Componente formativo

Procesamiento y análisis de datos

Breve descripción:

Un sistema de gestión eficiente es cuando los indicadores obtenidos son el resultado de un adecuado procesamiento y análisis de los datos. Es necesario desarrollar conocimiento y habilidades en diferentes métodos de procesamiento y análisis de datos para tener gran variedad al momento de determinar cuál es el adecuado dependiendo de los requerimientos del sistema.

Área ocupacional:

Procesamiento, fabricación y ensamble

Junio 2023

Grupo de Ejecución de la Formación Virtual

Tabla de contenido

Introducción

El procesamiento y el análisis de datos conllevan, anticipadamente, procedimientos de recopilación, limpieza y clasificación de estos, para luego llevar a cabo la transformación y el modelado de datos, bajo diferentes métodos tanto cualitativos como cuantitativos con el objetivo de descubrir la información requerida.

Los resultados así obtenidos se comunican, sugiriendo conclusiones y apoyando la toma de decisiones. La visualización de datos se utiliza, a veces, para representar los datos para facilitar el descubrimiento de patrones útiles en los datos. Los términos Modelado de datos y Análisis de datos significan lo mismo.

Video 1.

SINTESIS DEL VIDEO

1. Probabilidad y estadística

La estadística forma parte del comportamiento histórico de todos los fenómenos naturales que ocurren en el mundo:

El procesamiento y el análisis de datos conllevan, anticipadamente, procedimientos de recopilación, limpieza y clasificación de estos, para luego llevar a cabo la transformación y el modelado de datos, bajo diferentes métodos tanto cualitativos como cuantitativos con el objetivo de descubrir la información requerida.

Los resultados así obtenidos se comunican, sugiriendo conclusiones y apoyando la toma de decisiones. La visualización de datos se utiliza, a veces, para representar los datos

para facilitar el descubrimiento de patrones útiles en los datos. Los términos Modelado de datos y Análisis de datos significan lo mismo.

Moda, media, mediana y desviación típica, estudio de variables continuas.

La media, la mediana y la moda son tres tipos de "promedios". Hay muchos "promedios" en estadística, pero estos son los tres más comunes, y ciertamente son los tres más utilizados al momento de llevar a cabo aplicaciones y procesos estadísticos.

Media

La media resume un conjunto de datos completo con un solo número que representa el punto central de los datos o el valor típico. También se conoce como promedio aritmético y es una de varias medidas de tendencia central.

Encontrar la media es muy sencillo. Simplemente sume todos los valores y divídalos por el número de observaciones; la fórmula está a continuación.

$$\frac{X_1+X_2+...+X_n}{n}$$

Por ejemplo, si las alturas de cinco personas son 48, 51, 52, 54 y 56 pulgadas, su altura promedio es de 52,2 pulgadas.

$$48 + 51 + 52 + 54 + 56/5 = 52,2$$

Idealmente

La media indica la región donde caen la mayoría de los valores en una distribución. Los estadísticos se refieren a ella como la ubicación central de una distribución. Puede pensar en ello como la tendencia de los datos a agruparse en torno a un valor medio. Sin

embargo, la media no siempre encuentra el centro de los datos. Es sensible a datos sesgados y valores extremos. Por ejemplo, cuando los datos están sesgados, pueden fallar.

Media aritmética (AM)

La media aritmética (o simplemente media) de una lista de números, es la suma de todos los números divididos por el número de números. De manera similar, la media de una muestra x1, x2, ..., xn, generalmente denotado por es la suma de los valores muestreados dividida por el número de elementos de la muestra:

$$\overline{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Por ejemplo, la media aritmética de cinco valores: 4, 36, 45, 50, 75 es:

$$\frac{4 + 36 + 45 + 50 + 75}{5} = \frac{210}{5} = 42$$

Media geométrica (GM)

La media geométrica es una media útil para conjuntos de números positivos, que se interpretan según su producto (como es el caso de las tasas de crecimiento) y no su suma (como es el caso de la media aritmética):

$$\overline{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \left(x_1 x_2 \ldots x_n\right)^{\frac{1}{n}}$$

Por ejemplo, la media geométrica de cinco valores: 4, 36, 45, 50, 75 es:

$$(4 \times 36 \times 45 \times 50 \times 75)^{\frac{1}{5}} = \sqrt[5]{24\,300\,000} = 30.$$

Media armónica (HM)

La media armónica es un promedio que es útil para conjuntos de números que se definen en relación con alguna unidad, como en el caso de la velocidad (es decir, distancia por unidad de tiempo):

$$\bar{x} = n\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$$

Por ejemplo, la media armónica de los cinco valores: 4, 36, 45, 50, 75 es:

$$\frac{5}{\frac{1}{4} + \frac{1}{36} + \frac{1}{45} + \frac{1}{50} + \frac{1}{75} +} = \frac{5}{\frac{1}{3}} = 15$$

Relación entre AM, GM y HM

AM, GM y HM satisfacen estas desigualdades:

$$AM \geq GM \geq HM$$

La igualdad se mantiene si todos los elementos de la muestra dada son iguales.

1.1 Moda, media, mediana y desviación típica, estudio de variables continuas

A continuación, se definirá la moda, media, mediana y desviación típica, así como el estudio de variables continuas.

Mediana

La mediana es la observación intermedia en un conjunto de datos. Para contextualizar qué significa, se calcula la mediana de un conjunto de datos de muestra sobre el peso infantil, así:

13 36 98 77 42
50110 22 49 81 26 38

En el conjunto de datos anterior, ¿cuál es la observación intermedia? Bueno, antes que pueda resolver esto, tenemos que ordenar correctamente las observaciones de una manera lógica para que tengan sentido. Los ordenaremos de menor a mayor, como se muestra a continuación:

13 22 26 36 38 42
49 50 77 81 98 110

Ahora que nuestros datos están ordenados correctamente, podemos encontrar la observación intermedia.

