhatsApp
Diagrama de tallo y hojas – Geo
https://www.geogebra.org/m/zjab5cqx#material/qv4gnckb
GeoGebra
CREA UNA LECCIÓN
Métodos estadisticos I para Economí...
Introducción a la estadística
Agrupacion de datos
Datos sin procesar
Ordenamiento de los datos
Distribución de frecuencias
Representación gráfica de distrib...
Diagrama de barras y grafica de ...
Histograma
Polígono de frecuencias
Ojivas
estos.
38. Utilice la flecha para revisar los diferentes gráficos que puede generar
A B C D
1 44
2 58
3 61
4 63
5 63
6 64
7 65
8 69
9 70
10 75
11 75
12 75
13 76
14 84
15 85
16 87
17 89
Tallo y hojas
-1 0 +1
4 | 4
5 | 8
6 | 1 3 3 4 5 9
7 | 0 5 5 5 6
8 | 4 5 7 9
9 | 5 6 6 7 9
La clave 3|1 significa 31
Copiar
Export
ESP
LAA
12:21
25/8/2022

CAPITULO III

Medidas descriptivas

Estas medidas sirven para describir los datos o información de interés en un estudio, permitiendo hacer una especie de retrato de la información, ya que nos proporcionan aspectos esenciales de la misma, sin tener que ver uno a uno los datos.

Medidas de posición y dispersión

https://youtu.be/1MY0IVs2-fk

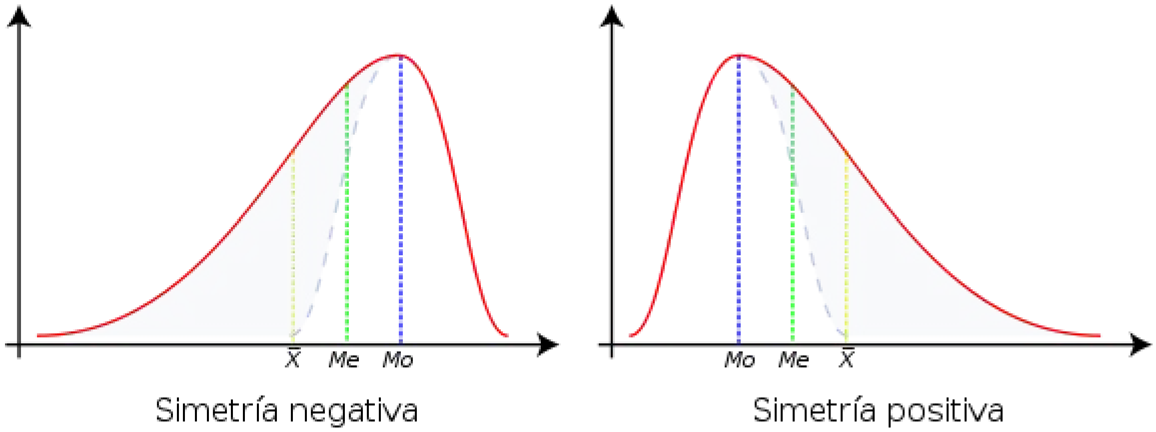
Introducción a la estadística

La estadística sumaria o descriptiva permite describir las características de un conjunto de datos - Las características de una población se denominan **Parámetros** - Las características de una muestra se denominan **estadísticos** - Resumen de la Información: Dos de estas características son de particular importancia para los responsables de tomar decisiones: la **tendencia central** y la **dispersión**.

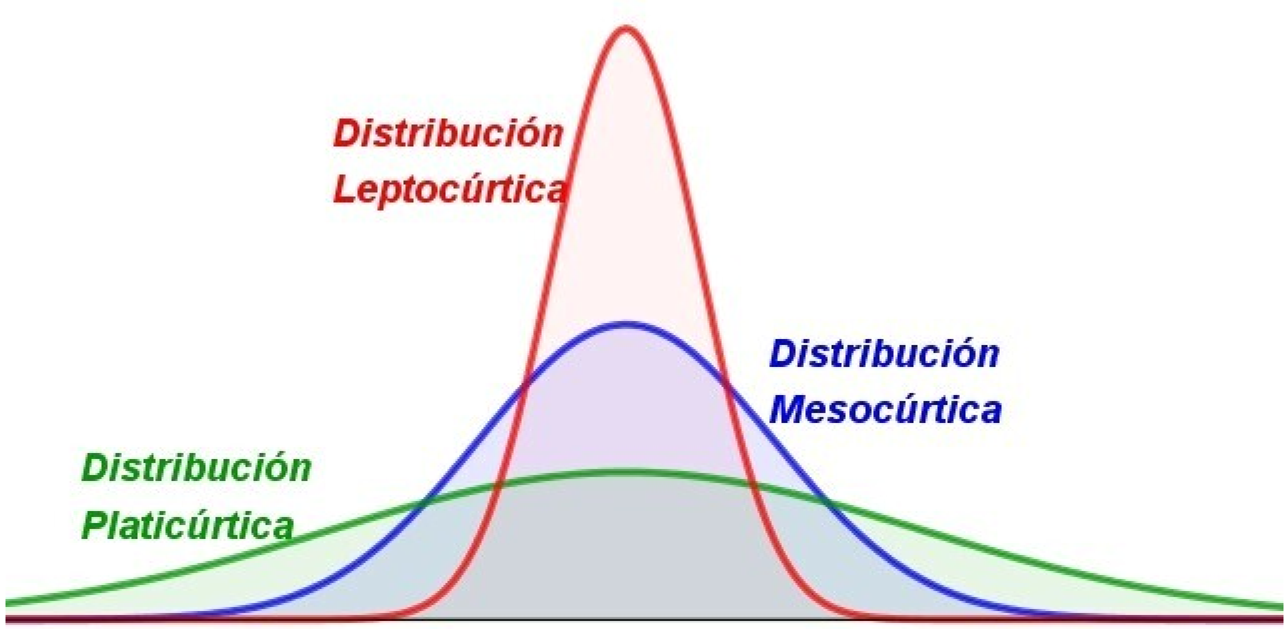
Tendencia central Las medidas de tendencia central, son medidas que describen un conjunto de datos, indicando cuales son los valores que tienden a ubicarse en la parte central de los mismos; y alrededor de ellas se agrupa la mayor parte de los valores de la información analizada. Las más usadas son: La media, La mediana y La moda.

Dispersión La dispersión se refiere a la separación de los datos en una distribución, es decir, al grado en que las observaciones se separan. Existen otras dos características de los conjuntos de datos que proporcionan información útil: el sesgo y la curtosis.

Sesgo Las curvas que representan los datos puntuales de un conjunto de datos pueden ser simétricas o sesgadas. Las curvas simétricas, tienen una forma tal que una línea vertical que pase por el punto más alto de la curva dividirá su área en dos partes iguales. Cada parte es una imagen de espejo de la otra.



https://sites.google.com/site/probyestacecytechig/_/rsrc/1317829385907/parcial-ii/medidas-de-forma/medidas-de-sesgo-o-asimetria/GrafMedCorr2.png **Curtosis** Cuando medimos la curtosis de una distribución, estamos midiendo qué tan puntiaguda es.



https://www.lifeder.com/wp-content/uploads/2020/03/curtosis.jpg

La curtosis (o apuntamiento) es una medida de forma que mide cuán escarpada o achatada está una curva o distribución. Este coeficiente indica la cantidad de datos que hay cercanos a la media, de manera que, a mayor grado de curtosis, más escarpada (o apuntada) será la forma de la curva. Las curvas se pueden clasificar en tres grupos según el signo de su curtosis, es decir, según la forma de la distribución:

- **Leptocúrtica**: la *Curtosis*>0. Los datos están muy concentrados en la media, siendo una curva muy apuntada.
- **Mesocúrtica:** la *Curtosis*=0. Distribución normal.
- **Platicúrtica:** la *Curtosis*<0. Muy poca concentración de datos en la media, presentando una forma muy achatada.

Medidas descriptivas de posición

La media aritmética

https://youtu.be/7fCHY4-neV8

$$\bar{X} = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_n}{N}$$

Es una medida de tendencia central que indica el valor promedio de un conjunto de datos, de allí que a veces nos referimos a la media como el promedio de los datos. Características • La media sólo puede utilizarse para datos cuantitativos y es única. • Tiene la desventaja que es afectada por valores extremos. La media aritmética (o media simplemente) de un conjunto de datos es la suma de los valores de los datos dividida por el número de observaciones. Si el conjunto de datos es toda la población de datos, la media poblacional es un parámetro que viene dado por donde N tamaño de la población.

71	67	65	73	96	77	58	70	64	70
76	74	87	63	58	64	58	71	61	67
66	71	66	75	69	80	70	73	63	82
79	55	75	55	75	71	61	60	58	65
72	72	78	70	57	70	85	96	57	79
75	61	87	71	79	69	78	86	64	84
59	65	81	63	86	71	64	89	64	77
78	67	67	69	81	49	78	56	84	55
73	74	75	79	56	57	68	63	55	70
77	83	65	53	71	68	69	76	74	69

$$\bar{X} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + x_4 f_4 + \cdots + x_n f_n}{N} = \frac{\sum_{i=1}^{n} x_i f_i}{N}$$

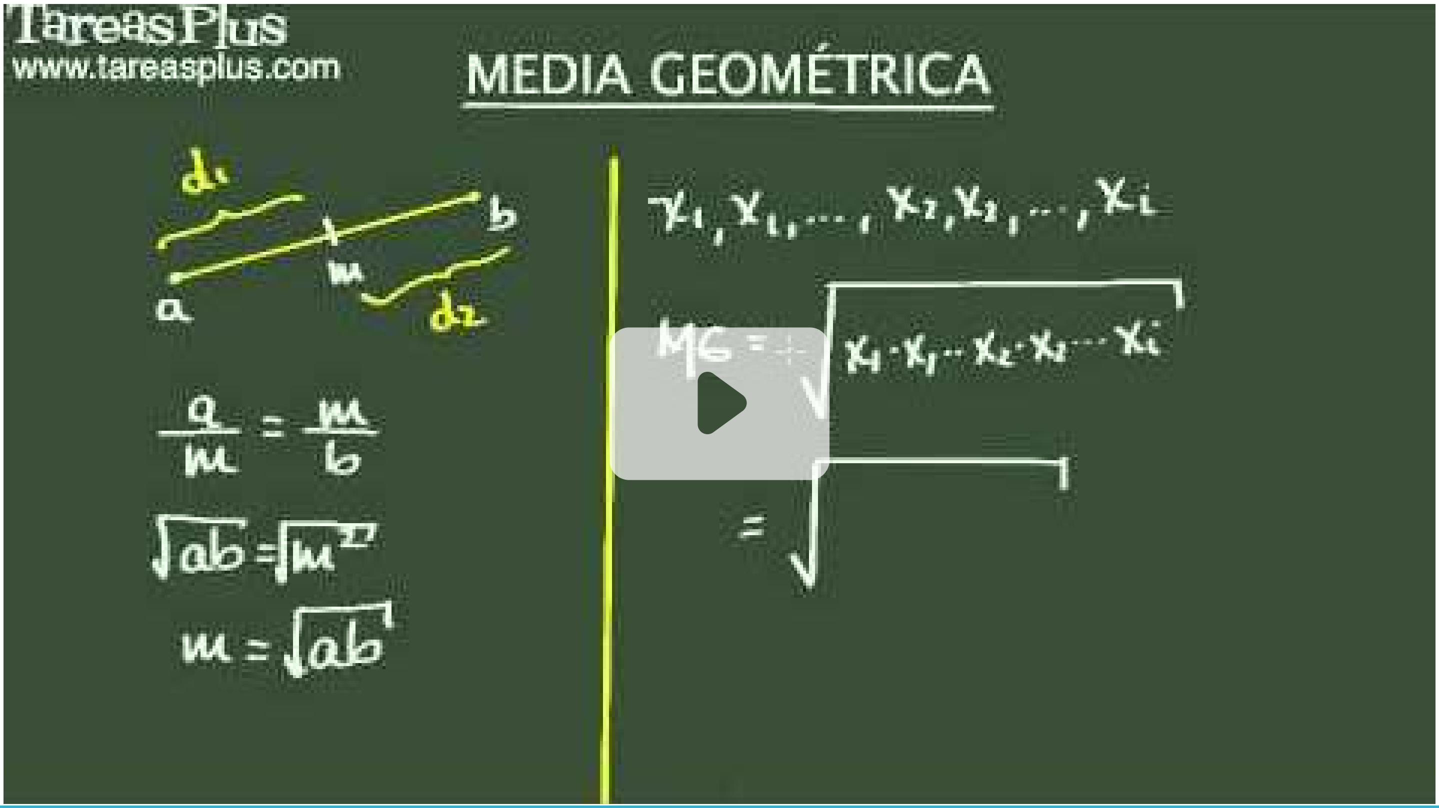
Si los datos están agrupados la media se calcula de la siguiente manera: Si el conjunto de datos procede de una muestra, entonces formula cambia N por n.

Media geométrica

Algunas veces, cuando trabajamos con cantidades que cambian en cierto periodo, necesitamos conocer una tasa promedio de cambio, como la tasa de crecimiento promedio en un periodo de varios años. En tales casos, la media aritmética simple resulta inapropiada, pues proporciona resultados equivocados. Lo que debemos encontrar es la media geométrica, llamada simplemente la M.G. Considere, por ejemplo, el crecimiento de una cuenta de ahorros. Suponga que inicialmente depositamos $100 y dejamos que acumule intereses a diferentes tasas durante cinco años. La fórmula para el cálculo es:

$$M \cdot G = \sqrt[n]{x_1 \times x_2 \ldots \times x_n}$$

Después de ver el video conteste la pregunta: Suponga que los porcentajes de inflación de los últimos 3 años han sido 18%, 15% y 12%. ¿Cuál es el porcentaje promedio de inflación en este periodo?



Media geométrica - YouTube

La media ponderada

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} = \frac{x_1 w_1 + x_2 w_2 + x_3 w_3 + \ldots + x_n w_n}{w_1 + w_2 + w_3 + \ldots + w_n}$$

La media ponderada nos permite calcular un promedio que toma en cuenta la importancia de cada valor con respecto al total. Esta es la forma como se calcula tu promedio de notas en la universidad. (“La media ponderada – GeoGebra”)

O lo que es lo mismo

$$\bar{x}_\omega = \frac{\sum(\omega \times x)}{\sum \omega}$$

Haz un ejercicio en Excel con las notas de tu primer periodo en la UNAH

La mediana

La mediana es una medida de tendencia central diferente a cualquiera de las que hemos tratado hasta ahora. La mediana es un solo valor del conjunto de datos que mide la observación central del conjunto de datos ordenados. Esta sola observación es el elemento que está más al centro del conjunto de números. La mitad de los elementos están por arriba de este punto y la otra mitad está por debajo. **La mediana es una medida de tendencia central que no es afectada por valores extremos o atípicos como lo es la media.** (“Frecuencias: Estadísticos - IBM”) A esta propiedad se la conoce como robustez.

Cálculo de la mediana a partir de datos no agrupados

Para hallar la mediana de un conjunto de datos, primero se organizan en orden descendente o ascendente. Si el conjunto de datos contiene un número impar de elementos, el de en medio en el arreglo es la mediana; si hay un número par de observaciones, la mediana es el promedio de los dos elementos de en medio. En lenguaje formal, la mediana es:

Cálculo de la mediana a partir de datos agrupados A menudo, tenemos acceso a los datos hasta después de agruparlos en una distribución de frecuencias.

$$\bar{m} = \left(\frac{(n+1)/2 - (F+1)}{fm}\right) w + L_m$$

En ese caso utilizamos esta formula

La moda

La moda es una medida de tendencia central diferente de la media, pero un tanto parecida a la mediana, pues en realidad no se calcula mediante algún proceso aritmético ordinario. La moda es el valor que más se repite en el conjunto de datos.

https://cdn.cutypaste.com/wp-content/uploads/2014/03/anigif_enhanced-buzz-1567-1374293192-4.gif

$$M_0 = L_{M_0} + \left(\frac{d_1}{d_1 + d_2}\right)\omega$$

Fórmula para calcular la moda si los datos están agrupados.

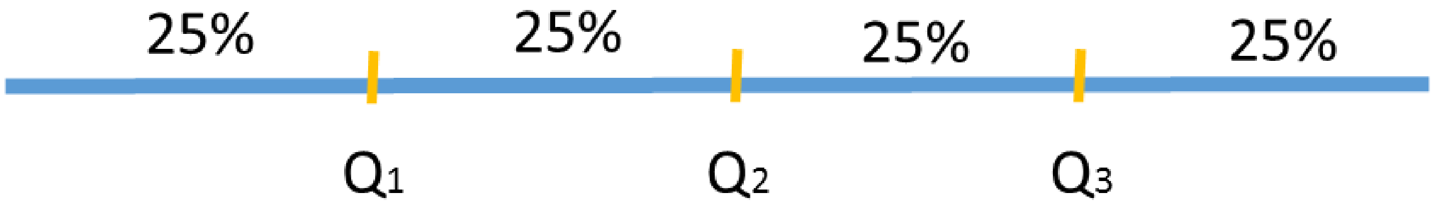
Ejemplo:

En los números 23 24 25 20 23 24 20 23 26 23, la moda es 23 ya que se repite 4 veces.

En los números 1 2 3 3 4 5 6, la moda es 3 ya que es el único número que se repite.

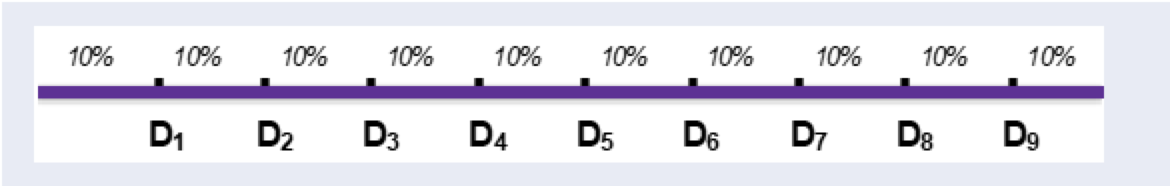
Media, mediana y moda de datos agrupados

Cuantiles, Deciles y percentiles



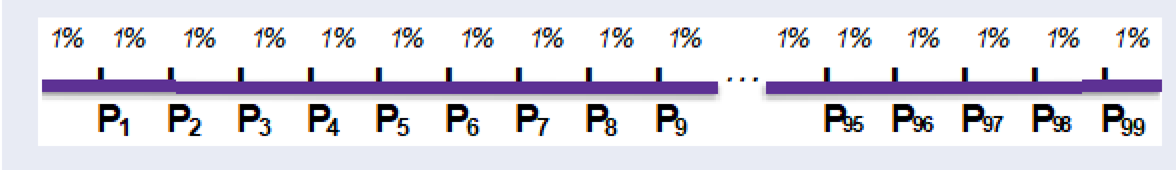
Para un conjunto de datos numéricos en forma ascendente o en una tabla de distribución de frecuencias, se llaman Cuartiles denotados por Q_1 Q_2 y Q_3 a los valores que dividen el conjunto de datos en 4 partes iguales. En la siguiente figura se ubica el Cuartil uno (Q_1) Cuartil dos (Q_2) y Cuartil tres (Q_3). La línea representa un conjunto de datos en orden ascendente, cada porción entre los cuartiles contiene el 25 de los datos, de manera que se puede observar lo siguiente: • A la izquierda de Q 1 se acumula el 25 de los datos y a la derecha de Q 1 está el 75 de los datos • A la izquierda de Q 2 se acumula el 50 de los datos y a la derecha de Q 2 está el 50 de los datos • A la izquierda de Q 3 se acumula el 75 de los datos y a la derecha de Q 3 está el 25 de los datos

Deciles



Para un conjunto de datos numéricos en forma ascendente o en una tabla de distribución de frecuencias, se llaman Deciles denotados por D 1 D 2 D 3 hasta D 9 a los valores que dividen el conjunto de datos en 10 partes iguales. En la siguiente figura se ubica el Decil uno (D 1) Decil dos (D 2) Decil tres (D 3) y así sucesivamente hasta el Decil nueve (D 9) La línea representa un conjunto de datos en orden ascendente, cada porción entre los deciles contiene el 10 de los datos, de manera que se puede observar lo siguiente: • A la izquierda de D 1 se acumula el 10 de los datos y a la derecha de D 1 está el 90 de los datos • A la izquierda de D 2 se acumula el 20 de los datos y a la derecha de D 2 está el 80 de los datos • A la izquierda de D 3 se acumula el 30 de los datos y a la derecha de D 3 está el 70 de los datos, y así sucesivamente • A la izquierda de D 9 se acumula el 90 de los datos y a la derecha de D 9 está el 10 de los datos

Percentiles



Para un conjunto de datos numéricos en forma ascendente o en una tabla de distribución de frecuencias, se llaman Percentiles denotados por P 1 P 2 P 3 hasta P 99 a los valores que dividen el conjunto de datos en 100 partes iguales. En la siguiente figura se ubica el Percentil uno (P 1) Percentil dos (P 2) Percentil tres (P 3) y así sucesivamente hasta el Percentil noventa y nueve (P 99). La línea representa un conjunto de datos en orden ascendente, cada porción entre los percentiles contiene el 1 de los datos, de manera que se puede observar lo siguiente: • A la izquierda de P 1 se acumula el 1 de los datos y a la derecha de P 1 está el 99 de los datos • A la izquierda de P 2 se acumula el 2 de los datos y a la derecha de P 2 está el 98 de los datos • A la izquierda de P 3 se acumula el 3 de los datos y a la derecha de P 3 está el 97 de los datos, y así sucesivamente • A la izquierda de P 99 se acumula el 99 de los datos y a la derecha de P 99 está el 1 de los datos.

Encuentre en el siguiente conjunto de datos ordenados que valores dividen los datos en cuartiles, deciles y percentiles.

49	57	63	65	68	70	72	75	78	84
53	58	63	65	69	70	73	75	79	84
55	58	63	65	69	71	73	76	79	85
55	58	63	66	69	71	73	76	79	86
55	58	**64**	66	69	71	74	**77**	79	86
55	59	64	67	69	71	74	77	80	87
56	60	64	67	70	71	74	77	81	87
56	61	64	67	70	71	75	78	81	89
57	61	64	67	70	71	75	78	82	96
57	**61**	**65**	**68**	**70**	**72**	**75**	**78**	**83**	96

Medidas de dispersión

La media no es por sí sola una descripción completa o suficiente de los datos. En este apartado presentamos números descriptivos que miden la variabilidad o dispersión de las observaciones con respecto a la media. En todas las áreas hay variaciones. En los deportes, el jugador estrella de futbol puede anotar cinco goles en un partido y ninguno en el siguiente o puede jugar 90 minutos en un partido y sólo 15 en el siguiente. La variación es obvia en el sector de la música; el tiempo meteorológico varía mucho de un día a otro e incluso de una hora a otra; las calificaciones de un examen varían de unos alumnos a otros dentro de un mismo curso con un mismo profesor; la presión sanguínea, el pulso, el nivel de colesterol de una persona varían diariamente.

Las medidas de dispersión nos informan sobre cuánto se alejan del centro los valores de la distribución. Para aumentar nuestro entendimiento del patrón de los datos, debemos medir su dispersión o variabilidad. La dispersión es importante porque proporciona información adicional que permite juzgar la confiabilidad de la medida de tendencia central. Si los datos se encuentran ampliamente dispersos, la posición central es menos representativa de los datos. Aunque dos conjuntos de datos tuvieran la misma media, las observaciones individuales de uno de ellos podrían variar con respecto a la media más que las del segundo.

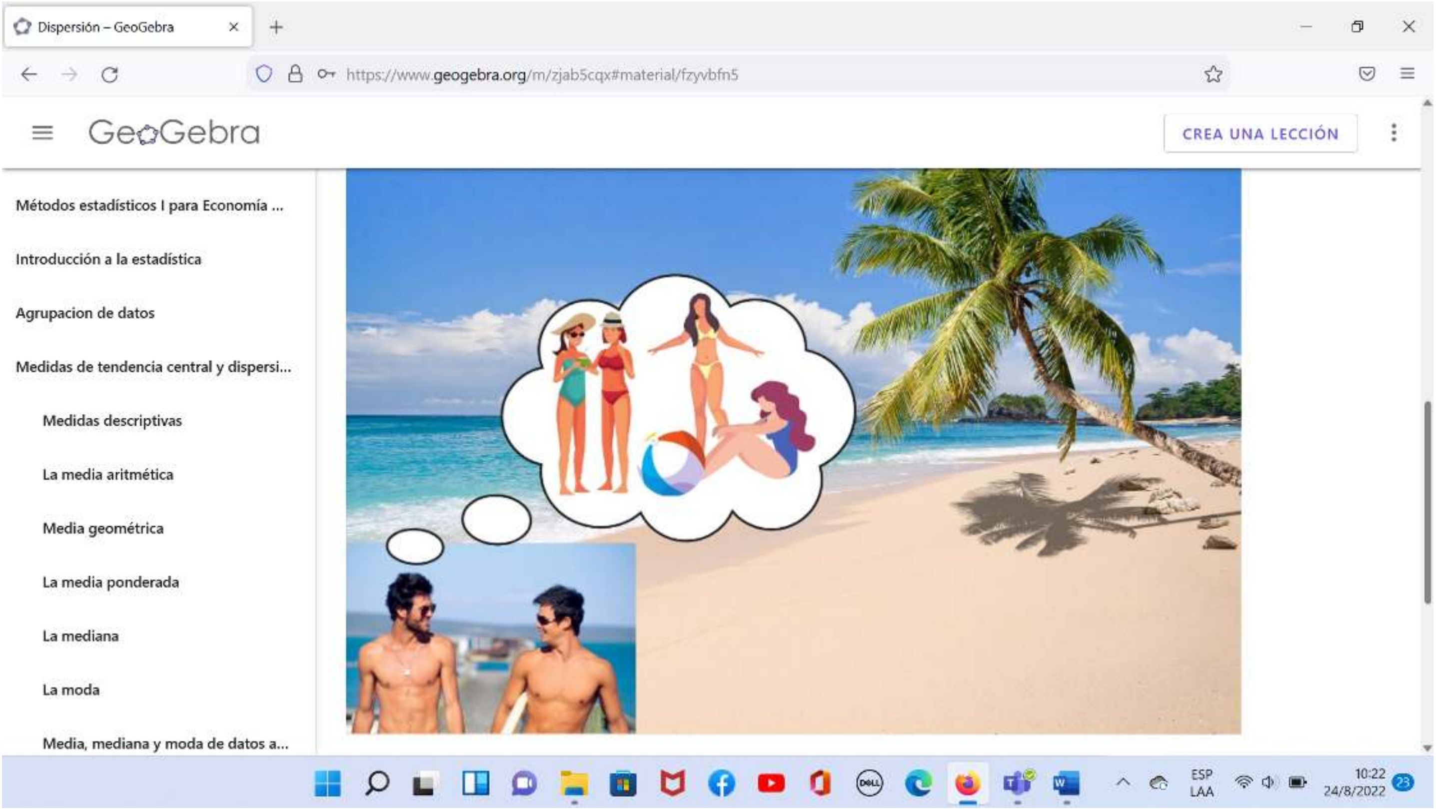
Consideremos los dos conjuntos siguientes de datos muestrales:

Muestra A 1 2 1 36

Muestra B 8 9 10 13

Aunque la media es 10 en ambas muestras, es evidente que los datos de la muestra A están más alejados de 10 que los de la muestra B. Necesitamos números descriptivos para medir esta dispersión.

Un poco de humor nos ayudara a entender la importancia del concepto de dispersión



Un joven insistía en que su mejor amigo lo acompañase a una isla paradisiaca. El amigo no mostraba mucho interés en acompañarlo, pero le pregunto cuál era la edad promedio de las mujeres en esa isla, a lo cual su amigo le contesto que la edad promedio de las mujeres era de **25** años. Con este dato el amigo decide acompañarlo. Al llegar a la isla el amigo se sorprende de no ver personas por ningún lado, hasta que llegan a una pensión atendida por una señora de 48 años. A punto estaba de reclamarle a su amigo por el engaño, cuando la señora le presenta a su nieta de 2 años. En este punto el amigo se percata que eran ellas las únicas habitantes de la isla y que efectivamente su promedio de edad era de 25 años. ¿Qué otra información debió solicitar el amigo antes de tomar una decisión?

El rango

Rango es una medida de dispersión. Es la diferencia entre la observación mayor y la menor. Cuanto mayor es la dispersión de los datos con respecto al centro de la distribución, mayor es el rango. Como el rango sólo tiene en cuenta la observación mayor y la menor, puede estar muy distorsionado si hay una observación excepcionalmente extrema.

Aunque el rango mide la dispersión total de los datos, puede ser una medida insatisfactoria de la variabilidad (dispersión) debido a que los casos atípicos, o bien muy altos o bien muy bajos, influyen en él. Una manera de evitar esta dificultad es ordenar los datos en sentido ascendente o descendente, descartar algunos de los números más altos y algunos de los más bajos y hallar el rango del resto.

¿Cuál era el rango en el caso de las edades de las mujeres de la isla?

¿Cuál es el rango en el caso de las notas de los 100 estudiantes?

La varianza. Segunda medida de dispersión

Aunque el rango mide la dispersión de los datos, sólo tienen en cuenta dos de los valores de los datos. Necesitamos una medida que considere cada uno de los valores de los datos. Esa medida promediaría la distancia total entre cada observación y la media. Esta distancia sería negativa en el caso de los valores menores que la media (y la distancia no es negativa). Si se eleva al cuadrado cada una de, cada observación (tanto por encima como por debajo de la media) contribuye a la suma de los términos al cuadrado. La media de la suma de los términos al cuadrado se llama varianza.

$$\sigma^2 = \frac{\sum (x-u)^2}{N} = \frac{\sum x^2}{N} - u^2$$

$$\sigma^2 = \frac{\sum (x-u)^2}{N-1} = \frac{\sum x^2}{N-1} - u^2$$

En el caso de la varianza de la muestra la fórmula es:

Obsérvese que, en el caso de los datos muestrales, en la ecuación la varianza se halla dividiendo el numerador por (n -1), y no por n. Como nuestro objetivo es hallar una media de los cuadrados de las desviaciones en torno a la media, sería de esperar que hubiera que dividir por n. ¿Por qué se calcula entonces la varianza muestral dividiendo por (n -1)?

Si tomáramos un número muy grande de muestras, cada una del tamaño n, de la población y calculáramos la varianza muestral, como se hace en la ecuación para cada una de estas muestras, la media de todas estas varianzas muestrales sería la varianza poblacional, esta propiedad indica que la varianza muestral es un «estimador insesgado» de la varianza poblacional, estadísticos matemáticos que han demostrado que, si no se conoce la varianza poblacional, una varianza muestral es un estimador mejor de la varianza poblacional si el denominador de la varianza muestral es (n -1), en lugar de n.

Para calcular la varianza hay que elevar al cuadrado las distancias, lo que altera la unidad de medición, que ahora son unidades al cuadrado. La desviación típica, que es la raíz cuadrada de la varianza, hace que los datos vuelvan a su unidad original de medición. Si las mediciones originales estuvieran en pies, la varianza estaría en pies cuadrados, pero la desviación típica estaría en pies. La desviación típica mide la dispersión media en torno a la media.

La desviación estándar (la medida de dispersión más utilizada)

La desviación típica poblacional es la raíz cuadrada (positiva) de la varianza poblacional y se define de la forma siguiente:

$$\sigma = \sqrt{\sigma^2}$$

En las fórmulas de variabilidad, varianza y desviación estándar, la media de la población ha sido substituida por su estimación en la muestra y el promedio se calcula dividiendo por n-1 en vez de simplemente por n. Esto, que puede causar confusión, no es otra cosa que la compensación por el hecho que la estimación de la variabilidad siempre tiende a subestimar aquella de la población.

Usos de la desviación estándar

El teorema de Chebyshev

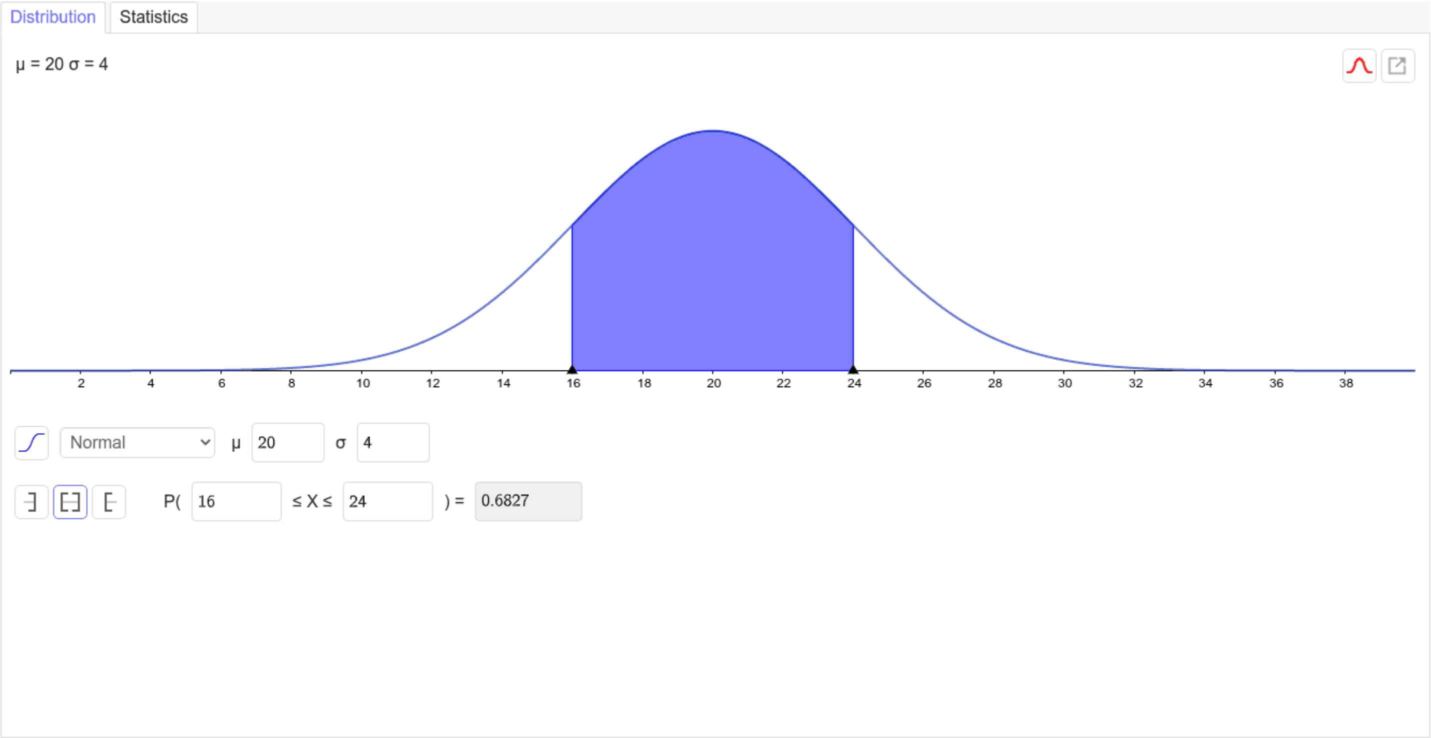
La desviación estándar nos permite determinar, con un buen grado de precisión, dónde están localizados los valores de una distribución de frecuencias con relación a la media. Podemos hacer esto de acuerdo con un teorema establecido por el matemático ruso P. L. Chebyshev (1821-1894).

El teorema de Chebyshev establece que independientemente de la forma de la distribución, al menos 75% de los valores caen dentro de 2 desviaciones estándar a partir de la media de la distribución, y al menos 89% de los valores caen dentro de 3 desviaciones estándar a partir de la media. Podemos medir aún con más precisión el porcentaje de observaciones que caen dentro de un rango específico de una curva simétrica con forma de campana. En estos casos, podemos decir que: 1. Aproximadamente 68% de los valores de la población cae dentro de ±1 desviación estándar a partir de la media. 2. Aproximadamente 95% de los valores estará dentro de ±2 desviaciones estándar a partir de la media 3. Aproximadamente 99% de los valores estará en el intervalo que va desde 3 desviaciones estándar a la izquierda de la media hasta 3 desviaciones estándar a la derecha de la media.