Número impar de observaciones

En un conjunto de datos que tiene un número impar de observaciones, esto es muy fácil; es simplemente el número justo en el medio (el que tiene el mismo número de observaciones arriba y abajo).

13	22	26	36	38	42	49	50	77	81	98
				MEDIANA						
				42						

$$Mediana\ (X) = X\ \frac{N+1}{2}$$

Número par de observaciones

Sin embargo, en nuestro caso tenemos 12 observaciones, que es un número par. Esto significa que debemos tomar las dos observaciones del centro y promediarlas. En este caso, las dos observaciones en el medio son 42 y 49. Cuando tomamos el promedio de estos dos números (recuerde, para hacer un promedio, sume los dos números (42 + 49 = 91) y divida ese número por la cuenta, que en este caso es 2), obtenemos 45,5. Entonces nuestra mediana es 45,5.

13	22	26	36	38	42	49	50	77	81	98	110
					MEDIANA						
					45,5						

$$Mediana\ (X) = Media\left(X\frac{N}{2}, X\frac{N}{2}+1\right) = \frac{X\frac{N}{2}+X\frac{N}{2}+1}{2}$$

Entonces, ¿qué significa la mediana?

Bueno, al igual que la media, proporciona una medida útil del centro de nuestro conjunto de datos. Ahora sabemos que el peso medio de los niños de nuestro grupo es de 45,5. Pero también es útil comparar la mediana con la media. 45,5 es obviamente menor que la media, que fue 53,5. A menudo, la media y la mediana serán las mismas en un conjunto de datos, pero a veces son diferentes, como en nuestro caso. Cuando la media y la mediana son iguales, se sabe que el conjunto de datos está " distribuido normalmente ". Cuando la media y la mediana son diferentes, sabe que los datos están " sesgados " de alguna manera.

¿Qué se quiere decir con sesgado?

Bueno, a diferencia de la media, que era un cálculo matemático que usaba todas las observaciones del conjunto de datos, la mediana ignora lo que dicen los números y solo usa la observación del medio. ¿Cuál es la correcta? Ambos lo son. Ninguno es necesariamente mejor que el otro. Entonces, ¿por qué usar una mediana? Bueno, hay ciertos tipos de datos en los que le preocupará el sesgo. El sesgo es cuando la media sube o baja por encima de la mediana debido a valores muy altos o muy bajos.

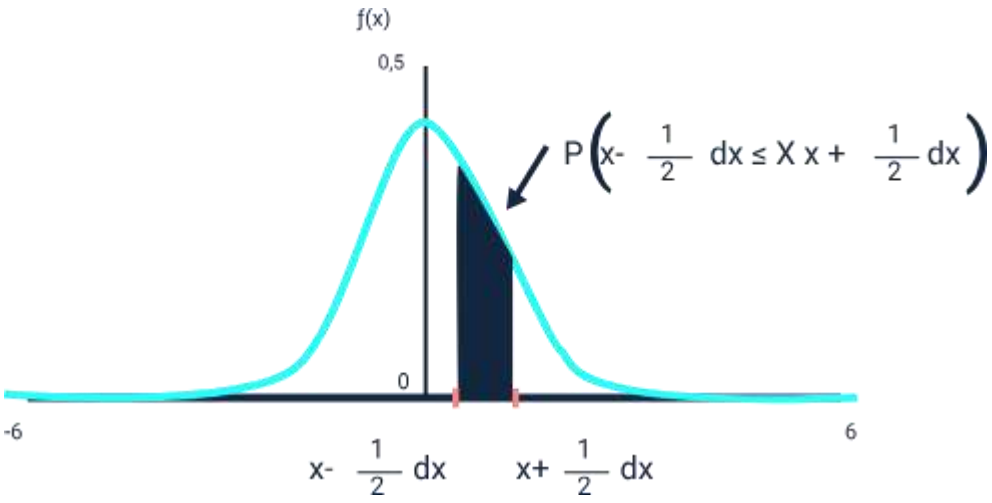
Moda

Es el valor que aparece con mayor frecuencia en un conjunto de valores de datos. Si X es una variable aleatoria discreta, la moda es el valor x (es decir, X = x) en el que la función de masa de probabilidad toma su valor máximo. En otras palabras, es el valor que es más probable que se muestree. De esta manera:

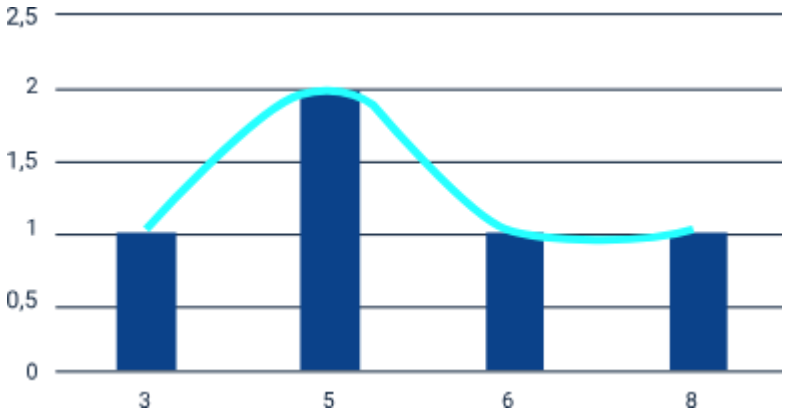
Al igual que la media y la mediana estadísticas, la moda es una forma de expresar, en un número (normalmente) único, información importante sobre una variable aleatoria o una población. El valor numérico de la moda es el mismo que el de la media y la mediana en una distribución normal y puede ser muy diferente en distribuciones muy asimétricas. La moda no

es necesariamente única para una distribución discreta dada, ya que la función de masa de probabilidad puede tomar el mismo valor máximo en varios puntos x1 x2, etc. El caso más extremo ocurre en distribuciones uniformes, donde todos los valores ocurren con la misma frecuencia.

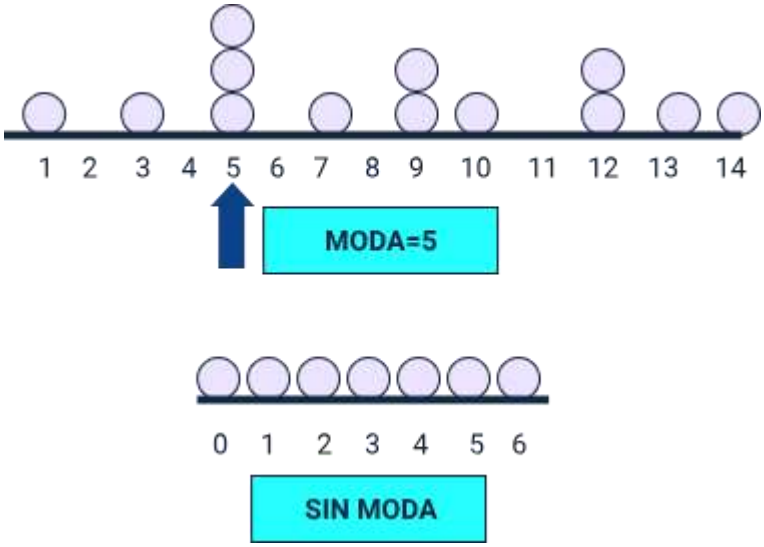
Cuando la función de densidad de probabilidad de una distribución continua tiene múltiples máximos locales, es común referirse a todos los máximos locales como modos de la distribución. Esta distribución continua se denomina multimodal (en contraposición a unimodal). Un modo de una distribución de probabilidad continua se considera a menudo como cualquier valor x en el que su función de densidad de probabilidad tiene un valor máximo localmente, por lo que cualquier pico es un modo.



En distribuciones unimodales simétricas, como la distribución normal, la media (si está definida), la mediana y la moda coinciden. Para las muestras, si se sabe que se extraen de una distribución unimodal simétrica, la media de la muestra se puede utilizar como una estimación de la moda poblacional.

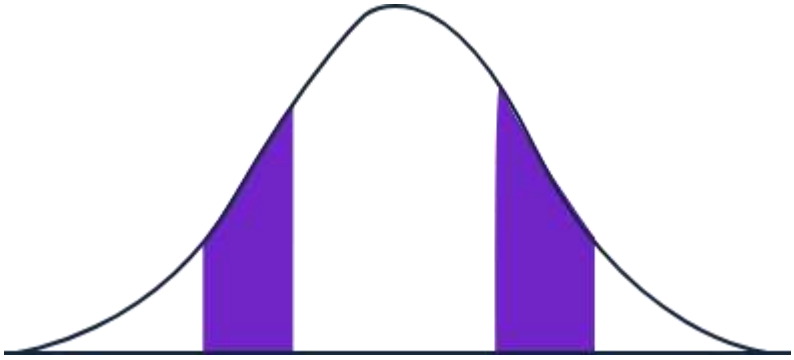


Moda de una muestra: La moda de una muestra es el elemento que aparece con mayor frecuencia en la colección. Por ejemplo, la moda de la muestra [1, 3, 6, 6, 6, 6, 7, 7, 12, 12, 17] es 6. Dada la lista de datos [1, 1, 2, 4, 4] su modo no es único. En tal caso, se dice que un conjunto de datos es bimodal, mientras que un conjunto con más de dos modos puede describirse como multimoda.

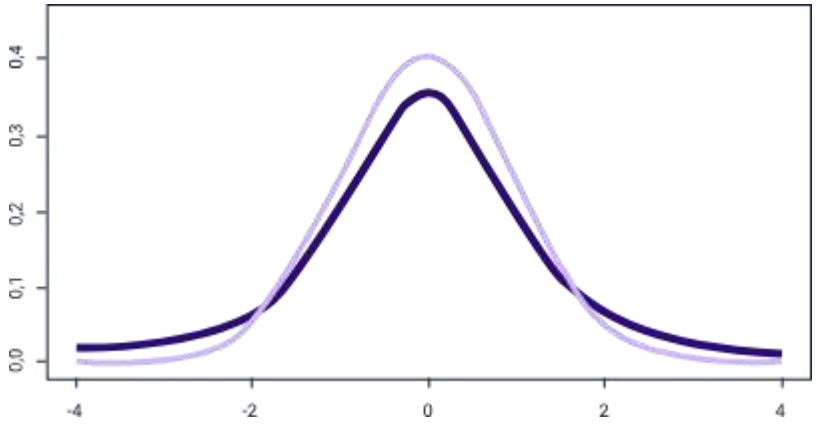


Para una muestra de una distribución continua, como [0.935 ..., 1.211 ..., 2.430 ..., 3.668 ..., 3.874 ...], el concepto es inutilizable en su forma bruta, ya que no hay dos valores será exactamente el mismo, por lo que cada valor ocurrirá exactamente una vez. Para estimar la moda de la distribución subyacente, la práctica habitual es discretizar los datos

asignando valores de frecuencia a intervalos de igual distancia, como para hacer un histograma, reemplazando efectivamente los valores por los puntos medios de los intervalos a los que están asignados. El modo es entonces el valor en el que el histograma alcanza su pico.