La desviación estándar es útil también para describir cuánto se apartan las observaciones individuales de una distribución de la media de esta. Una medida que se conoce como resultado estándar nos da el número de desviaciones estándar que una observación en particular ocupa por debajo o por encima de la media. Si x simboliza la observación, entonces el resultado estándar calculado a partir de los datos de la población es:

$$Z_i = \frac{(X_i - \mu)}{\sigma}$$

Distribución normal. Observa el applet



Coeficiente de variación

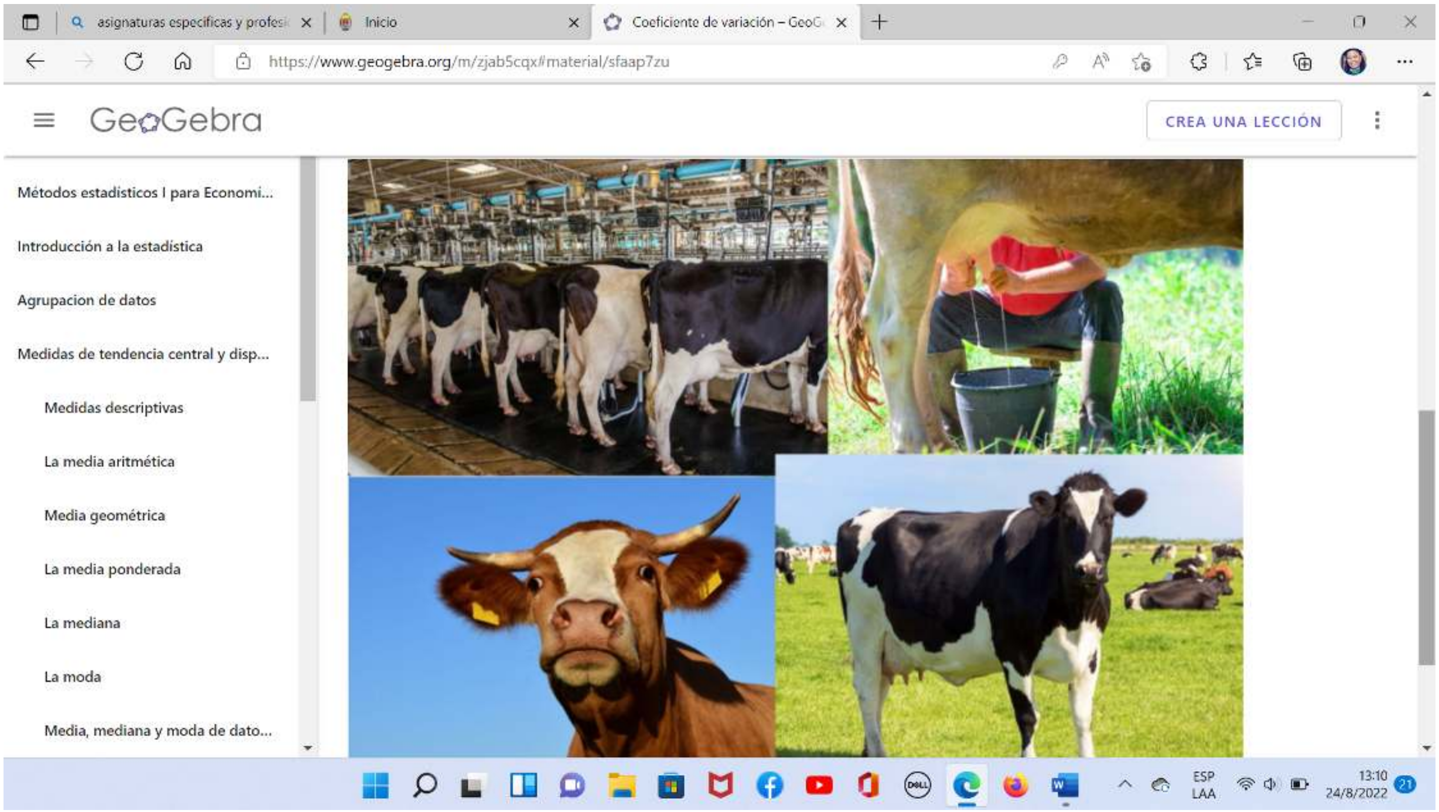
La desviación estándar es una medida absoluta de la dispersión que expresa la variación en las mismas unidades que los datos originales. La desviación estándar no puede ser la única base para la comparación de dos distribuciones. Si tenemos una desviación estándar de 10 y una media de 5, los valores varían en una cantidad que es el doble de la media. Si, por otro lado, tenemos una desviación estándar de 10 y una media de 5,000, la variación relativa a la media es insignificante. En consecuencia, no podemos conocer la dispersión de un conjunto de datos hasta que conocemos su desviación estándar, su media y cómo se compara la desviación estándar con la media. Lo que necesitamos es una medida relativa que nos proporcione una estimación de la magnitud de la desviación respecto a la magnitud de la media. El coeficiente de variación es una de estas medidas relativas de dispersión. Relaciona la desviación estándar y la media, expresando la desviación estándar como porcentaje de la media. La unidad de medida, entonces, es "porcentaje", en lugar de las unidades de los datos originales. Para una población, la fórmula para el coeficiente de variación es:

$$CV = \frac{S_x}{\overline{X}} \cdot 100$$

En una lechería la producción por vaca es 15: 2, entonces se entiende que la producción promedio por vaca es 15 litros, con una desviación estándar de 2 litros y un CV=13.3%.

Si en otra lechería la producción por vaca es 14: 0,5, entonces en ésta la producción promedio por vaca es de 14 litros con una desviación estándar de 0,5 litros y un CV=3.6%.

En consecuencia, la producción en la lechería B es más que en la lechería A.



Ejercicio

Encuentre medidas de tendencia central y dispersión para los siguientes datos: 1,000 2,500 3,000 3,500 3,600 3,800 4,000 4,200 4,200 5,000

CAPITULO IV

Probabilidad: Ideas introductorias

Terminología básica de probabilidad

Jacob Bernoulli (1654-1705), Abraham de Moivre (1667-1754), el reverendo Thomas Bayes (1702- 1761) y Joseph Lagrange (1736-1813) desarrollaron fórmulas y técnicas para el cálculo de la probabilidad. En el siglo XIX, Pierre Simon, marqués de Laplace (1749-1827), unificó todas estas ideas y compiló la primera teoría general de probabilidad. La teoría de la probabilidad fue aplicada con éxito en las mesas de juego y, lo que es más importante en nuestra área de estudio, a problemas sociales y económicos. La industria de seguros, que surgió en el siglo XIX, requería un conocimiento preciso acerca de los riesgos de pérdida, con el fin de calcular las primas. Medio siglo más tarde, muchos centros de aprendizaje estaban estudiando la probabilidad como una herramienta para el entendimiento de los fenómenos sociales. En la actualidad, la teoría matemática de la probabilidad es la base para las aplicaciones estadísticas, tanto en investigaciones sociales como en la toma de decisiones. La probabilidad de un evento solo puede ser un número entre 0 y 1 y también puede escribirse como un porcentaje. En general, la probabilidad es la posibilidad de que algo pase. Las probabilidades se expresan como fracciones (1/6, 1/2, 8/9) o como decimales (0.167, 0.500, 0.889) que están entre cero y uno. Tener una probabilidad de cero significa que algo nunca va a suceder; una probabilidad de uno indica que algo va a suceder siempre.

Eventos En la teoría de la probabilidad, un evento es uno o más de los posibles resultados de hacer algo. Al lanzar una moneda al aire, si cae escudo es un evento, y si cae cara es otro. De manera similar, si sacamos una carta de un mazo de naipes, el tomar el as de corazones es un evento. Un ejemplo de evento que, quizá, esté más cercano a su quehacer diario es ser elegido de entre diez estudiantes para que responda a una pregunta. Espacio Muestral - El espacio muestral de un experimento es el conjunto que contiene solamente a todos los eventos simples posibles. De aquí en adelante utilizaremos la letra S para referirnos al espacio muestral. Ejemplo 5: Halle el espacio muestral de lanzar al azar un dado. Respuesta: S = {1, 2, 3, 4, 5, 6} Ejemplo 6: Halle el espacio muestral de lanzar al azar dos monedas americanas. Respuesta: S = {(cara-cara), (cara-cruz), (cruz-cara), (cruz-cruz)}

Experimento En la teoría de probabilidad, la actividad que origina uno de dichos eventos se conoce como experimento. En Devore (2008), se dice que un *experimento* es cualquier acción o proceso cuyo resultado está sujeto a la incertidumbre.

Ejemplos: Lanzar un dado y observar la cantidad de puntos que aparecen en la cara superior. Lanzar una moneda tres veces y contar la cantidad total de caras obtenidas. Lanzar cuatro veces una moneda y observar la sucesión de escudos y caras obtenidas. Registrar la temperatura de un laboratorio cada 20 minutos, durante una semana de control. Los experimentos anteriores se consideran *experimentos aleatorios*, ya que:

1. Es posible repetir cada experimento en forma indefinida sin cambiar esencialmente las condiciones,
2. Podemos describir el conjunto de todos los posibles resultados del experimento.
3. A medida que el experimento se repite aparece un patrón definido de los resultados.

Utilizando un lenguaje formal, podríamos hacer la siguiente pregunta: en un experimento de lanzar una moneda, ¿cuál es la probabilidad del evento cara? Y, desde luego, si la moneda no está cargada y tiene la misma probabilidad de caer en cualquiera de sus dos lados (sin posibilidades de que caiga parada), podríamos responder, “1/2” o “0.5”. Al conjunto de todos los resultados posibles de un experimento se le llama **espacio muestral** del experimento. Considérese un experimento que consta de la observación de 3 semillas en un cierto orden, cada una de las cuales puede estar sana (situación que se representará con el signo “+”) o bien enferma (situación que se representará con el signo “-”). Hay 8 resultados posibles en el experimento, los que conforman un conjunto que se denomina espacio muestral y que a continuación se representa: **Ω** = {+ + +, + - -, + + -, - - +, + - +, - + -, - + +, - - -} En el de lanzar una moneda, el espacio muestral es S {cara, escudo}

En el experimento de sacar una carta, el espacio muestral tiene 52 elementos: as de corazones, dos de corazones, etcétera. A la mayoría de las personas les emocionan menos el lanzamiento de monedas o las cartas que las preguntas como, ¿cuáles son las posibilidades de poder tomar ese avión a tiempo?, o ¿cuáles son mis posibilidades de conseguir una segunda entrevista de trabajo? En resumen, estamos preocupados por la probabilidad de que ciertos eventos sucedan.

Se dice que los eventos son **mutuamente excluyentes** si uno y sólo uno de ellos puede tener lugar a un tiempo. Considere de nuevo el ejemplo de la moneda. Tenemos dos resultados posibles, cara y escudo. En cualquier lanzamiento obtendremos una cara o un escudo, nunca ambas. En consecuencia, se dice que los eventos cara y escudo en un solo lanzamiento son mutuamente excluyentes.

De manera parecida, usted puede pasar o reprobar una materia o, antes de que termine el curso, desertar y no obtener calificación. Solamente uno de esos tres resultados es posible, por tanto, se dice que son eventos mutuamente excluyentes. La pregunta fundamental que se debe formular al decidir si ciertos eventos son mutuamente excluyentes es: ¿pueden ocurrir dos o más de tales eventos al mismo tiempo? Si la respuesta es afirmativa, los eventos no son mutuamente excluyentes.

Fuente: Levin, Richard (2004). Estadística para administración y economía. Séptima ed.

Objetivos de la unidad

Luego de finalizar el estudio de este módulo estarás capacitado para 1. Distinguir entre probabilidad empírica, teórica y subjetiva. 2. Determinar el espacio muestral generado en un experimento. 3. Determinar si un evento dado de un espacio muestral es simple o no. 4. Determinar la probabilidad asociada a un evento simple. 5. Determinar la probabilidad asociada al complemento de un evento. 6. Definir y distinguir eventos compuestos. 7. Determinar la probabilidad asociada a un evento compuesto. 8. Definir eventos mutuamente excluyentes y eventos independientes. 9. Distinguir entre eventos dependientes e independientes. 10. Calcular la probabilidad de eventos mutuamente excluyentes y eventos independientes. 11. calcular la probabilidad condicional de un evento. 12. utilizar las reglas de la suma y la multiplicación para hallar probabilidades.

Tres tipos de probabilidad

De acuerdo con Levin y Rubin existen tres maneras básicas de clasificar la probabilidad; éstas representan planteamientos conceptuales bastante diferentes para el estudio de la teoría de probabilidad.

1. El planteamiento clásico.
2. El planteamiento de frecuencia relativa.
3. El planteamiento subjetivo.

$$\mathbb{P}(5) = \frac{1}{1+1+1+1+1+1}$$

Probabilidad clásica. Definición Clásica de Laplace, "A Priori" o Teórica Este define la probabilidad de que un evento ocurra como = número de resultados en los que se presenta el evento/número total de resultados posibles. A esta probabilidad también se le conoce como probabilidad *a priori*, debido a que, si empleamos ejemplos como monedas no alteradas, dados no cargados y mazos de barajas normales, entonces podemos establecer la respuesta de antemano (*a priori*) *sin* necesidad de lanzar una moneda, un dado o tomar una carta. No tiene que efectuarse el experimento para poder llegar a conclusiones. Ejemplo: ¿Cuál es la probabilidad de que al lanzar un dado n cantidad de veces obtengamos un 5? y es la misma probabilidad para cada cara del dado. Te invito a ver la simulación del lanzamiento de un dado.

Definición Empírica, "A Posteriori", Experimental o de Frecuencia Relativa

Suponga que empezamos por hacernos preguntas complejas como: ¿Cuál es la probabilidad de que yo viva hasta los 100 años?, ¿Cuáles son las posibilidades de que la consulta del médico este llena un lunes? o ¿Cuál es la probabilidad de que la instalación de un nuevo proyecto a las orillas del río cercano a nuestro pueblo ocasione una significativa muerte de peces? Rápidamente nos damos cuenta de que no somos capaces de emitir una respuesta por adelantado, sin antes hacer algo de experimentación, sobre cuáles son esas probabilidades. Otros planteamientos pueden resultar de más utilidad.

En el siglo XIX, los estadísticos británicos, interesados en la fundamentación teórica del cálculo del riesgo de pérdidas en las pólizas de seguros de vida y comerciales, empezaron a recoger datos sobre nacimientos y defunciones (Levin y Rubin).

En la actualidad, a este planteamiento se le llama frecuencia relativa de presentación de un evento y define la probabilidad como: 1. La frecuencia relativa observada de un evento durante un gran número de intentos o; 2. la fracción de veces que un evento se presenta a la larga, cuando las condiciones son estables. Este método utiliza la frecuencia relativa de las presentaciones pasadas de un evento como probabilidad.

Determinamos qué tan frecuentemente ha sucedido algo en el pasado y usamos esa cifra para predecir la probabilidad de que suceda de nuevo en el futuro. La definición frecuencial de probabilidad es distinta ya que se refiere a una serie repetida de estudios aleatorios. Generalmente se usa cuando el espacio muestral es infinito y por tanto no se pueden enumerar todos los resultados posibles del estudio. Así, se repite el estudio un número grande de veces y se registra la frecuencia relativa de ocurrencia de cada resultado, la que es luego usada como un estimador de probabilidad. La definición frecuencial de probabilidad establece que: Si A es un evento y nA es el número de veces que A ocurre en N repeticiones independientes del experimento, la probabilidad del evento A, denotada por P(A), se define como el límite, cuando el número de repeticiones del experimento es grande, de la frecuencia relativa asociada con el evento.

Por ejemplo, consideremos que la germinación de una semilla es un experimento aleatorio (puede germinar o no). Supongamos que con A se representa el evento “encontrar la semilla germinada”. Si se observan 1000 semillas, es decir se repite 1000 veces el ensayo de germinación (N=1000) en condiciones tales que cada observación no afecte a las otras y 600 semillas germinan (nA=600), se dice que la probabilidad estimada de observar una semilla germinada está dada por: P(A) = P (observar una semilla germinada) = nA/N = 600 / 1000 = 0,6

Entonces, surge la pregunta: ¿Qué diferencia existe entre el concepto de frecuencia relativa y el de probabilidad? Si bien la analogía es fundamental, las frecuencias se entienden como probabilidades sólo cuando N tiende a infinito. Si el número de veces que se repite un experimento no es grande, entonces hablaremos de frecuencia relativa y diremos que ésta “aproxima” una probabilidad.

Probabilidades subjetivas

Se puede definir como la probabilidad asignada a un evento por parte de un individuo, basada en la evidencia que tenga disponible. Esta evidencia puede presentarse en forma de frecuencia relativa de presentación de eventos pasados o puede tratarse, simplemente, de una creencia meditada. Quizá la más antigua estimación de probabilidad subjetiva de la posibilidad de que fuera a llover se dio cuando alguna tía anciana dijo: “Me duelen los huesos, creo que se avecina lluvia.”

Las valoraciones subjetivas de la probabilidad permiten una más amplia flexibilidad que los otros dos conceptos analizados. Los tomadores de decisiones pueden hacer uso de cualquier evidencia que tengan a mano y mezclarla con los sentimientos personales sobre la situación. Las asignaciones de probabilidad subjetiva se dan con más frecuencia cuando los eventos se presentan sólo una vez o un número muy reducido de veces. Ejemplo. Un juez debe decidir si permite la construcción de una planta de energía nuclear en un lugar donde hay evidencias de que existe una falla geológica. Debe preguntarse a sí mismo: “¿Cuál es la probabilidad de que ocurra un accidente nuclear grave en este sitio?” El hecho de que no exista una frecuencia relativa de presentación de la evidencia de accidentes anteriores en ese sitio no es suficiente para liberarlo de tomar la decisión. Debe utilizar su mejor sentido común para determinar la probabilidad subjetiva de que suceda un accidente nuclear. (Ampliar más en la página 133 del libro de texto)

Reglas de probabilidad

Axiomas y reglas de probabilidad

Sean A y B eventos, S el espacio muestral. (A C S y B C S) entonces:

$P(A) \geq 0$ (No negativa)

$0 \leq P(A) \leq 1$

$P(S) = 1$, S: espacio muestral

$P(AUB) = P(A) + P(B)$, si $A \cap B = \{\}$

$P(\emptyset) = 0$, $\emptyset$: vacío

$P(AUB)= P(A)+P(B) - P(A \cap B)$

$P(A \cap B)= P(A)+P(B)- P(AUB)$

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

$P((AUB)^c= P(A^cUB^c)= 1-P(AUB)$

$P((A \cap B)^c = P(A^c \cup B^c)= 1-P(A \cap B)$

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

Si $P(A \cap B)= P(A)*P(B)$ entonces A y B son independientes

$P(A/B) = P(A \cap B)/P(B)$, $P(B)>0$. Probabilidad condicional

Considérese un experimento en el que se analizan 3 pariciones de una vaca (n=3), registrándose el sexo del ternero nacido. Como los resultados posibles de cada parición son dos (N=2), los resultados posibles del experimento son Nn =23 =8. Estos son: HHH, HHM, HMH, MHH, HMM, MHM, MMH y MMM donde M representa una cría macho y H una cría hembra y se asume que estos resultados son igualmente probables. Defina los eventos A como "una cría hembra nace en cada uno de los dos primeros partos"; B como "un macho nace en el tercer parto" y C como "exactamente 2 machos ocurren en los tres partos". Mostrar que A y B son dos eventos independientes, mientras que B y C no lo son:

A = {HHH, HHM} B = {HHM, HMM, MHM, MMM} C = {HMM, MHM, MMH}

A y B son independientes ⇔ $P(A \cap B) = P(A) . P(B)$

Como $A \cap B$ = {HHM} ⇒ $P(A \cap B) = 1/8$.

Por otra parte, $P(A) = 2/8$ y $P(B) = 4/8$ y $P(A).P(B) = 2/8 \cdot 4/8 = 1/8$.

Luego, $P(A \cap B) = P(A) .P(B)$, por consiguiente, A y B son independientes.

Si B y C son independientes ⇒ $P(B \cap C) = P(B) . P(C)$.

En este ejemplo, P(B) = 4/ 8 y P(C) = 3 /8 . Así, P(B) . P(C) = 4 /8 . 3/ 8 = 3 /16.

Por otro lado, B ∩ C = {HMM, MHM} ⇒ P(B ∩ C) = 2/ 8 ≠ P(B) . P(C).

Luego B y C no son independientes.

De la producción de tornillos de cierta magnitud resulta que el 5 % de ellos no tienen el largo especificado, el 7 % no tienen el diámetro especificado y el 2 % tiene ambos defectos. Se elige un tornillo al azar de la producción de estas magnitudes.

¿Cuál es la probabilidad que: a) tenga al menos uno de los dos defectos?

b) tenga sólo el defecto del largo?

c) tenga sólo uno de los defectos

L: Defecto de largo
D:Defecto de diametro

P(L)=0.05
P(D)=0.07
$P(L \cap D) = P(D \cap L) = 0.02$

Dados:

$$P(LUD)$$
$$P(L-D)$$
$$P(L-D)+P(D-L)$$

= 0.1 b) =0.03 c) = 0.08

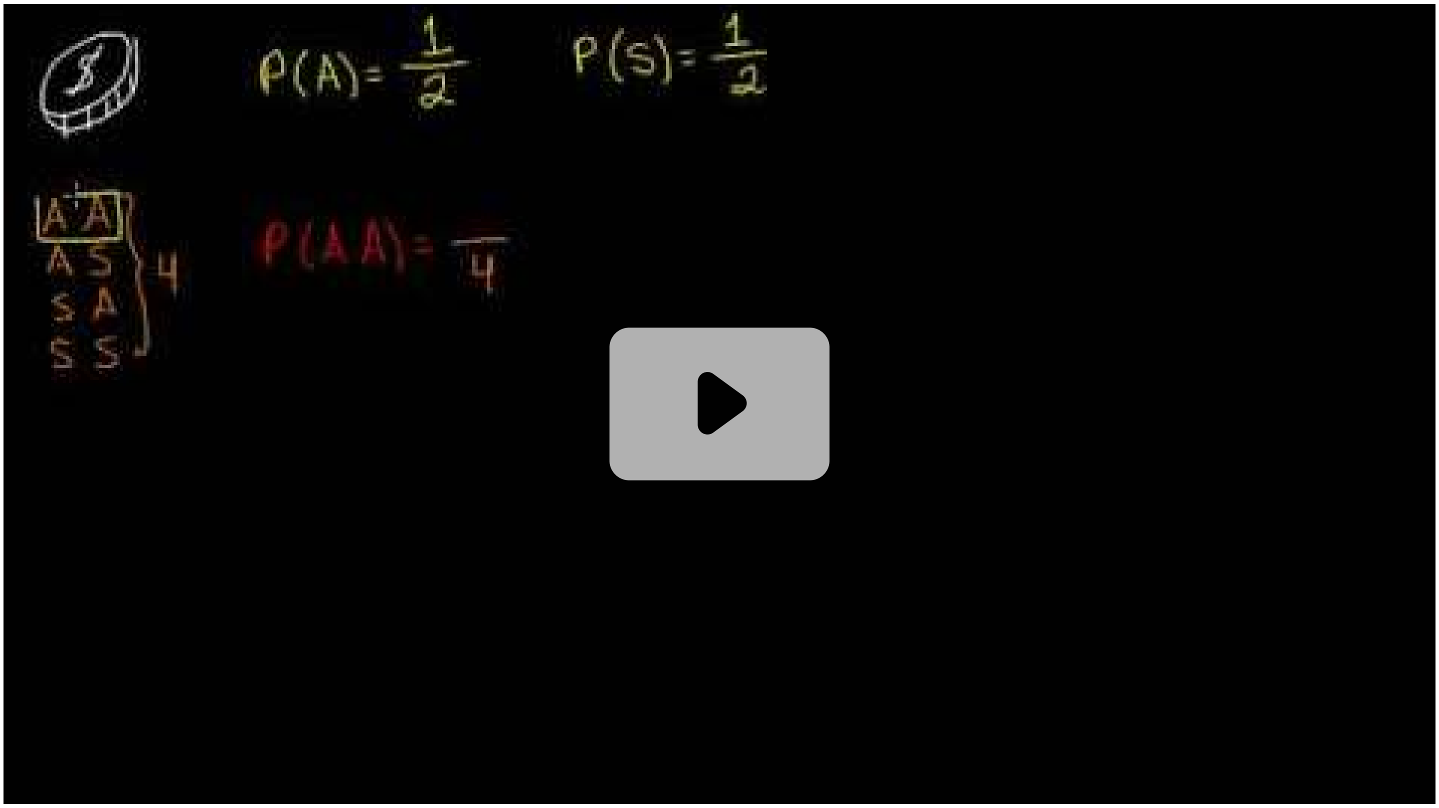
Supongamos que tenemos dos eventos A y B, los cuales son mutuamente excluyentes. La probabilidad de A es de 0.17 y la probabilidad de B es de 0.5.

a) ¿Cuál es la probabilidad de que ocurra A o B?

b) Ahora supongamos que los eventos A y B son independientes. ¿Cuál es la probabilidad de que ocurra A y B?

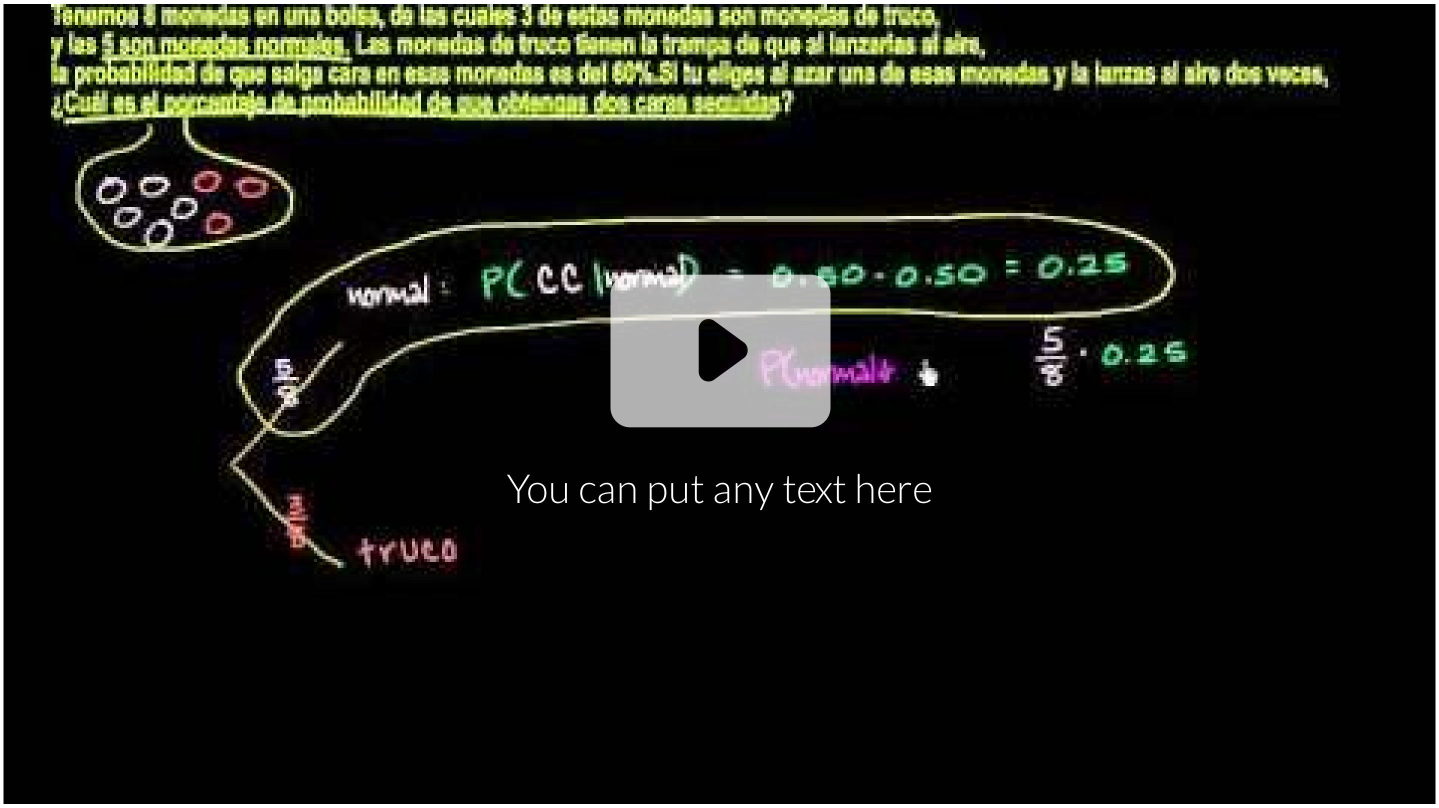
Video Probabilidad compuesta de eventos independientes

Probabilidades bajo condiciones de independencia estadística



Probabilidades bajo condiciones de dependencia estadística

Video Introduccion a la probabilidad dependiente



Probabilidad dependiente -- Ejemplo

Combinaciones y permutaciones

Una persona desea armar una computadora, para lo cual considera que puede seleccionar la Motherboard de entre las dos disponibles, mientras que el procesador puede ser seleccionado de un Pentium IV, un Celeron o un Athlon, la tarjeta de video puede ser una ATI Radeon o una GForce y por último hay disponible un solo modelo de gabinete (Tower). ¿Cuántas maneras tiene esta persona de armar su PC?" Respuesta. Aquí se utiliza el principio muiltiplicativo

Motherboard	2
Procesador	3
Tarjeta de video	2
Gabinete	1
Principio multiplicativo	12

Una pareja quiere ir al sur de Chile. Para ir en avión hay 2 compañías y para ir en autobús hay 3 compañías. ¿Cuántas maneras tienen para ir al sur? Principio sumativo

Avión	2
Autobús	3
Principio de suma	5

Tenemos las letras A, B, C (n=3) quiero escoger 2 (r=2) *a.

Si me importa el orden es una permutación =PERMUTACIONES (3;2) AB, AC, BA, BC, CA, CB * b.

Si no me importa el orden es una combinación AB, AC, BC

Ejemplo 1: En una rifa se van a vender 100 boletos al primer lugar se le darán mil lempiras, al segundo 500 y al tercer lugar 100, cuantas formas de escoger a los ganadores

Respuesta: =PERMUTACIONES (100;3) = 970,200 Si los 3 premios fuesen iguales Respuesta: =COMBINAT(100;3) = 161,700

¿Cuántos comités diferentes de 3 personas puede haber a partir de un grupo de 10 individuos?

Teorema de Bayes

Se observa que hombres y mujeres reaccionan de modo diferente a un conjunto determinado de circunstancias; se sabe que 70% de las mujeres reaccionan positivamente a estas circunstancias mientras que de este mismo modo reaccionan sólo 40% de los hombres. Un grupo de 20 personas, 15 mujeres y 5 hombres se sometió a estas circunstancias y a los sujetos se les pidió describieran sus reacciones en un cuestionario escrito.

Calcule a. Probabilidad de obtener una reacción positiva

$$P(A) = \sum_{1=1}^{k} P(A|B_i)\,P(B_i) \quad PROBABILIDAD\ TOTAL$$

$$P(B_i|A) = \frac{P(A y\ B_i)}{P(A)} = \frac{P(A|B_i\ )P(B_i)}{\sum_{1=1}^{k} P(A|B_i)} \quad \text{TEOREMA DE BAYES}$$

b. Probabilidad de que sea hombre dado que se obtuvo una reacción positiva.

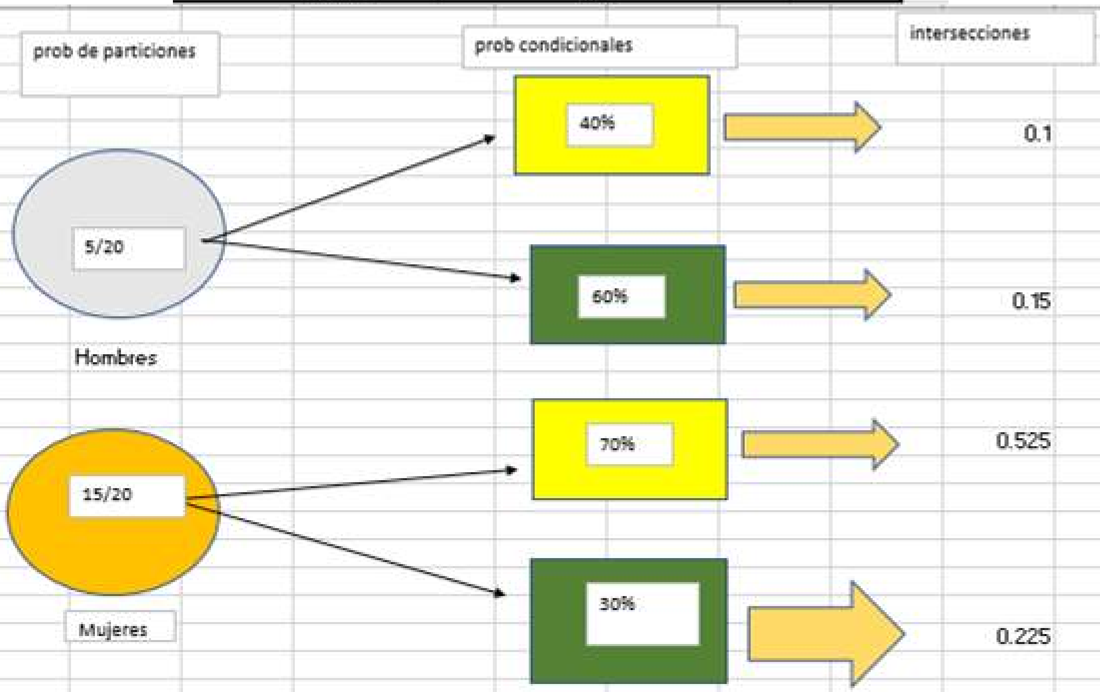
Particiones	Prob	Probabilidades condicionales		
H	0.25	P(B/H)	0.4	
		P(N/H)	0.6	
M	0.75	P(B/M)	0.7	
		P(N/M)	0.3	

Sea A la reaccion es positiva

	P(A)	0.625

La probabilidad de que sea hombre dado que reacciona de manera positiva

	P(H/A)	0.16



	reaccion positiva(B)
	reaccion negativa(N)

P(A)	0.625	P(H/A)=P(Hy A)/P(A)	0.16	

CAPITULO V

Distribuciones de probabilidad

¿Qué es una distribución de probabilidad?

Las distribuciones de probabilidad están relacionadas con las distribuciones de frecuencias. De hecho, podemos pensar que una distribución de probabilidad es una distribución de frecuencias teórica, por lo tanto, una distribución de frecuencias teórica es una distribución de probabilidades que describe la forma en que se espera varíen los resultados. Como estas distribuciones representan expectativas de que algo suceda, resultan modelos útiles para hacer inferencias y tomar decisiones en condiciones de incertidumbre.

Tipos de distribuciones de probabilidad

Las distribuciones de probabilidad se clasifican como discretas y continuas. En la distribución de probabilidad **discreta** está permitido considerar sólo un número limitado de valores. En una distribución de probabilidad **continua**, por otro lado, la variable que se está considerando puede tomar cualquier valor dentro de un intervalo dado. Para una total comprensión de este párrafo debes recordar los diferentes tipos de variables.

Anteriormente, analizamos los resultados posibles de lanzar dos dados y calculamos algunas probabilidades asociadas con los diferentes resultados. Construya una tabla y una gráfica de la distribución de probabilidad que represente los resultados (en términos del número total de puntos que salen cara arriba en ambos dados) de este experimento. ¿De qué tipo de variable estamos hablando? ¿Qué tipo de resultados esperamos?

Distinguir el tipo de variable es útil no solo en la etapa exploratoria del análisis de datos sino también en etapas donde se quiera asignar probabilidades a eventos relacionados con la variable. Para ciertos tipos de variables aleatorias ya se conocen modelos probabilísticos teóricos que ajustan razonablemente bien sus distribuciones empíricas y por tanto se usan estos modelos para el cálculo de probabilidades. Para una variable continua y de distribución simétrica unimodal, es común el uso del modelo **Normal**; mientras que para proporciones se piensa en el modelo probabilístico **Binomial** y para conteos no acotados en el modelo **Poisson**.

Cuando estudiamos una variable aleatoria, es de interés calcular probabilidades sobre la ocurrencia de ciertos valores (eventos). Por ejemplo, podríamos estimar la probabilidad de obtener un rendimiento de maíz superior a 100 qq/ha, de tomar 100 semillas y que no germinen más de 90, o de tomar una muestra de insectos con golpes de red y capturar menos de 20 insectos. Los cálculos de probabilidad pueden hacerse luego de enumerar todo el espacio muestral, cuando esto es posible, usando información sobre las frecuencias con que ocurren los distintos eventos o bien usando un modelo de distribución teórico que ajuste relativamente bien a la distribución empírica de la variable. Para la elección del modelo de probabilidad teórico, es importante considerar características de la variable tales como la forma en que se cuantifica (medición, proporción, conteo, etc.). La naturaleza de la variable, es decir si es discreta o continua las condiciones en que se realiza el experimento y el registro de los valores son determinantes para la selección de un modelo probabilístico.

http://www.agro.unc.edu.ar/~mcia/archivos/Estadistica%20y%20Biometria.pdf

Variables aleatorias

Una variable es aleatoria si toma diferentes valores como resultado de un experimento aleatorio. Esta variable aleatoria puede ser discreta o continua. Si puede tomar sólo un número limitado de valores, entonces es una variable aleatoria discreta. En el otro extremo, si puede tomar cualquier valor dentro de un intervalo dado, entonces se trata de una variable aleatoria continua.

Una variable aleatoria es una especie de valor o magnitud que cambia de una ocurrencia a otra sin seguir una secuencia predecible. Una variable aleatoria X es una función que asocia a cada elemento del espacio muestral S de un experimento aleatorio un número real. $\rightarrow \mathbb{R}$ E$\rightarrow$x Se utilizan letras mayúsculas X para designar variables aleatorias, y las respectivas minúsculas x para designar valores concretos de las mismas.

Se clasifica en variables **aleatorias discretas y continuas.** Variable aleatoria discreta es aquella variable aleatoria que toma un número finito o infinito, pero numerable, de valores. Ejemplos: El número de llamadas realizadas en una hora o la puntuación obtenida al lanzar un dado. Variable aleatoria continua es aquella variable aleatoria que puede tomar un número infinito de valores no numerables en un intervalo en particular. Ejemplos: Los tiempos de los vuelos comerciales entre Tegucigalpa y La Ceiba. Los valores de una variable aleatoria son los valores numéricos correspondientes a cada posible resultado del experimento aleatorio.

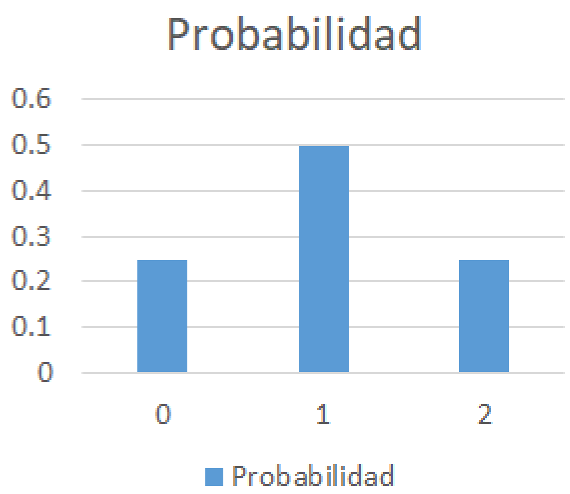
¿Cuántas veces esperarías que caiga cara en el lanzamiento de una moneda 2 veces?

En la tabla ordenamos los resultados para enfatizar el número de caras contenidas en cada resultado. En este punto, debemos tener cuidado y considerar que la tabla no representa el resultado real de lanzar una moneda no alterada dos veces.

1 lanzamiento	2o. lanzamiento	# caras en 2 lanzamientos	Probabilidad de los 4 resultados posibles
Cara Cara	Cara Escudo	2 1 0 1	0.5x0.5 = 0.25 0.5x0.5 = 0.25
Escudo Escudo	Escudo Cara		0.5x0.5 = 0.25 0.5x0.5 = 0.25 1

Distribución de probabilidad del número posible de caras que se obtienen al lanzar dos veces una moneda no alterada.