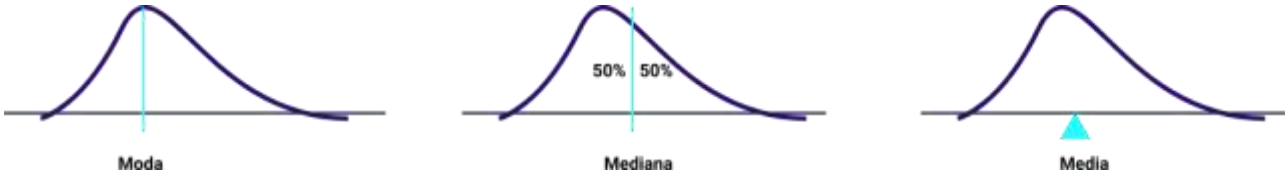
Para muestras pequeñas o medianas, el resultado de este procedimiento es sensible a la elección del ancho del intervalo si se elige demasiado estrecho o demasiado ancho; por lo general, se debe tener una fracción considerable de los datos concentrada en un número relativamente pequeño de intervalos (5 a 10), mientras que la fracción de los datos que quedan fuera de estos intervalos también es considerable. Un enfoque alternativo es la estimación de la densidad del núcleo, que esencialmente difumina las muestras puntuales para producir una estimación continua de la función de densidad de probabilidad que puede proporcionar una estimación del modo.



Comparación de media, mediana y moda

A diferencia de la media y la mediana, el concepto de moda también tiene sentido para los “datos nominales” (es decir, que no constan de valores numéricos en el caso de la media, ni siquiera de valores ordenados en el caso de la mediana). Por ejemplo, tomando una muestra de apellidos coreanos, uno podría encontrar que “Kim” aparece con más frecuencia que cualquier otro nombre. Entonces “Kim” sería la moda de la muestra. En cualquier sistema de votación donde una pluralidad determina la victoria, un único valor modal determinar al vencedor, mientras que un resultado multimodal requeriría algún procedimiento de desempate.

Figura 1. Diferencias entre mediana, media y moda



A diferencia de la mediana, el concepto de moda tiene sentido para cualquier variable aleatoria que asuma valores de un espacio vectorial, incluidos los números reales (un espacio vectorial unidimensional) y los enteros (que pueden considerarse incrustados en los reales). Por ejemplo, una distribución de puntos en el plano normalmente tendrá una media y una moda, pero el concepto de mediana no se aplica. La mediana tiene sentido cuando existe un orden lineal en los valores posibles.

Las generalizaciones del concepto de mediana a espacios de dimensiones superiores son la mediana geométrica y el punto central.

Tipo	Descripción	Ejemplo	Resultado
Media Aritmética	Suma de valores de un conjunto de datos dividida por el número de elementos o cantidad de números.	(1 + 2 + 2 + 3 + 4 + 7 + 9) / 7	4
Mediana	Valor medio o central que separa las mitades menor y mayor de un conjunto de elementos numéricos.	1, 2, 2, 3, 4, 7, 9	3
Moda	Valor más frecuente en un conjunto de elementos numéricos.	1, 2, 2, 3, 4, 7, 9	2

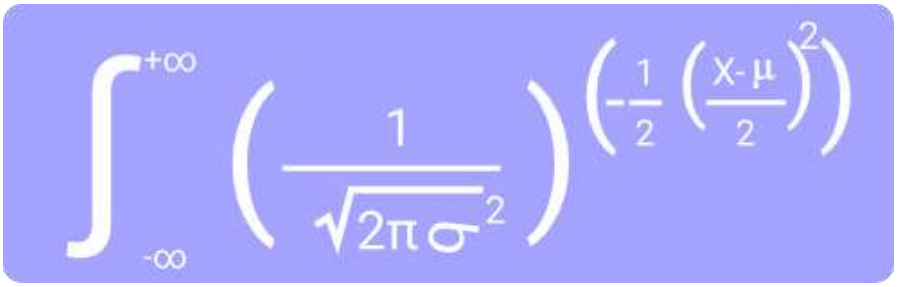
Varianza y desviación típica o estándar

La varianza es la expectativa de la desviación al cuadrado de una variable aleatoria de su media poblacional o muestral. La varianza es una medida de dispersión, lo que significa que es una medida de qué tan lejos se separa un conjunto de números de su valor promedio y se define de manera detallada así:

Varianza y desviación típica o estándar

La varianza es la expectativa de la desviación al cuadrado de una variable aleatoria de su media poblacional o muestral. La varianza es una medida de dispersión, lo que significa que es una medida de qué tan lejos se separa un conjunto de números de su valor promedio.

La **varianza** tiene un papel central en la estadística, donde algunas ideas que la usan incluyen estadística descriptiva, inferencia estadística, prueba de hipótesis y muestreo de Monte Carlo.



La varianza…

Es una herramienta importante en las ciencias, donde el análisis estadístico de datos es común. La varianza es el cuadrado de la desviación estándar, el segundo momento central de una distribución y la covarianza de la variable aleatoria consigo misma, y a menudo se representa por σ2, s2, Var(x) o V(x).