Numero de caras	Lanzamientos	Probabilidad de este resultado
0 1 2	escudo, escudo (cara, escudo) (escudo, cara) cara, cara	0.25 0.50 0.25



Se trata del resultado teórico, es decir, representa la forma en que esperamos que se comporte nuestro experimento de dos lanzamientos. Podemos representar gráficamente la distribución de probabilidad de la tabla. Para ello, colocamos en una gráfica el número de caras que deberíamos ver en dos lanzamientos contra la probabilidad de que este número se presente.

Sea X variable aleatoria que a cada cara del dado le asocia el número de puntos, construya su distribución de probabilidad.

El valor esperado de una variable aleatoria

¿Cuál es el resultado más probable al lanzar 2 dados al aire?

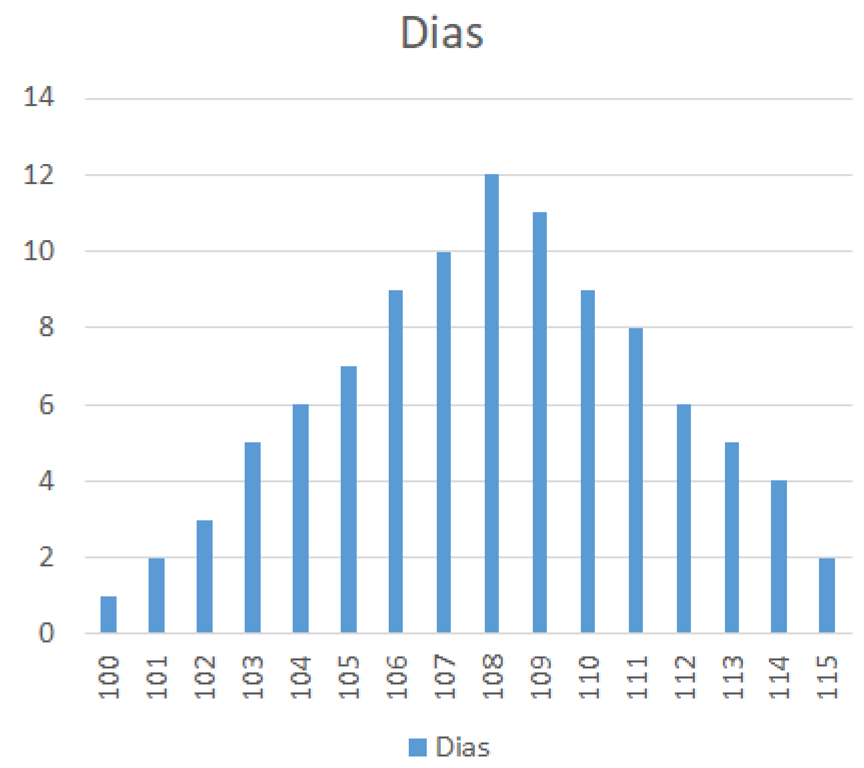
Solución.

A continuación presentamos todos los sucesos que pueden ocurrir al lanzar dos dados y el valor que para cada uno de estos sucesos tiene la variable suma:

(1,1) 2	(2,1) 3	(3,1) 4	(4,1) 5	(5,1) 6	(6,1) 7
(1,2) 3	(2,2) 4	(3,2) 5	(4,2) 6	(5,2) 7	(6,2) 8
(1,3) 4	(2,3) 5	(3,3) 6	(4,3) 7	(5,3) 8	(6,3) 9
(1,4) 5	(2,4) 6	(3,4) 7	(4,4) 8	(5,4) 9	(6,4) 10
(1,5) 6	(2,5) 7	(3,5) 8	(4,5) 9	(5,5) 10	(6,5) 11
(1,6)) 7	(2,6) 8	(3,6) 9	(4,6) 10	(5,6) 11	(6,6) 12

Como todos estos sucesos tienen la misma probabilidad 1/36, la distribución de la suma será:

X	2	3	4	5	6	7	8	9	10	11	12
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36



Número de mujeres examinadas durante 100 días en una clínica Examinadas Número de días que se observó esta cantidad Probabilidad 100 1 0.01 101 2 0.02 102 3 0.03 103 5 0.05 104 6 0.06 105 7 0.07 106 9 0.09 107 10 0.10 108 12 0.12 109 11 0.11 110 9 0.09 111 8 0.08 112 6 0.06 113 5 0.05 114 4 0.04 115 2 0.02 100 1.00

El valor esperado es una idea fundamental en el estudio de las distribuciones de probabilidad. Durante muchos años, el concepto ha sido puesto en práctica con bastante regularidad por las compañías aseguradoras y, en los últimos años, también ha sido utilizado ampliamente por muchas de las personas que deben tomar decisiones en condiciones de incertidumbre. Para obtener el valor esperado de una variable aleatoria discreta, multiplicamos cada valor que la variable puede tomar por la probabilidad de ocurrencia de ese valor y luego sumamos los productos. Revisar página 184 del libro de texto. ¿Qué significa esto? Significa que, en un periodo largo, el número de mujeres examinadas diariamente deberá tener un promedio de aproximadamente 108.02. Recuerde que un valor esperado de 108.02 no significa que mañana 108.02 mujeres asistan a la clínica. La directora de la clínica podría basar sus decisiones en el valor esperado del número de mujeres examinadas diariamente debido a que éste es un promedio ponderado de los resultados que espera en el futuro.

El valor esperado pesa cada resultado posible con respecto a la frecuencia con que se espera se presente. En consecuencia, las ocurrencias más comunes tienen asignado un peso mayor que las menos comunes. Conforme van cambiando las condiciones, la directora podría recalcular el valor esperado de los exámenes diarios y utilizar el nuevo resultado como base para tomar decisiones.

Uso del valor esperado en la toma de decisiones

En este apartado analizaremos cómo los tomadores de decisiones combinan las probabilidades de que una variable aleatoria asuma ciertos valores con las ganancias o pérdidas monetarias que se dan cuando efectivamente toma estos valores. De esta forma, los responsables son capaces de decidir inteligentemente en condiciones de incertidumbre.

Combinación de probabilidades y valores monetarios

Un vendedor al mayoreo de frutas y legumbres que comercia con frambuesas. Este producto tiene una vida útil muy limitada: si no se vende el día que llega, ya no tiene valor. Una caja de frambuesas cuesta $20 y el vendedor recibe $50 por ella. Éste no puede especificar el número de cajas que un cliente pedirá en cualquier día dado, pero su análisis de registros pasados ha producido la información que presentamos en la tabla.

Ventas durante 100 días

Ventas diarias	Número de días de ventas	Probabilidad de venta de cada cantidad
10 11 12 13	15 20 40 25	0.15 0.20 0.40 0.25

El vendedor al mayoreo ha sufrido dos tipos de pérdidas: 1) pérdidas por obsolescencia, ocasionadas por tener en existencia demasiada fruta en un día y tener que tirarla al siguiente, y 2) pérdidas de oportunidad, ocasionadas por no tener en existencia el producto al momento en que un cliente lo solicita (los clientes no esperan más allá del día en que solicitan una caja de bananas).

Los valores que se tienen en la tabla incluyen no solamente las pérdidas por la fruta descompuesta, sino también las que se derivan de los ingresos perdidos cuando el vendedor no es capaz de suministrar un pedido. Cuando el número de cajas en existencia en un día cualquiera es igual al número de cajas solicitadas no ocurre ninguno de estos dos tipos de pérdida. En tales casos, el vendedor vende todo lo que tiene almacenado y no sufre pérdidas. Esta situación se indica con el cero en negrita que aparece en la columna correspondiente. Las cifras que se encuentren por encima de un cero cualquiera representan las pérdidas sufridas al tener que tirar la fruta. En este ejemplo, el número de cajas almacenadas es mayor al de cajas solicitadas. Por ejemplo, si el vendedor tiene en existencia 12 cajas, pero recibe solicitud para sólo 10 de ellas, pierde $40 (o $20 por caja no vendida ese mismo día).

Los valores que se encuentran debajo de los ceros en negrita representan las pérdidas de oportunidad derivadas de pedidos que no se pueden cumplir. Si, un cierto día, el vendedor tiene en existencia solamente 10 cajas de bananas y le solicitan 11, éste sufre una pérdida de oportunidad de $30 por la caja que le faltó ($50 por caja menos $20 de su costo, igual a $30).

Cálculo de pérdidas esperadas

La tabla B es una tabla de pérdidas condicionales. Cada valor en ella está condicionado a un número específico de cajas que se encuentran en existencia y a un número específico de solicitudes.

La acción de almacenamiento óptima es aquella que minimiza las pérdidas esperadas. Tener en existencia 12 cajas diariamente constituye esta opción, en cuyo caso las pérdidas esperadas toman el valor mínimo de $17.50. Con la misma facilidad, se pudo haber resuelto este problema tomando un camino alternativo, es decir, maximizando la ganancia esperada ($50 recibidos por caja de fruta, menos $20 del costo de cada caja), en lugar de minimizar la pérdida esperada. En cualquier caso, habríamos obtenido la misma respuesta: 12 cajas en existencia. **Para ampliar este tema remítase a la página 187 del libro de texto.**

Ejercicio

La compañía Airport Rent-a-Car opera de manera local y compite con varias alquiladoras más grandes. Airport Rent-a-Car está planeando ofrecer un nuevo contrato a los clientes potenciales que deseen alquilar un automóvil por sólo un día para devolverlo en el aeropuerto. La tarifa será de $35 y el automóvil, un modelo compacto económico; el único gasto adicional del cliente será llenar el tanque del automóvil al término del día. Airport Rent-a-Car tiene planeado comprar cierto número de automóviles compactos al precio especial de $6,300. La pregunta que se tiene que responder es: ¿cuántos automóviles deben comprar? Los ejecutivos de la compañía han estimado la siguiente distribución para la demanda diaria del servicio:

Número de automóviles alquilados 13 14 15 16 17 18 Probabilidad 0.08 0.15 0.22 0.25 0.21 0.09

La compañía pretende ofrecer el servicio seis días a la semana (312 días al año) y estima que el costo por automóvil por día será de $2.50. Al término de un año, la compañía espera vender los automóviles y recuperar el 50% del costo original. Sin tomar en cuenta el valor temporal del dinero ni cualquier otro gasto que no sea en efectivo, utilice el método de pérdida esperada para determinar el número óptimo de automóviles que debe comprar la compañía.

Tipos de distribuciones de probabilidad

Las distribuciones de probabilidad se clasifican como discretas y continuas. En la distribución de probabilidad discreta está permitido considerar sólo un número limitado de valores. La probabilidad de que usted haya nacido en un mes dado es discreta, puesto que sólo hay 12 posibles valores (los 12 meses del año).

En una distribución de probabilidad continua, por otro lado, la variable que se está considerando puede tomar cualquier valor dentro de un intervalo dado. Las distribuciones continuas son una forma conveniente de presentar distribuciones discretas que tienen muchos resultados posibles, todos muy cercanos entre sí.

Distribución binomial

Una distribución de probabilidad de variable aleatoria discreta utilizada ampliamente es la distribución binomial. Esta distribución describe **datos discretos**, no continuos, que son resultado de un experimento conocido como proceso de Bernoulli, en honor del matemático suizo nacido en el siglo XVII, Jacob Bernoulli. El lanzamiento de la moneda no alterada un número fijo de veces es un proceso de Bernoulli, y los resultados de tales lanzamientos pueden representarse mediante la distribución binomial de probabilidad. El éxito o fracaso de los aspirantes en la Prueba de aptitud de la UNAH, también puede ser descrito como un proceso de Bernoulli.

Cada proceso de Bernoulli tiene su propia probabilidad característica. Considere una situación en la que, a lo largo del tiempo, siete décimas partes de todos los aspirantes aprueban el examen de aptitudes. Diríamos que, en este caso, la probabilidad característica es de 0.7, pero podríamos describir el resultado del examen como de Bernoulli sólo si tenemos la certeza de que la fracción de los que aprueban el examen (0.7) permanece constante en el tiempo. Desde luego que las otras características del proceso de Bernoulli también deben cumplirse.

Cada examen debería tener solamente dos resultados (éxito o fracaso) y los resultados de cada prueba deberían ser estadísticamente independientes. En un lenguaje más formal, el símbolo p representa la probabilidad de tener éxito (0.7 en este ejemplo) y el símbolo q (q=1 - p) es la probabilidad de que resulte en un fracaso (0.3). Para representar un cierto número de éxitos, utilizaremos el símbolo r, y para representar el número total de intentos o de ensayos utilizamos el símbolo n. En las situaciones que analizaremos, el número de ensayos está fijo desde antes de empezar el experimento.

Un ejemplo: Calculemos, para utilizar este lenguaje en un problema sencillo, las posibilidades de obtener exactamente dos caras (en cualquier orden) en tres lanzamientos de una moneda no alterada. Simbólicamente, expresamos los valores de la forma siguiente:

P= probabilidad característica o probabilidad de tener éxito 0.5

q =1- p probabilidad de fracaso 0.5 •

r =número de éxitos deseados = 2

n =número de intentos hechos = 3

$$\frac{n!}{r!\,(n-r)!}\,p^r\,q^{n-r}$$

Fórmula binomial Podemos resolver el problema utilizando la fórmula binomial: Probabilidad de r éxitos en n intentos =

El símbolo **!** significa **factorial** y se calcula de la manera siguiente: 3! significa 3 X 2 X 1 igual a 6.

Para calcular 5!, multiplicamos 5 X 4 X 3 X 2 X 1 = 120.

Los matemáticos definen 0! igual a 1.

$$\frac{3!}{2!\,(3-2)!}\,(0.5)^2(0.5)^1$$

Utilizando la fórmula binomial para resolver nuestro problema, descubrimos que Probabilidad de 2 éxitos en 3 intentos Al realizar la operación nos da 0.375 Por tanto, existe una probabilidad de 0.375 de obtener dos caras en tres lanzamientos de una moneda no alterada.

Uso de Excel

Es frecuente que los 5 empleados de una finca lleguen tarde al trabajo. El propietario ha estudiado la situación durante cierto periodo y determino que hay una probabilidad de 0.4 de que cualquier empleado llegue tarde y que las llegadas tarde de los mismos son independientes. ¿Cómo podría trazar una distribución binomial de probabilidades que ejemplifique las probabilidades de que 0, 1, 2, 3, 4 o 5 empleados lleguen tarde simultáneamente?

En excel tenemos 2 opciones
DIST.BINOM.N con opciones
*FALSO nos da la probabilidad en un punto
*VERDADERO nos da la probabilidad acumulada hasta ese punto

n	5
éxito	llegar tarde
fracaso	no llegar tarde
p	0.4
q	0.6

x Cuenta el # de empleados que llegan que llegan tarde x va desde 0 hasta 5

x	p(x) con Excel opción falso	F(x)	Con Excel opción verdadero

¿Cuál es la probabilidad de que a lo más 2 empleados lleguen tarde? o sea P(X≤2) ¿Cuál es la probabilidad de que exactamente 2 empleados lleguen tarde? P(X=2) ¿Cuál es la probabilidad de que entre 2 y 4 empleados lleguen tarde? ¿Cuál es la probabilidad de que más de 2 empleados lleguen tarde? ¿Cuál es la probabilidad de que 2 empleados o más lleguen tarde?

En unas pruebas de alcoholemia se ha observado que el 5% de los conductores controlados dan positivo en la prueba y que el 10% de los conductores controlados no llevan puesto el cinturón de seguridad. También se ha observado que las dos infracciones son independientes. Un guardia de tráfico para cinco conductores al azar. Si tenemos en cuenta que el número de conductores es suficientemente importante como para estimar que la proporción de infractores no varía al hacer la selección. Calcule lo siguiente: De ser necesario repase las reglas de probabilidad o vea el video en youtube Ampliar tema en página 195 del libro de texto

Distribucion binomial. Ejercicio resuelto paso a paso

Presentación gráfica de la distribución binomial

Es frecuente que los empleados lleguen tarde a trabajar a la Farmacia Kerr y hay cinco empleados en ella. El propietario ha estudiado la situación durante cierto periodo y determinó que hay una probabilidad de 0.4 de que cualquier empleado llegue tarde y que las llegadas de estos son independientes entre sí. ¿Cómo podríamos trazar una distribución binomial de probabilidad que ejemplifique las probabilidades de que 0, 1, 2, 3, 4 o 5 empleados lleguen tarde simultáneamente?

Para hacerlo, necesitaríamos utilizar la fórmula binomial donde: p 0.4 q 0.6 n 5* y efectuar cálculos separados para cada r, desde 0 hasta 5. Recuerde que, matemáticamente, cualquier número elevado a la cero potencia es igual a 1.

Grafica

Sin efectuar todos los cálculos necesarios, podemos ilustrar la apariencia general de una familia de distribuciones binomiales de probabilidad. En la figura, por ejemplo, cada distribución representa n=5. A partir de la figura, podemos hacer las siguientes generalizaciones:

1. Cuando p es pequeña (0.l), la distribución binomial está sesgada hacia la derecha.

2. Conforme p aumenta (a 0.3, por ejemplo), el sesgo es menos notable.

3. Cuando p 0.5, la distribución binomial es simétrica.

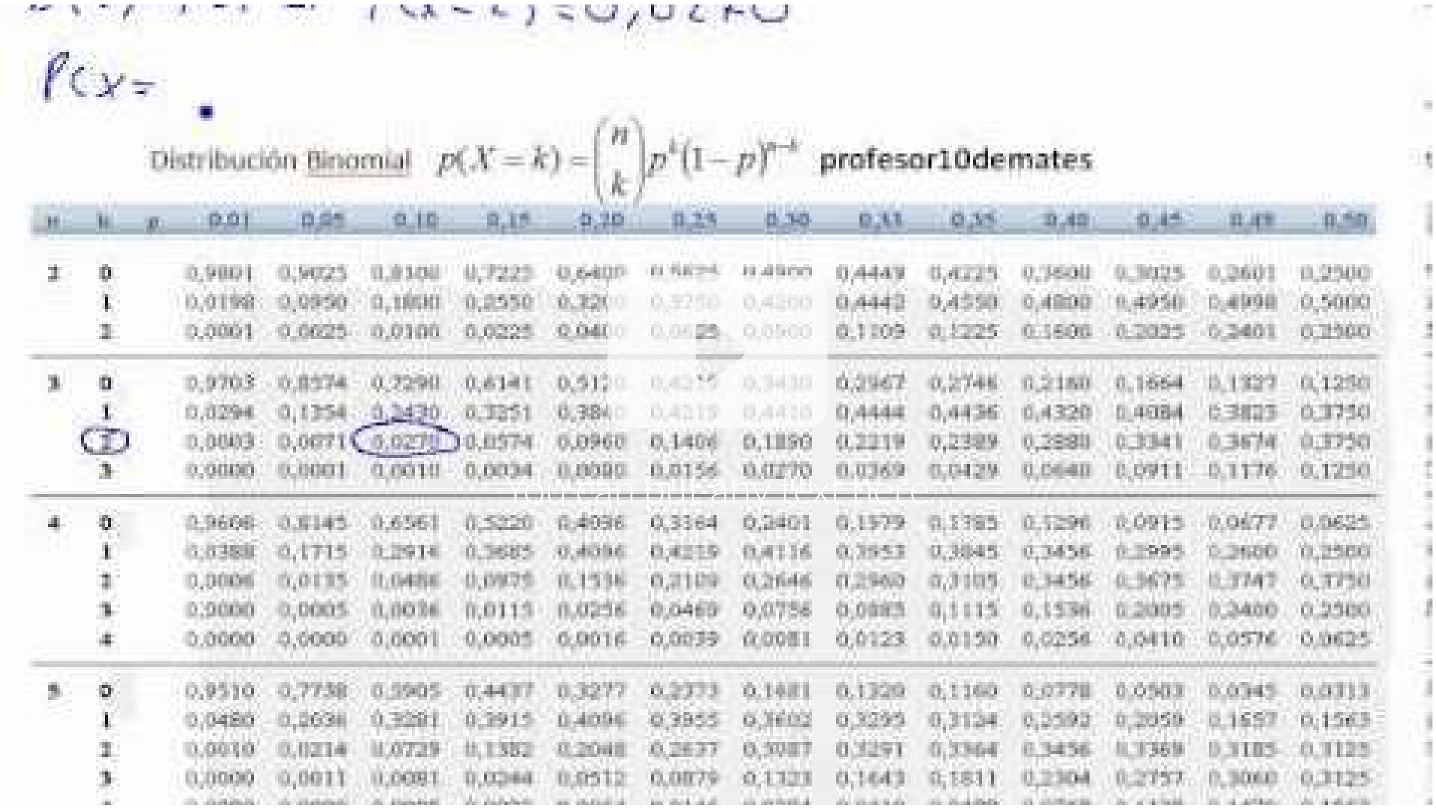
4. Cuando p es mayor que 0.5, la distribución está sesgada hacia la izquierda.