Una ventaja de la varianza como medida de dispersión es…

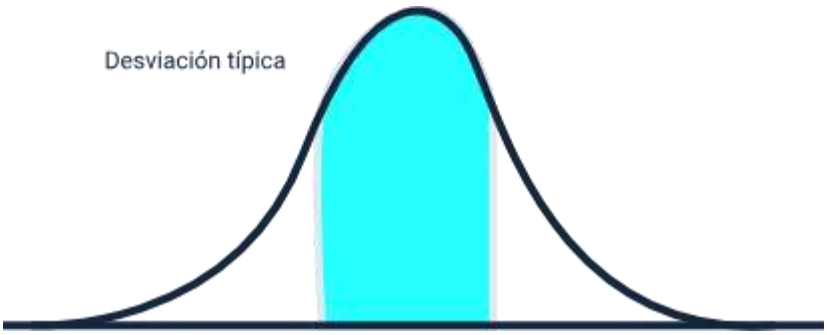
Que es más susceptible de manipulación algebraica que otras medidas de dispersión, como la desviación absoluta esperada; por ejemplo, la varianza de una suma de variables aleatorias no correlacionadas es igual a la suma de sus varianzas. Una desventaja de la varianza para aplicaciones prácticas es que, a diferencia de la desviación estándar, sus unidades difieren de la variable aleatoria, por lo que la desviación estándar se informa más comúnmente como una medida de dispersión una vez finalizado el cálculo.

$$R = L_m - L_i \qquad D\bar{x} = \frac{\sum_{i=1}^{N} |\bar{x}_i - \bar{x}| f_i}{N} \qquad CV = \frac{\sigma}{X}$$

Rango | Desviación media | Coeficiente de variación

La desviación estándar en las estadísticas

Está típicamente denotada por σ, es una medida de variación o dispersión (se refiere al grado de estiramiento o compresión de una distribución) entre valores en un conjunto de datos. Cuanto menor es la desviación estándar, más cerca tienden a estar los puntos de datos de la media (o valor esperado), μ. Por el contrario, una desviación estándar más alta indica una gama más amplia de valores. De manera similar a otros conceptos matemáticos y estadísticos, existen muchas situaciones diferentes en las que se puede usar la desviación estándar y, por lo tanto, muchas ecuaciones diferentes.



Además de expresar la variabilidad de la población

La desviación estándar también se usa a menudo para medir resultados estadísticos como el margen de error. Cuando se usa de esta manera, la desviación estándar a menudo se denomina error estándar de la media o error estándar de la estimación con respecto a una media.

Desviación estándar de población

La desviación estándar de la población, la definición estándar de σ se usa cuando se puede medir una población completa y es la raíz cuadrada de la varianza de un conjunto de datos dado.

En los casos en que se pueda muestrear a cada miembro de una población, se puede utilizar la siguiente ecuación para encontrar la desviación estándar de toda la población:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\bar{x}_i - \mu\right)^2}$$

Donde:

xi es un valor individual

μ es el valor medio / esperado

N es el número total de valores

Para aquellos que no están familiarizados con la notación de suma, la ecuación anterior puede parecer desalentadora, pero cuando se aborda a través de sus componentes individuales, esta suma no es particularmente complicada.

Para entender más…

El i = 1 en la suma indica el índice inicial, es decir, para el conjunto de datos 1, 3, 4, 7, 8, i = 1 sería 1, i = 2 sería 3, y así sucesivamente. Por lo tanto, la notación de suma simplemente significa realizar la operación de (xi – μ)2 en cada valor a través de N, que en este caso es 5 ya que hay 5 valores en este conjunto de datos.

Ejemplo:

μ = (1 + 3 + 4 + 7 + 8) / 5 = 4,6.

σ = √ [(1 - 4,6)2 + (3 - 4,6)2 + … + (8 - 4,6)2] / 5.

σ = √ (12,96 + 2,56 + 0,36 + 5,76 + 11,56) / 5 = 2,577.

Desviación estándar de la muestra

En muchos casos, no es posible muestrear a todos los miembros dentro de una población, lo que requiere que se modifique la ecuación anterior para que la desviación estándar se pueda medir a través de una muestra aleatoria de la población que se está estudiando. Un estimador común para σ es la desviación estándar de la muestra, normalmente denotada por s. Vale la pena señalar que existen muchas ecuaciones diferentes para calcular la desviación estándar de la muestra, ya que, a diferencia de la media de la muestra, la desviación estándar de la muestra no tiene un estimador único que sea insesgado, eficiente y con una probabilidad máxima.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}{n-1}}$$

La ecuación proporcionada a continuación es la “desviación estándar de la muestra corregida”. Es una versión corregida de la ecuación obtenida al modificar la ecuación de desviación estándar de la población utilizando el tamaño de la muestra como el tamaño de la población, lo que elimina parte del sesgo de la ecuación. Sin embargo, la estimación imparcial de la desviación estándar es muy complicada y varía según la distribución.

$$S = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}$$

Como tal la “desviación estándar de la muestra corregida” es el estimador más comúnmente utilizado para la desviación estándar de la población, y generalmente se la denomina simplemente “desviación estándar de la muestra”. Es una estimación mucho mejor que su versión sin corregir, pero aún tiene un sesgo significativo para tamaños de muestra pequeños (N <10).

Donde:

xi es un valor de la muestra

$\bar{x}$ es la media de la muestra

N es el tamaño de la muestra

La desviación estándar se usa ampliamente en entornos experimentales e industriales para probar modelos con datos del mundo real; un ejemplo de esto es el control de calidad de algunos productos. A continuación, se informa acerca de sus usos:

Aplicaciones de la desviación estándar

La desviación estándar se usa ampliamente en entornos experimentales e industriales para probar modelos con datos del mundo real. Un ejemplo de esto en aplicaciones industriales es el control de calidad de algunos productos. La desviación estándar se puede utilizar para calcular un valor mínimo y máximo dentro del cual algún aspecto del producto debería caer en un alto porcentaje del tiempo. En los casos en que los valores estén fuera del rango calculado, puede ser necesario realizar cambios en el proceso de producción para garantizar el control de calidad.

La desviación estándar también se usa en el clima para determinar las diferencias en el clima regional. Imagínese dos ciudades, una en la costa y otra tierra adentro, que tienen la

misma temperatura media de 75 ° F. Si bien esto puede generar la creencia de que las temperaturas de estas dos ciudades son prácticamente las mismas, la realidad podría enmascararse si solo se aborda la media y se ignora la desviación estándar.

Las ciudades costeras tienden a tener temperaturas mucho más estables debido a la regulación de grandes masas de agua, ya que el agua tiene una mayor capacidad calorífica que la tierra; esencialmente, esto hace que el agua sea mucho menos susceptible a los cambios de temperatura, y las áreas costeras permanecen más cálidas en invierno y más frescas en verano debido a la cantidad de energía requerida para cambiar la temperatura del agua.

Otra área en la que se usa ampliamente la desviación estándar es la de las finanzas, donde a menudo se usa para medir el riesgo asociado en las fluctuaciones de precios de algún activo o cartera de activos.

El uso de la desviación estándar en estos casos proporciona una estimación de la incertidumbre de los rendimientos futuros de una inversión determinada.

Por ejemplo, al comparar la acción A que tiene un rendimiento promedio del 7% con una desviación estándar del 10% contra la acción B, que tiene el mismo rendimiento promedio pero una desviación estándar del 50%, la primera acción sería claramente la opción más segura. dado que la desviación estándar de la acción B es significativamente mayor, para el mismo rendimiento exacto.

Eso no quiere decir que la acción A sea definitivamente una mejor opción de inversión en este escenario, ya que la desviación estándar puede sesgar la media en cualquier dirección.

Estos son solo algunos ejemplos de cómo se puede usar la desviación estándar, pero existen muchos más. Generalmente, calcular la desviación estándar es valioso en cualquier

momento que se desee saber qué tan lejos de la media puede estar un valor típico de una distribución.

Estudio de variables continuas

Una variable continua es un tipo específico de variable cuantitativa que se utiliza en estadística para describir datos que se pueden medir de alguna manera. Si sus datos tratan de medir la altura, el peso o el tiempo, entonces tiene una variable continua.

Definamos más un par de términos utilizados en nuestra definición. Una variable en estadística no es lo mismo que una variable en álgebra. En estadística, una variable es algo que nos da datos. Algunos ejemplos de variables en las estadísticas pueden incluir edad, color de ojos, altura, número de hermanos, sexo o número de mascotas. Nuestra definición de variable continua también menciona que es cuantitativa. Los datos cuantitativos involucran cantidades o números. En los ejemplos de variables enumerados anteriormente, su edad, altura, número de hermanos y número de mascotas son todas variables cuantitativas.

La definición también menciona que los datos se pueden medir de alguna manera. Para comprender esto, debe comprender las variables discretas. Los datos se consideran discretos si son un recuento. Cuando contamos cosas, usamos números enteros como 0, 1, 2 y 3. Mi ejemplo favorito de una variable discreta es cuántos huevos pone una gallina. Cada día una gallina puede o no poner un huevo, pero hay dos cosas que nunca pueden suceder. Nunca puede haber un número negativo de huevos, y nunca puede haber una fracción o una porción de un huevo.

Las variables continuas son variables que miden algo. Un buen ejemplo de variable continua es cuántos galones de leche da una vaca. Hasta donde se sabe, las vacas no saben cómo dejar de producir o dar leche después de exactamente 4 galones. Bessie puede dar 4,17

galones el lunes, 3,89 galones el martes y 4,2 galones el miércoles. Observe que las variables continuas nos permiten tener decimales o fracciones.

1.2 Distribuciones bidimensionales, diagramas de dispersión y rectas de regresión

En el siguiente recurso, se invita a descubrir todo sobre las distribuciones bidimensionales, diagramas de dispersión y rectas de regresión.

a. Distribuciones Bidimensionales

En la estadística, a la distribución donde interceden dos variables, x e y, se le denomina distribución bidimensional, esto significa que a cada individuo le pertenecen dos valores, xi, yi. Y al momento de ser representados en un plano o diagrama cartesiano se les debe considerar como coordenadas de un solo punto (xi, yi). Así, toda esta distribución se visualizará como un conjunto de puntos y cada punto corresponde a un individuo de la distribución.

Como ejemplo típico de distribución bidimensional, se pueden tomar los errores de precisión del fuego de artillería. El error total se compone de dos desviaciones independientes: error en la distancia de disparo X y desviación lateral Y de la dirección de disparo. Un número tan relativamente pequeño de puntos clasifica esta distribución entre las distribuciones discretas.

b. Correlación

La correlación es una medida estadística que indica la relación lineal entre dos variables. No implica una relación de causa y efecto, sino que cuantifica la fuerza de la relación. Se utiliza un coeficiente de correlación, r, para medir la correlación. Además, se

pueden realizar pruebas estadísticas para determinar la significancia de la correlación. Sin embargo, la correlación tiene limitaciones, ya que no puede tener en cuenta otras variables que puedan influir en la relación entre las dos variables analizadas.