5. Las probabilidades para 0.3, por ejemplo, son las mismas para 0.7, excepto que los valores de p y q están invertidos.

Esto se aplica a cualquier pareja de valores p y q complementarios (0.3 y 0.7; 0.4 y 0.6; 0.2 y 0.8).

Tablas de distribución binomial

Uso de tablas binomiales Si quieres evitar los cálculos tediosos, puedes utilizar las tablas de distribuciones Aprende a utilizar la table de distribución binomial



Distribución binomial 07 Tabla

Distribución de Poisson

Distribución de Poisson | Ejercicios resueltos | Intro

La distribución de Poisson también sirve como modelo probabilístico para variables discretas de tipo conteo. A diferencia de la Binomial, donde el conteo se realizaba sobre n experimentos independientes, en el caso de la Poisson, los conteos se refieren al número de veces que un evento ocurre en una unidad de tiempo o espacio dada (hora, kilo, m2, m3, planta, etc.) y por tanto los valores de la variable no están acotados. Es decir, mientras los valores de Y en una Binomial podían pertenecer a los naturales entre 0 y n inclusive, en el caso de una Poisson pueden pertenecer a los naturales entre 0 e infinito.

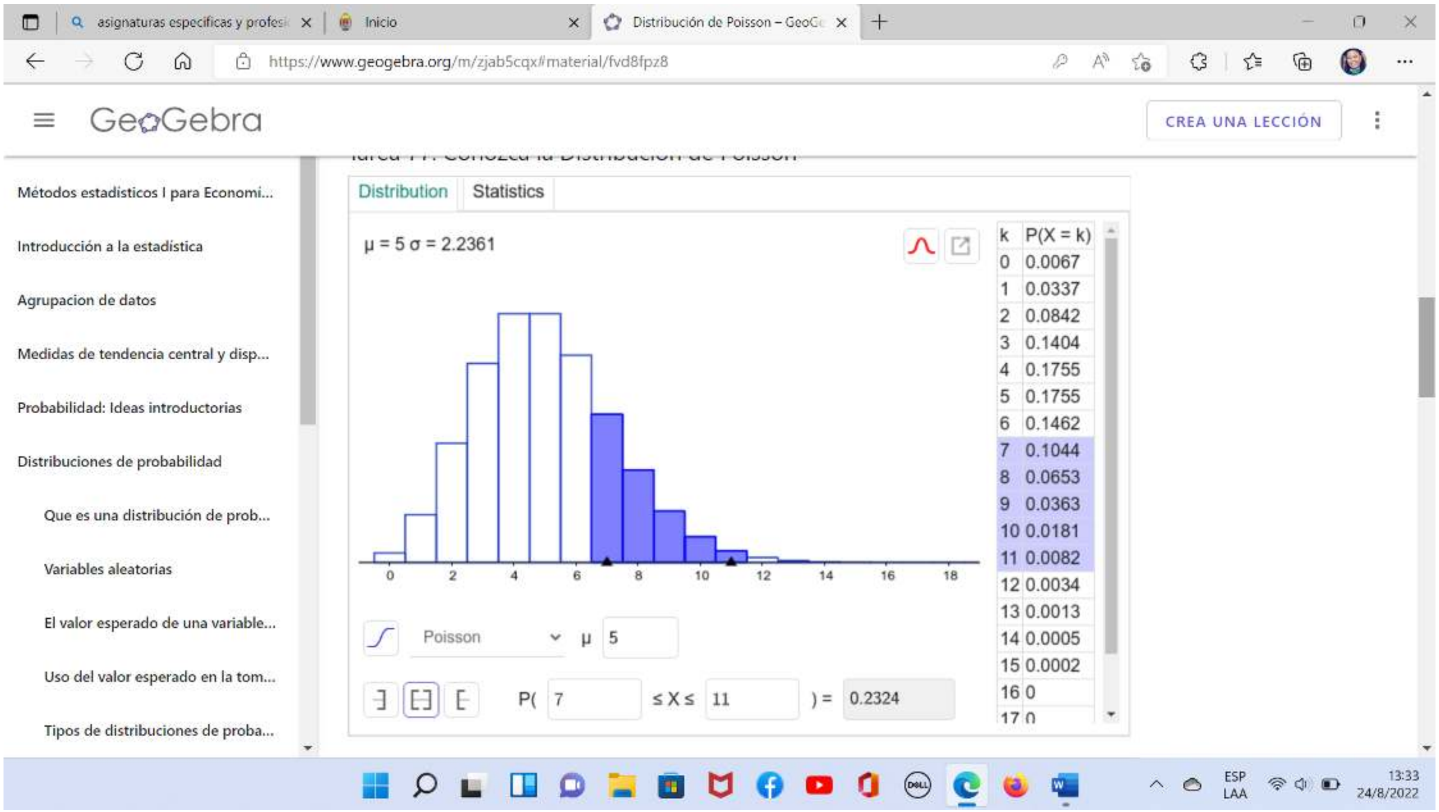
La distribución Poisson facilita el cálculo de probabilidades de variables aleatorias que provienen de conteos no acotados; mientras que la distribución binomial asigna probabilidades a variable aleatorias que cuentan la cantidad de éxitos y donde el máximo de la variable está acotado por n, el número de observaciones de tipo éxito/fracaso que se realicen.

En Agronomía, la distribución Poisson suele usarse para modelar el número de insectos sobre una planta, o en un golpe de red, el número de manchas defectuosas en un mosaico, o en un metro cuadrado de piso, el número de colémbolos en 100 g de suelo, o en 1000 cm3 de suelo o el número de coliformes en 1 ml de agua, entre otros conteos de interés.

Para ejemplificar un cálculo de probabilidad bajo el modelo Poisson, supongamos que el número promedio de picaduras de gorgojo por semilla es 0.2 (es decir, por ejemplo, que, en promedio, cada 100 semillas se cuentan 20 picaduras). El modelo Poisson podría ayudarnos a resolver estas preguntas ¿cuántas de 100 semillas no tendrán picaduras?, ¿cuántas 1 picadura? y ¿cuántas 2 o más? La fórmula que se aplica en esta distribución es

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

Donde landa es la media y x el valor para el cual encontraremos la probabilidad y e es equivalente a 2.71828



Agregar video Distribución de Poisson con tabla | Ejemplo 1 - YouTube

Distribución de Poisson con tabla | Ejemplo 1

λ = μ	4,5	5	5,5	6	6,5	7	7,5	8	8,5	9
x										
0	0,0111	0,0067	0,0041	0,0025	0,0015	0,0009	0,0006	0,0003	0,0002	0,0001
1	0,0611	0,0404	0,0266	0,0174	0,0113	0,0073	0,0047	0,0030	0,0019	0,0012
2	0,1736	0,1247	0,0884	0,0620	0,0430	0,0296	0,0203	0,0138	0,0093	0,0062
3	0,3423	0,2650	0,2017	0,1512	0,1118	0,0818	0,0591	0,0424	0,0301	0,0212
4	0,5321	0,4405	0,3575	0,2851	0,2237	0,1730	0,1321	0,0996	0,0744	0,0550
5	0,7029	0,6160	0,5289	0,4457	0,3690	0,3007	0,2414	0,1912	0,1496	0,1157
6	0,8311	0,7622	0,6860	0,6063	0,5265	0,4497	0,3782	0,3134	0,2562	0,2068
7	0,9134	0,8666	0,8095	0,7440	0,6728	0,5987	0,5246	0,4530	0,3856	0,3239
8	0,9597	0,9319	0,8944	0,8477	0,7916	0,7291	0,6620	0,5925	0,5231	0,4557
9	0,9829	0,9682	0,9462	0,9161	0,8774	0,8305	0,7764	0,7166	0,6530	0,5874
10	0,9933	0,9863	0,9747	0,9574	0,9332	0,9015	0,8622	0,8159	0,7634	0,7060
11	0,9976	0,9945	0,9890	0,9799	0,9661	0,9467	0,9208	0,8881	0,8487	0,8030
12	0,9992	0,9980	0,9955	0,9912	0,9840	0,9730	0,9573	0,9362	0,9091	[illegible]
13	0,9997	0,9993	0,9983	0,9964	0,9929	0,9872	0,9784	0,9658	0,9486	0,9261
14	0,9999	0,9998	0,9994	0,9986	0,9970	0,9943	0,9897	0,9827	0,9726	[illegible]
15	1,0000	0,9999	0,9998	0,9995	0,9988	0,9976	0,9954	0,9918	0,9862	[illegible]
16		1,0000	0,9999	0,9998	0,9996	0,9990	0,9980	0,9963	0,9934	[illegible]
17			1,0000	0,9999	0,9998	0,9996	0,9992	0,9984	0,9970	0,9947

Distribución de Chi cuadrado

Nota técnica

Si Y tiene una distribución χ^2 con 10 gl, encuentre en Excel:

- $\chi^2_{0.90} = 4.86518$
  INV.CHICUAD.CD(α;gl)
- $\chi^2_{0.10} = 4.86518$
  INV.CHICUAD(α;gl)
- $P(Y > 4.86518) = 0.90$
  DISTR.CHICUAD.CD(x;α)
- $P(Y < 4.86518) = 0.10$
  DISTR.CHICUAD(x;α)

Distribución T de student

Aplicación en Excel

- El valor t con $v = 14$ grados de libertad que deja un área de 0.025 a la izquierda, es:
  =INV.T(0.025;14)=-2.1448
- El valor t con $v = 14$ grados de libertad que deja un área de 0.975 a la derecha, es:
  =INV.T(0.975;14)=2.1448
- Sea X una va de una población normal. Se toma una muestra de 25, encuentre la $P(-1.711 < \overline{X} < 1.711) = 0.90$
  =DISTR.T.N(-1.711;24;VERDADERO)
  =DISTR.T.N(1.711;24;VERDADERO)

Distribución F

Ejercicio en Excel

Por ejemplo, si la variable F de interés tiene 5 grados de libertad en el numerador y 7 grados de libertad en el denominador, entonces

- $F_{0.1} = 2.8833$ **INV.F.CD(0.1;5;7)**
- $F_{0.025} = 5.2852$ **INV.F.CD(0.025;5;7)**
- $P(F > 5.2852) = 0.025$
  DISTR.F.CD(5.2852;5;7)
- $P(F < 5.2852) = 0.975$
  DISTR.F.N(5.2852;5;7;VERDADERO)

La distribución normal: distribución de una variable aleatoria continua

Esta distribución es la más utilizada en las ciencias biológicas, agronómicas y forestales ya que usualmente ajusta bien histogramas de frecuencias de variables como el peso y la altura de seres vivos, así como otras mediciones morfométricas además del rendimiento. Estas características, particularmente interesantes en agronomía, son producidas por el resultado de la acción conjunta de muchos factores y por tanto asumen muchos valores distintos (en un continuo de valores posibles) entre las unidades de análisis. No obstante, algún valor o intervalo de valores se repite con mayor frecuencia, mientras que otros muy alejados de estos valores centrales (por ser mucho mayores o menores) aparecen con menor frecuencia. La distribución normal se usa para el **cálculo de probabilidades de variables continuas**, cuyos histogramas tienen forma "acampanada", por eso y porque su expresión matemática fue estudiada por Gauss, también se conoce como modelo Gaussiano.

La distribución de frecuencias de esta variable tiene ciertas características: es aproximadamente **simétrica**, posee una gran cantidad de valores cerca del centro. La media, la moda y la mediana son prácticamente iguales y los valores extremos, tanto inferiores como superiores, tienen menor frecuencia de ocurrencia que los valores centrales. Además, la distribución es simétrica, es decir con distribución de valores superiores a la media igual a la de valores por debajo de la media.

La localización del centro de la campana está dada por el parámetro μ (también conocido como esperanza de Y) y la mayor o menor amplitud de la campana viene dada por el parámetro (la varianza de Y en la población). Como la función es simétrica respecto a la media, ésta **divide a la gráfica en partes iguales**. Está definida para valores en la abscisa que tienden a infinito y a menos infinito, se aproxima al eje horizontal sin tocarlo (curva asintótica). Como toda función de densidad, el área comprendida entre el eje de las abscisas y la curva es igual a la unidad. La función de densidad de una variable aleatoria normal tendrá distintas formas dependiendo de sus parámetros que son la esperanza y varianza.

Introducción a la estadística

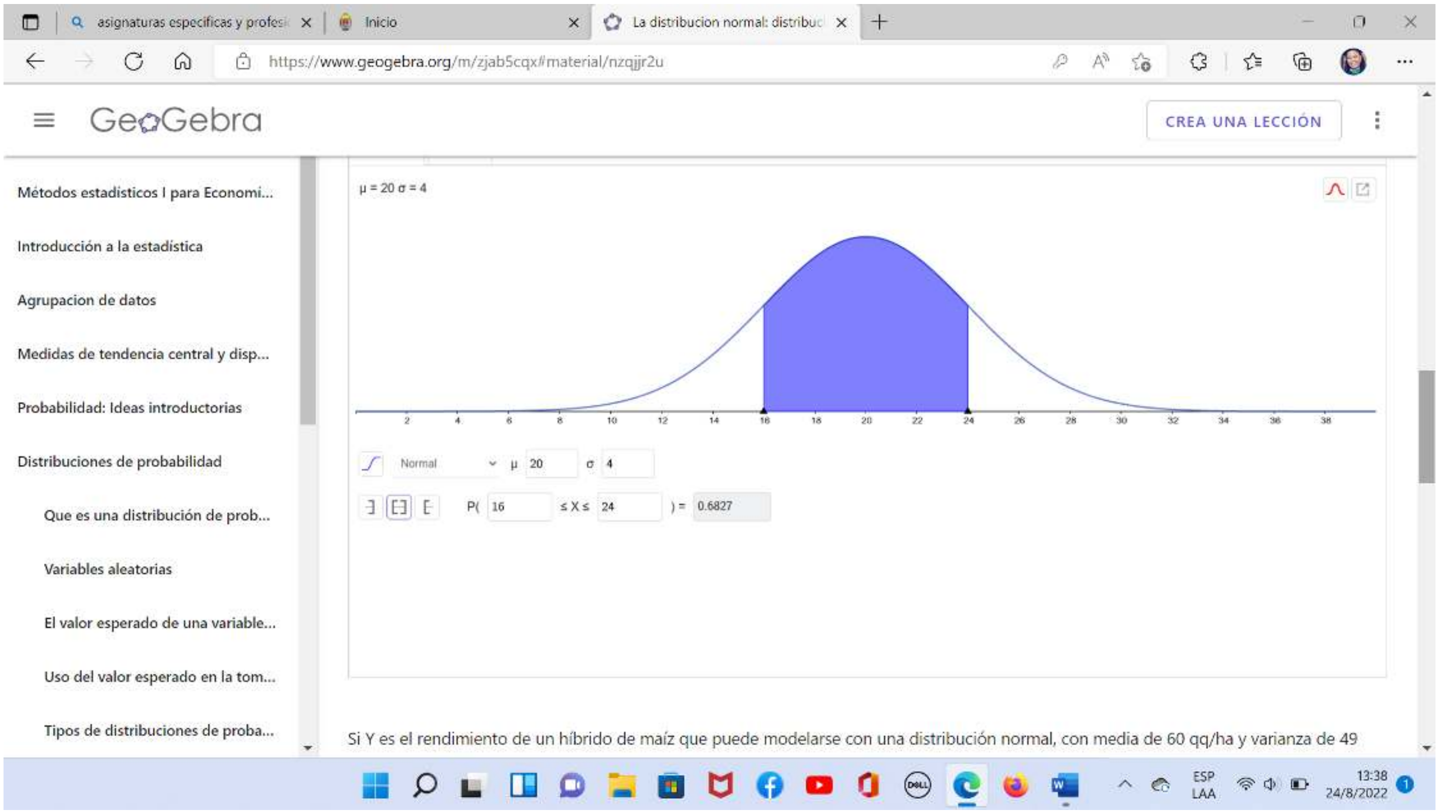
La distribución normal es un modelo de probabilidad y una vez adoptado el modelo es posible responder a las siguientes preguntas: - ¿Cuál es la probabilidad de que la variable en estudio tome valores menores a un valor determinado?

Por ejemplo, si la variable es el rendimiento de un cultivar, el responder a esta pregunta podría indicar la posibilidad de obtener rendimientos que no justifiquen el costo de producción. - ¿Cuál es la probabilidad de que la variable en estudio tome valores mayores a un valor determinado? Si la variable aleatoria en estudio es la cantidad de semillas de maleza en el suelo antes de la siembra, el responder a esta pregunta podría indicar si se necesitará o no aplicar herbicida (este podría ser el caso de modelación de una variable aleatoria discreta como si se tratara de una continua).

- ¿Cuál es la probabilidad de que la variable en estudio tome valores entre 2 valores determinados? Esta probabilidad es de interés, por ejemplo, al clasificar tubérculos de papa dado que aquellos con volumen entre 59 cm3 y 80 cm3 son considerados de valor comercial.