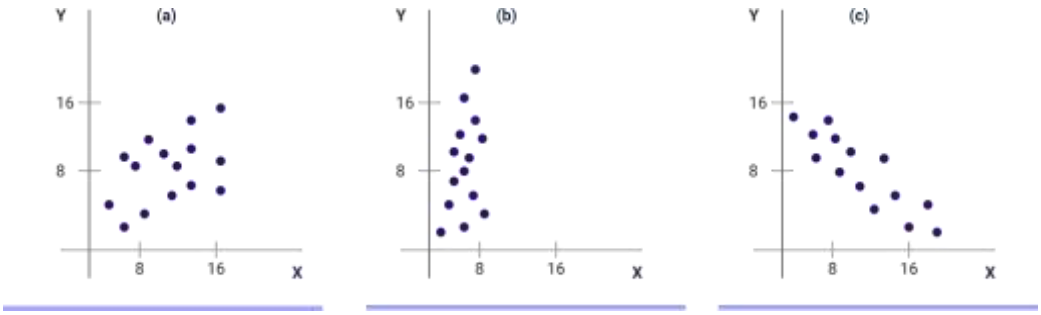
La correlación no implica causalidad y no describe relaciones curvilíneas. Sin embargo, es útil para analizar relaciones simples entre datos. Por ejemplo, al examinar datos de campings en un parque de montaña, se puede utilizar la correlación para determinar si hay una relación entre la elevación del camping y la temperatura alta en verano. Una correlación negativa indica que a medida que aumenta la elevación, la temperatura tiende a descender.

Diagramas de dispersión

Los diagramas de dispersión son herramientas matemáticas convenientes para estudiar la correlación entre dos variables aleatorias. Como su nombre indica, son una forma de hoja de papel sobre la que se encuentran dispersos los puntos de datos correspondientes a las variables de interés.

A juzgar por la forma del patrón que forman los puntos de datos en esta hoja de papel, podemos determinar la asociación entre las dos variables y aplicar la técnica de análisis de correlación más adecuada.

Figura 2. a. Correlación débil, b. Correlación fuerte, c. correlación inversa



Interpretación de diagramas de dispersión

Los diagramas de dispersión entre dos variables aleatorias presentan las variables como sus ejes x e y. Podemos tomar cualquier variable como variable independiente en tal caso (la otra variable es la dependiente) y, en consecuencia, graficar cada punto de datos en el gráfico (xi, yi). La totalidad de todos los puntos graficados forma el diagrama de dispersión. Basándonos en las diferentes formas que puede asumir el diagrama de dispersión, podemos sacar diferentes inferencias. Podemos calcular un coeficiente de correlación para los datos dados. Es una medida cuantitativa de la asociación de las variables aleatorias. Su valor es siempre menor que 1 y puede ser positivo o negativo.

En el caso de una correlación positiva, los puntos graficados se distribuyen desde la esquina inferior izquierda a la esquina superior derecha (en el patrón general de estar distribuidos uniformemente sobre una línea recta con una pendiente positiva), y en el caso de una correlación negativa, los puntos trazados se extienden sobre una línea recta de una pendiente negativa) desde la parte superior izquierda hacia la parte inferior derecha.

Si los puntos están distribuidos aleatoriamente en el espacio, o distribuidos casi por igual en todos los lugares sin representar ningún patrón en particular, es el caso de una correlación muy pequeña, que tiende a 0.

Regresión

Consiste en una técnica estadística para estimar las relaciones entre variables. Incluye muchas técnicas para modelar y analizar varias variables cuando el foco está en la relación entre una variable dependiente y una o más variables independientes. Más específicamente, el análisis de regresión ayuda a comprender cómo cambia el valor típico de la variable dependiente cuando se varía cualquiera de las variables independientes, mientras que las otras variables independientes se mantienen fijas. Más comúnmente, el análisis de regresión estima la expectativa condicional de la variable dependiente dadas las variables independientes, es decir, el valor promedio de la variable dependiente cuando las variables independientes son fijas. Con menos frecuencia, la atención se centra en un cuantil, u otro parámetro de ubicación de la distribución condicional de la variable dependiente dadas las variables independientes.

En todos los casos, el objetivo de estimación es una función de las variables independientes, llamada función de regresión. En el análisis de regresión, también es interesante caracterizar la variación de la variable dependiente alrededor de la función de regresión, que puede describirse mediante una distribución de probabilidad. Una aplicación muy común de la regresión es ayudar a los administradores financieros y de inversiones a valorar los activos y comprender las relaciones entre las variables, como los precios de los productos básicos y las acciones de las empresas que comercian con esos productos básicos.

El software no es solo para buscar filtros de caras divertidas o aprender nuevos movimientos de baile. Puede ayudar a su equipo a aumentar su eficiencia y ser más productivo y capaz en su trabajo. Una solución de software personalizada puede ayudar a eliminar el cuello de botella del seguimiento de los recibos y gastos del personal, o puede facilitar que su equipo administre los contactos de marketing.

Y el software personalizado también puede servir como una solución para sus clientes. Por ejemplo, una opción de chat en vivo en su sitio web proporciona un contacto inmediato para clientes o prospectos con una necesidad urgente o que están listos para realizar una compra. Un software personalizado también podría ayudar a sus clientes a rastrear a su representante de servicio en el camino a su hogar, o enviar una solicitud de soporte técnico.