Uso de la tabla Z

Hallar areas bajo la curva de una distribucion normal - YouTube



Resumen de distribuciones de probabilidad y uso de las tablas

CAPITULO VI

Muestreo y distribuciones de muestreo

Introducción al muestreo

Los especialistas en estadística usan la palabra **población** para referirse no sólo a personas sino a **todos los elementos que han sido escogidos para su estudio** y la palabra **muestra** para describir una **porción escogida de la población**. Matemáticamente, podemos describir muestras y poblaciones al emplear mediciones como la media, la mediana, la moda y la desviación estándar. Cuando estos términos describen las características de una muestra, se denominan **estadísticas**. Cuando describen las características de una población, se llaman **parámetros**. Una estadística es una característica de una muestra y un parámetro es una característica de una población. Si estamos convencidos de que la edad promedio de los alumnos de tercer año de economía agrícola es una estimación exacta de la edad promedio de todos los alumnos de tercer año del CURLA, podríamos usar la **estadística de muestra** "edad promedio de los alumnos de economía agrícola" para estimar el **parámetro de población** "edad media de los alumnos de tercer año del CURLA", sin tener que entrevistar a todos los alumnos del CURLA. Para ser consecuentes, los especialistas en estadística emplean literales latinas minúsculas para representar estadísticas de muestra, y literales griegas o latinas mayúsculas para representar parámetros de población.

Repasar tipos de muestreo

Población y muestra

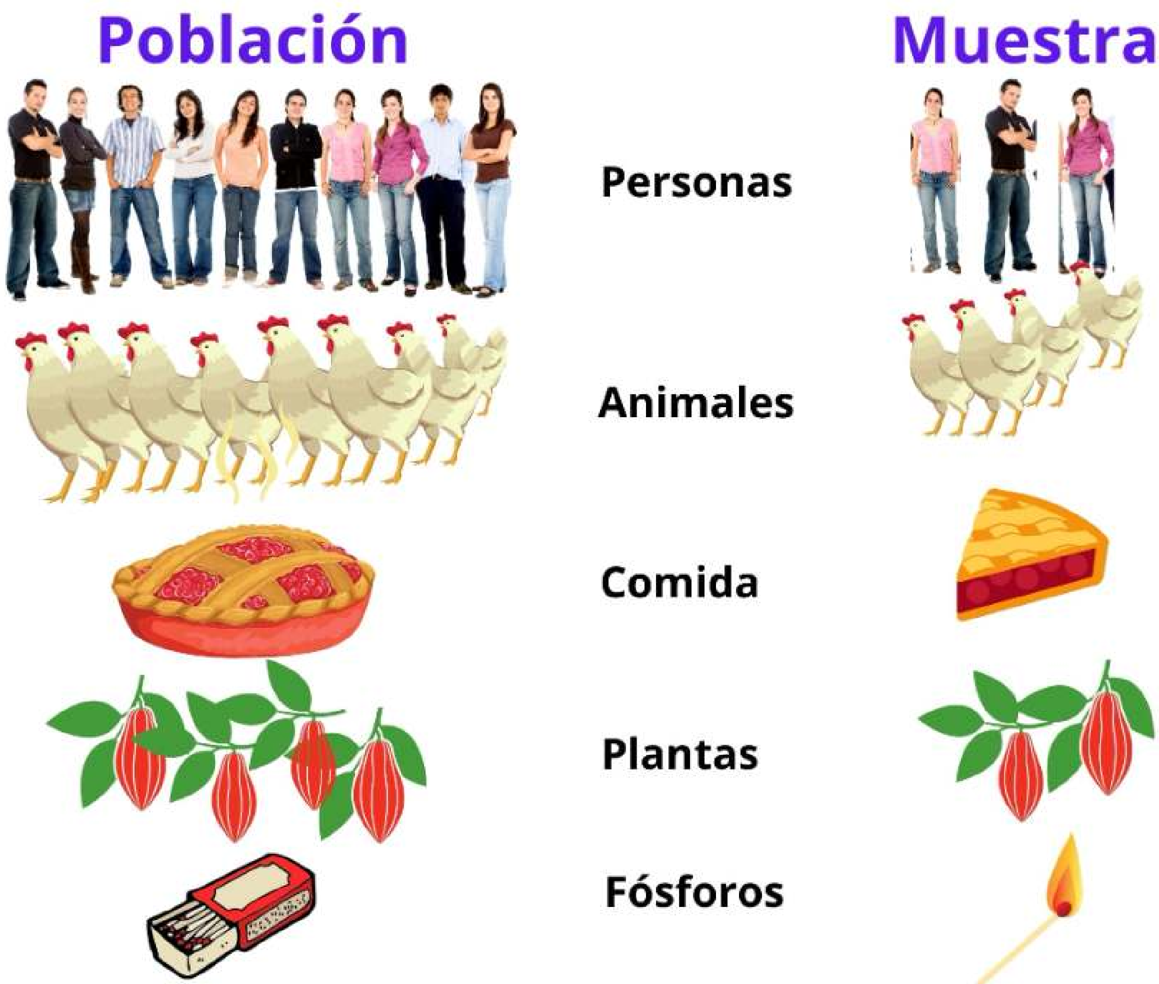
Población

Es un conjunto de elementos acotados en un tiempo y en un espacio determinados, con alguna característica común observable o medible que es interés de estudio en una investigación. El tamaño de la población, N, puede ser muy grande o incluso infinito. Ahora bien, una población no tiene por qué estar formada por personas. Pueden ser plantas, envases de leche, semillas en fin cualquier cosa de la cual podamos medir alguna característica, a la cual le vamos a llamar variable aleatoria. En una población de personas podemos medir ciertas características, su altura su peso su sexo femenino o masculino el número de hermanos que tiene la ciudad donde viven todos esos son ejemplos de variables aleatoria. Otros elementos considerados podrían ser días, animales, semillas, plantas, personas o localidades de una cierta región.

Tamaño poblacional Si la población es finita, diremos que el tamaño poblacional es el número de elementos de esta y lo denotaremos con **N**. Generalmente es imposible o impracticable examinar alguna característica en la población entera, por lo que se examina una parte de ella y en base a la información relevada en esa porción se hacen inferencias sobre toda la población.

Muestra Una muestra es un subconjunto observado de valores poblacionales que tiene un tamaño muestral que viene dado por **n**. **Unidad muestral** Es el elemento o entidad de la muestra. **Tamaño muestral** Es el número de elementos de la población que conforman la muestra y se denota con **n**. **Parámetro** Se llama así a cualquier valor característico de una población como ser el peso promedio, la altura máxima o el estado civil más frecuente. Este valor es una constante. **Estadígrafo o estadístico** Es un valor característico obtenido a partir de una muestra. Esta cantidad es variable en cada muestra tomada de una población.

Introducción a la estadística



video

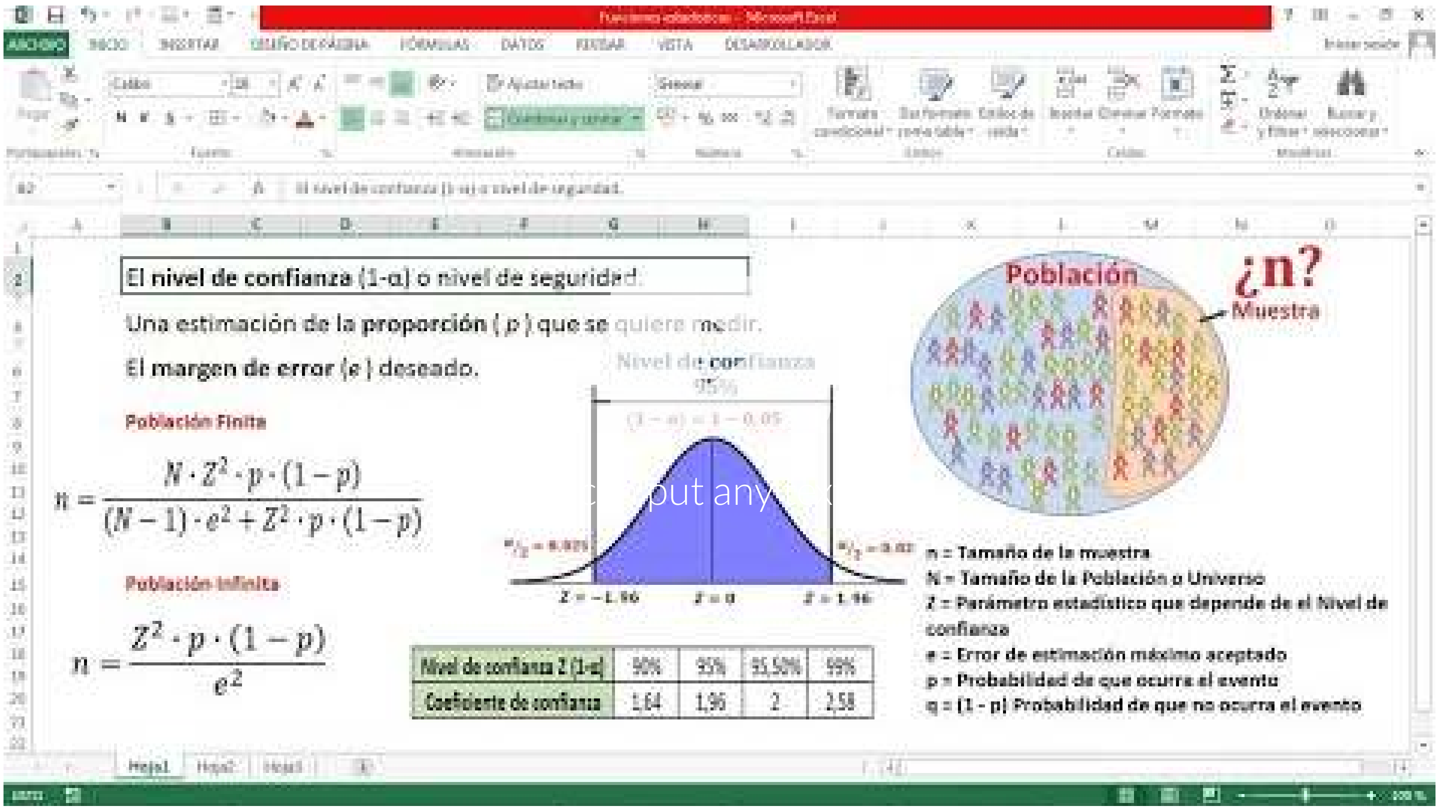
Después de ver el siguiente Video Poblacion y muestra responde las preguntas: Se quiere conocer cuál es el candidato presidencial por el cual votarán los hondureños en las próximas elecciones, para lo cual se toma una muestra de 3500 hondureños, y se les consulta por quién votarán en las próximas elecciones presidenciales. Pregunta. ¿Cuál es la población en este estudio?

Ejemplo de población y muestra

Observa el siguiente applet donde de una población de 100 elementos **(N=100)** se extraen diferentes muestras de tamaño 20 **(n=20)**. Actualiza el applet para observar que ocurre con las diferentes muestras. Puedes imaginar que es un campo de cultivos con 100 árboles y que solo necesitas tomar 20 de ellos. O una universidad con 100 estudiantes en la cual tomarás una muestra al azar de 20 alumnos para una investigación. Cada vez que actualizas los datos obtienes una muestra diferente y por lo tanto los valores de la muestra cambian. Ejemplo de población y muestra – GeoGebra

Tamaño de la muestra

Tamaño de la muestra – GeoGebra



https://www.youtube.com/watch?v=vqAjfz_0hOE&feature=emb_imp_woyt

Tipos de muestreo (repaso)

https://www.youtube.com/watch?t=403&v=-i9Lmuqhg90&feature=emb_imp_woyt

Muestreo aleatorio

En numerosas situaciones deseamos utilizar los resultados del análisis de datos muestrales para elaborar conclusiones que puedan ser extendidas a la población de la que proviene la muestra. A este proceso inductivo se lo denomina **Inferencia Estadística**. Si la muestra es una ventana a través de la cual observamos a la población podemos asegurar que aquello que vemos en la muestra está presente en la población; pero no podemos decir que aquello que no vemos, no está presente.

Esto sugiere que, si **toda muestra contiene una parte de la población**, dos muestras de una misma población podrían "mostrar" cosas diferentes e inclusive puede que la diferencia sea muy grande. ¿Cómo decidir en qué muestra confiaremos? ¿Podemos otorgar una medida de confiabilidad al cálculo obtenido en una muestra, para así establecer una medida del error potencial que podríamos tener al concluir sobre la población, de la mano de la muestra? Vemos que inferir acerca de una población en base a lo observado en solo una de las posibles muestras, implica riesgo: el riesgo de concluir erróneamente por haber seleccionado una muestra que no represente adecuadamente a la población, ya que **existe la posibilidad de que la estimación no sea buena por errores aleatorios debidos al muestreo**. En este sentido, se hace necesario conocer el comportamiento de los **estadísticos** obtenidos en las posibles muestras; es decir, conocer su **distribución en el muestreo**.

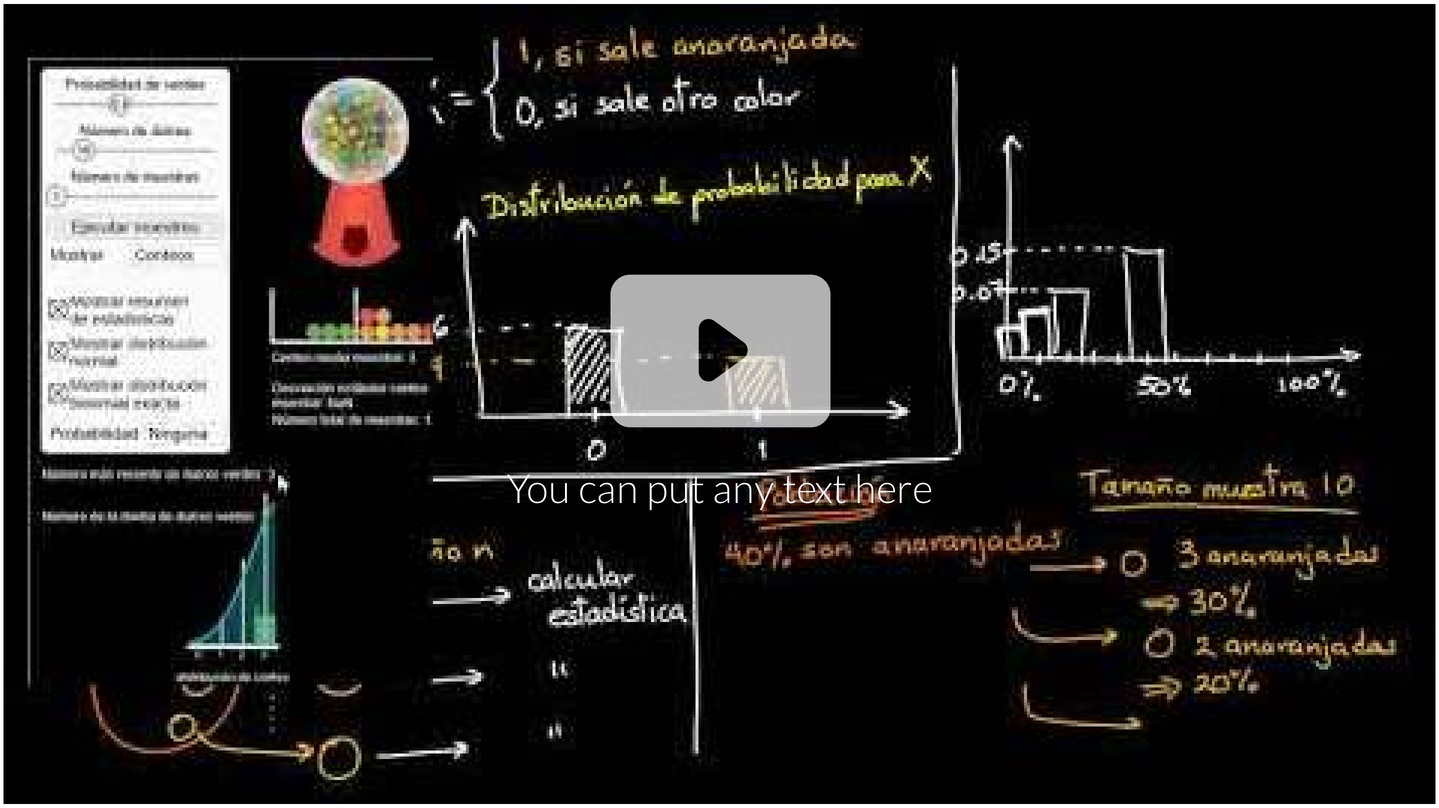
En este tema abordaremos las distribuciones de los estadísticos: **media y varianzas muestrales** y el **Teorema Central del Límit**e, que da sustento a las conclusiones que se obtienen en los estudios que se realizan con muestras.

Se desea un muestreo de 15 páginas del libro de texto. Use el generador de números aleatorios para seleccionar 15 páginas aleatorias y cuente el número de palabras en letras cursivas en cada página. Presente sus resultados. Utilice Excel.

Introducción a las distribuciones de muestreo

Una distribución de probabilidad de todas las medias posibles de las muestras es una distribución de las medias de las muestras. Los especialistas en estadística la conocen como distribución de muestreo de la media.

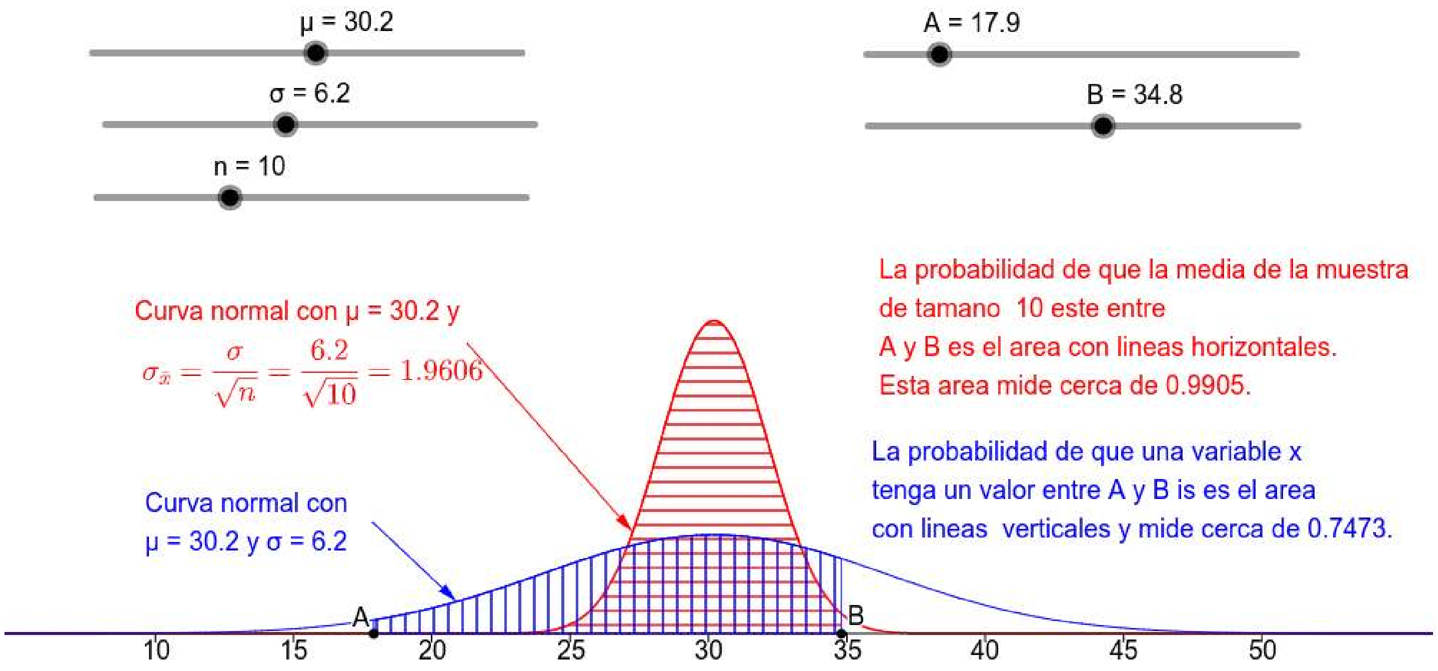
https://www.youtube.com/watch?v=PSO9WdraWrE&feature=emb_imp_woyt



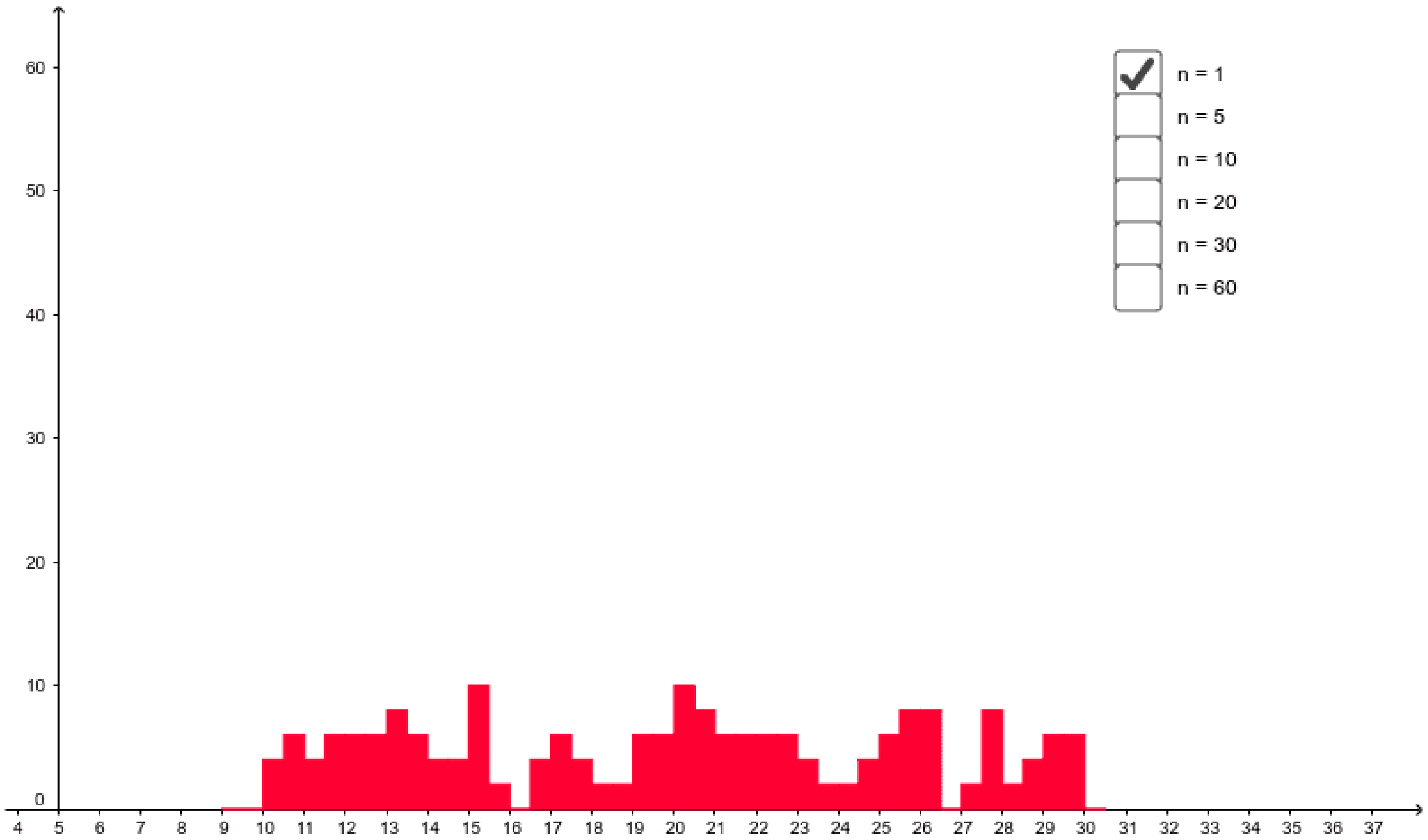
Teorema del límite central

El teorema del límite central es, tal vez, el más importante de toda la inferencia estadística, pues asegura que **la distribución de muestreo de la media se aproxima a la normal al incrementarse el tamaño de la muestra.** Hay situaciones teóricas en las que el teorema del límite central no se cumple, pero casi nunca se encuentran en la toma de decisiones práctica. De hecho, una muestra no tiene que ser muy grande para que la distribución de muestreo de la media se acerque a la normal. Los especialistas en estadística **utilizan la distribución normal como una aproximación a la distribución de muestreo siempre que el tamaño de la muestra sea de al menos 30**, pero la distribución de muestreo de la media puede ser casi normal con muestras de incluso la mitad de ese tamaño. La importancia del teorema del límite central es que nos permite **usar estadísticas de muestra para hacer inferencias con respecto a los parámetros de población**, sin saber sobre la forma de la distribución de frecuencia de esa población más que lo que podamos obtener de la muestra.

https://profefily.com/wp-content/uploads/2017/12/Estad%C3%ADstica-para-administraci%C3%B3n-y-economia-Richard-I.-Levin.pdf



Esta hoja de trabajo demuestra el Teorema del límite central usando una distribución de probabilidad uniforme donde la variable aleatoria X tiene la misma probabilidad de tomar un valor entre 10 y 30. Lo que está viendo: se extraen 100 muestras aleatorias de la distribución. Se calcula la media de cada muestra. Se construye una distribución de frecuencia de las 100 medias muestrales. Utilice las casillas de verificación para mostrar las distribuciones de frecuencia para diferentes tamaños de muestras de n = 1 a n = 60.



¿Qué sucede con la forma de la distribución de frecuencias a medida que aumenta el tamaño de la muestra, específicamente la altura del pico y el ancho de la distribución?

Error estándar

En vez de decir "la desviación estándar de la distribución de las medias de la muestra" para describir una distribución de medias de la muestra, los especialistas en estadística se refieren al error estándar de la media. De manera similar, la "desviación estándar de la distribución de las proporciones de la muestra" se abrevia como error estándar de la proporción. EL término error estándar se utiliza porque da a entender un significado específico. Supongamos que deseamos saber algo sobre la edad promedio de los estudiantes de Economía Agrícola. Podríamos tomar una serie de muestras y calcular la edad promedio de cada muestra.

Es altamente improbable que todas estas medias de muestra fueran iguales; es de esperar alguna variabilidad en las medias observadas. Esta variabilidad en las estadísticas de muestras proviene de un error de muestreo debido al azar; es decir; hay diferencias entre cada muestra y la población, y entre las diversas muestras, debido únicamente a los elementos que decidimos escoger para las muestras. La desviación estándar de la distribución de las medias de las muestras mide el grado hasta el cual es de esperar que varíen las medias de las diferentes muestras, debido a este error cometido en el proceso de muestreo.

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Por tanto, la desviación estándar de la distribución de una estadística de muestra se conoce como error estándar de la estadística. El error estándar indica no sólo el tamaño del error al azar que se ha cometido, sino también la probable precisión que puede obtenerse al utilizar una estadística de muestra para estimar un parámetro de población. Una distribución de medias de muestra que está menos extendida (y que tiene un error estándar pequeño) constituye una mejor estimación de la media de la población que una distribución de medias de muestra que está ampliamente dispersa y que tiene un error estándar más grande. El error estándar se calcula como Se debe destacar el hecho de que la varianza de las medias muestrales es inversamente proporcional al tamaño de la muestra. Esto tiene un importante resultado práctico y es que a través del tamaño muestral se puede controlar la variabilidad de la media resultante. Consecuentemente, si la muestra es grande es menos probable que se obtenga una media muestral muy alejada de la esperanza de la distribución que se está muestreando. Cuando la población es finita a la formula anterior se le agrega el multiplicador de población finita, quedando de la siguiente manera:

Usted acaba de comprar una caja de cereal con pasas y cuenta el número de pasas. La compañía afirma que la cantidad de pasas por caja es, en promedio, de 2.0 tasas, con una desviación estándar de 0.2 tasas. Su caja contenía sólo 1.9 tasas. ¿Puede la compañía asegurar que afirma lo correcto?

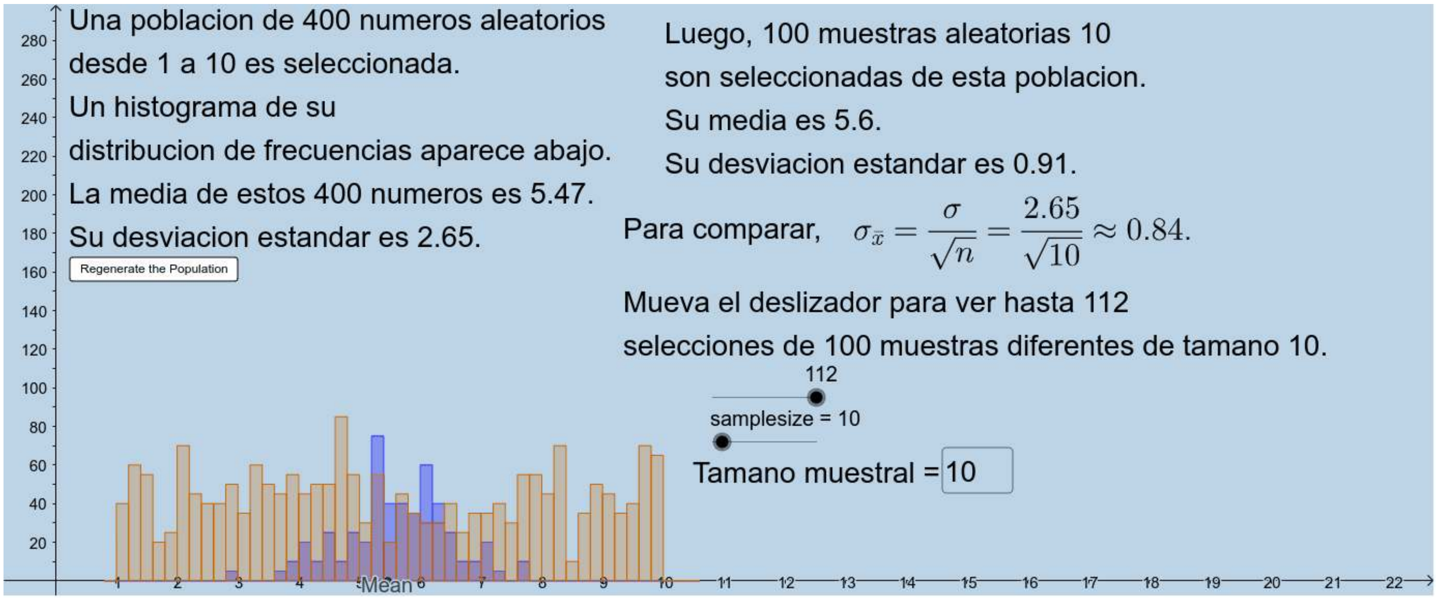
Relación entre el tamaño de la muestra y el error estándar

El error estándar , es una medición de dispersión de las medias de muestras alrededor de la media de población μ. Si la dispersión disminuye (si , se hace más pequeña), entonces los valores tomados por la media de la muestra tienden a agruparse más cercanamente alrededor de μ. Por el contrario, si la dispersión se incrementa (si , se hace más grande), los valores tomados por la media de la muestra tienden a agruparse menos cercanamente alrededor de μ. Podemos concebir esta relación así: al disminuir el error estándar, el valor de cualquier media de muestra probablemente se acercará al valor de la media de población. Los especialistas en estadística describen este fenómeno de otra manera: **al disminuir el error estándar, se incrementa la precisión con la que se puede usar la media de muestra para estimar la media de población.**

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

En el siguiente ejercicio tenemos una población de 400 elementos con media de 5.5 y desviación estándar de 2.69. Dado que la fórmula del error estándar es: a medida que aumente el tamaño de la muestra n el error estándar va a disminuir.

Cambie a diferentes tamaños de muestra y vea lo que ocurre con el error estándar para comparar



¿Cuánto es el error estándar cuando el tamaño de la muestra es de 10?

Al aumentar nuestro tamaño de muestra de 10 a 100), el error estándar disminuyó de 0.85 a 0.27, lo que es sólo aproximadamente un tercio de su valor inicial. Nuestros ejemplos muestran que, debido al hecho de que x, varía inversamente con la raíz cuadrada de n, hay una utilidad decreciente en el muestreo.

Es cierto que muestrear más elementos disminuye el error estándar, pero este beneficio puede no valer el costo. Un estadístico diría: “El aumento de precisión no vale el costo del muestreo adicional”. En un sentido estadístico, rara vez vale la pena tomar muestras excesivamente grandes. Los administradores debieran evaluar siempre tanto el valor como el costo de la precisión adicional que obtendrían de una muestra mayor antes de comprometer recursos para tomarla.

Juanita Martínez, investigadora de la Colombian Cofee Corporation, está interesada en determinar la tasa de uso de café por hogar en Estados Unidos. Ella cree que el consumo anual por hogar tiene distribución normal con media desconocida y desviación estándar cercana a 1.25 libras.

a) Si Juanita toma una muestra de 36 hogares y registra su consumo de café durante un año, ¿Cuál es la probabilidad de que la media de la muestra se aleje de la media de la población no más de media libra?

b) ¿Qué tan grande debe ser la muestra que tome para tener el 98% de certidumbre de que la media de la muestra no se aleja más de media libra de la media de la población?

Tipos de muestreo

Muestreo aleatorio

El muestreo aleatorio simple selecciona muestras mediante métodos que permiten que cada posible muestra tenga una igual probabilidad de ser seleccionada y que cada elemento de la población total tenga una oportunidad igual de ser incluido en la muestra. Supongamos que tenemos una población de 20 estudiantes en la clase de métodos estadísticos I y queremos muestras de cinco estudiantes cada vez para entrevistarlos. Cada vez que selecciones esta opción en el applet, diferentes estudiantes aparecerán en la muestra. Realiza el ejercicio.

Muestreo estratificado

Para utilizar el muestreo estratificado, dividimos la población en grupos relativamente homogéneos, llamados estratos. Después utilizamos uno de los dos planteamientos: o bien seleccionamos aleatoriamente, en cada estrato, un número específico de elementos correspondiente a la proporción de este en relación con la población completa, o extraemos el mismo número de elementos de cada estrato y después ponderamos los resultados considerando la proporción que el estrato representa con respecto a la población total. Con cualquiera de los planteamientos, el muestreo estratificado garantiza que cada elemento de la población tenga posibilidad de ser seleccionado. El muestreo estratificado resulta apropiado cuando la población ya está dividida en grupos de diferentes tamaños y deseamos tomar en cuenta esta condición.

Muestreo por conglomerado o racimo

En el muestreo de racimo dividimos la población en grupos, o racimos, y luego seleccionamos una muestra aleatoria de estos racimos, asumiendo que cada uno de ellos es representativo de la población. Si una investigación de mercado tiene la intención de determinar por muestreo el número promedio de televisores por casa en una ciudad grande, podrían usar un mapa de la ciudad para dividir el territorio en manzanas y luego escoger un cierto número de éstas (racimos) para entrevistar a sus habitantes. Cada casa perteneciente a cada una de estas manzanas sería considerada para entrevistar a sus habitantes. Un procedimiento de muestreo de racimo bien diseñado puede producir una muestra más precisa a un costo considerablemente menor que el de un muestreo aleatorio simple.

Tanto en el muestreo estratificado como en el de racimo, la población se divide en grupos bien definidos. Usamos el muestreo estratificado cuando cada grupo tiene una pequeña variación dentro de sí mismo, pero hay una amplia variación de un grupo a otro. Usamos el muestreo de racimo en el caso opuesto, cuando hay una variación considerable dentro de cada grupo, pero los grupos son esencialmente similares entre sí.

Muestreo sistemático

En el muestreo sistemático, los elementos son seleccionados de la población dentro de un intervalo uniforme que se mide con respecto al tiempo, al orden o al espacio. Si tuviera que entrevistar a cada quinto estudiante de la clase de 20 estudiantes, escogería un punto de inicio aleatorio entre los primeros 5 nombres del listado estudiantil y luego seleccionaría cada quinto nombre de ahí en adelante. El muestreo sistemático difiere del muestreo aleatorio simple en que cada elemento tiene igual oportunidad de ser seleccionado, pero cada muestra no tiene una posibilidad igual de ser seleccionada. Aun cuando este tipo de muestreo puede ser inapropiado cuando los elementos entran en un patrón secuencial, este método puede requerir menos tiempo y, algunas veces, tiene como resultado un costo menor que el método de muestreo aleatorio simple.

CAPITULO VII

Estimaciones

La estadística inferencial

Pocas veces cuando se hace investigación se trabaja con toda la población. Generalmente usamos muestras y a partir de ellas sacamos conclusiones con respecto a la población. No utilizamos a toda la población por diversas razones. El tiempo que tomaría podría ser excesivo, al igual que el costo. Otras veces resulta imposible sin dañar a la población. *Esto me recuerda la anécdota de la madre que mando a su hijo a comprar cerillos y al llegar a casa le pregunta si se aseguró que lo fósforos están buenos. y él contesta, si mama, los probé y todos encendieron.*

Al igual que en este caso muchas veces tomar toda la población haría que esta desapareciera. Si la muestra que se toma es representativa, bastara con una muestra significativa para poder sacar conclusiones sobre la población. La estadística inferencial es tema para la clase de Métodos estadísticos II.

Objetivos

Aprender cómo hacer estimaciones de parámetros a partir de una muestra

Aprender ventajas y desventajas de las estimaciones puntuales y de intervalo

Calcular que tan precisas son las estimaciones

Tipo de estimaciones

Se conocen dos tipos de estimación de parámetros de población a partir de una muestra: a) Estimaciones puntuales. Se trata de un solo valor que estima un parámetro desconocido de la población. Ejemplo de estimaciones puntuales:

- Llegare a las 4 en punto.
- Para el tercer periodo del 2022 esperamos 36 nuevos estudiantes
- El promedio de edad de los estudiantes de Economía Agrícola es de 23.5 años
- 70% de los estudiantes están a favor del cambio de plan de estudios

b) Estimaciones de intervalo. Se trata de un rango de valores que predice entre cuales valores se encuentra el verdadero parámetro de la población.

- Llegare entre 3:55 y 4:05
- Para el tercer periodo del 2022 esperamos entre 32 y 40 estudiantes
- La edad promedio de los estudiantes de Economía Agrícola esta entre los 22 y los 25 años
- Entre el 68% y el 72% de los estudiantes están a favor de un cambio en el plan de estudios.

Una estimación puntual muchas veces no es suficiente, pues es más fácil equivocarse. Es útil acompañarla del error implicado en la estimación. Por su parte en la estimación de intervalo se puede indicar el error de 2 maneras: Por lo ancho del intervalo y por la probabilidad de que el verdadero parámetro se encuentre dentro de la estimación de intervalo.

Mueve las piezas (valores de media, desviación y tamaño de la muestra)

PROYECTO FINAL DE CURSO

Trabajo en equipo (2 estudiantes)