Rectas de regresión

Están definidas de la siguiente manera:

a. Las rectas de regresión son líneas que se utiliza para describir el comportamiento de un conjunto de datos. En otras palabras, da la mejor tendencia de los datos proporcionados. Las rectas de regresión son útiles en los procedimientos de pronóstico. Su propósito es describir la interrelación de la variable dependiente (variable y) con una o muchas variables independientes (variable x).
b. El uso de la ecuación obtenida de la recta de regresión actúa como un analista que puede pronosticar comportamientos futuros de las variables dependientes ingresando diferentes valores para las independientes. Los dos tipos básicos de regresión son la recta de regresión normal o simple y la recta de regresión múltiple, aunque existen métodos de regresión no lineal para datos y análisis más complicados.
c. La recta de regresión simple usa una variable independiente para explicar o predecir el resultado de la variable dependiente Y, mientras que la recta de regresión múltiple usa dos o más variables independientes para predecir el resultado.
d. La forma general de cada tipo de regresión es: Fórmula de la recta de regresión simple: $Y = a + bX + u$

   Fórmula de la recta de regresión múltiple: $Y = a + b1X1 + b2\ X2 + b3\ X3 + \ldots + bt\ Xt + u$

Donde:

Y = la variable que está intentando predecir (variable dependiente).

X = la variable que está utilizando para predecir Y (variable independiente).

a = la intersección.

b = la pendiente.

u = el residual de regresión.

Para profundizar aún más en este ítem, diríjase al:

Anexo 1 - Distribuciones bidimensionales y rectas de regresión.

1.3 Distribuciones discretas, distribución binomial, distribuciones continuas, distribución normal

A continuación, se exponen los diferentes tipos de distribución y cómo aplicarlos en la vida profesional.

En estadística, las distribuciones parecen infinitas con docenas de distribuciones compitiendo por su atención y con poca o ninguna base intuitiva para diferenciarlas. Las descripciones tienden a ser abstractas y enfatizan propiedades estadísticas como los momentos, funciones características y distribuciones acumulativas.

Distribuciones discretas

Una distribución discreta es una distribución de probabilidad que representa la ocurrencia de resultados discretos (contables individualmente), como 1, 2, 3 … o 0 contra 1.

Por ejemplo, es una distribución discreta que evalúa la probabilidad de que ocurra un resultado “sí” o “no” en un número determinado de intentos, dada la probabilidad del evento en cada intento, como lanzar una moneda cien veces y teniendo el resultado “cabezas”.

Las distribuciones estadísticas pueden ser discretas o continuas. Una distribución continua se construye a partir de resultados que caen en un continuo, como todos los números mayores que 0 (que incluirían números cuyos decimales continúan indefinidamente, como pi = 3,14159265 …). En general, los conceptos de distribuciones de probabilidad discretas y continuas y las variables aleatorias que describen son la base de la teoría de la probabilidad y el análisis estadístico.

Características principales

Una distribución de probabilidad discreta cuenta las ocurrencias que tienen resultados contables o finitos.

Esto contrasta con una distribución continua, donde los resultados pueden caer en cualquier parte de un continuo.

Los ejemplos comunes de distribución discreta incluyen las distribuciones binomiales, binomial negativa, beta-binomial, Poisson, hipergeométrica y Bernoulli.

Estas distribuciones a menudo involucran análisis estadísticos de “conteos” o “cuántas veces” ocurre un evento.

En finanzas, las distribuciones discretas se utilizan en la fijación de precios de opciones y en la previsión de shocks o recesiones del mercado.

Distribución binomial

Suponer que X1, X2, …, Xn son variables aleatorias de Bernoulli independientes e idénticamente distribuidas (iid), cada una con la distribución.

$$F(x_i|\pi) = \pi^{x_i}(1-\pi)^{1-x_i} \quad \text{for} \quad X_i = 0,1 \text{ and } 0 \leq \pi \leq 1$$

Entonces se dice que X tiene una distribución binomial con parámetros n y p:

$$X \sim \text{Bin}(n, \pi)$$

Suponga que un experimento e x consiste en n ensayos repetidos tipo Bernoulli, cada ensayo resulta en un “éxito” con probabilidad π y un “fracaso” con probabilidad 1 – π. Por ejemplo, lanza una moneda 100 veces n = 100. Cuente el número de veces que observa cabezas, por ejemplo, X = # de cabezas. Si todos los ensayos son independientes, es decir, si la probabilidad de éxito en cualquier ensayo no se ve afectada por el resultado de cualquier otro ensayo, entonces el número total de éxitos en el experimento tendrá una distribución binomial, por ejemplo, dos lanzamientos de moneda no lo hacen afectarse unos a otros. La distribución binomial se puede escribir como:

$$F(x) = \frac{n!}{x!\,(n-x)!}\pi^{x}(1-\pi)^{n-x} \quad \text{for } x_i = 0,1,2 \ldots n, \text{ and } 0 \leq \pi \leq 1$$

La distribución de Bernoulli es un caso especial del binomio con n = 1. Es decir,

$$X \sim \text{Bin}(1, \pi)$$

Ello significa que X tiene una distribución de Bernoulli con probabilidad de éxito π. Se puede demostrar algebraicamente que si se cumple la ecuación anterior entonces:

$$E(X) = n\pi \text{ Y } V(x) = n\pi(1 - \pi)$$

Una forma más sencilla de llegar a estos resultados es tener en cuenta que dónde son variables aleatorias de Bernoulli. Entonces, por las propiedades aditivas de media y varianza, X= X1, X2, …, Xn donde X1, X2, …, Xn son variables aleatorias de Bernoulli. Entonces, por las propiedades aditivas de media y varianza,

$$E(X) = E(X_1) + E(X_2) + \ldots + E(X_n) = n\pi$$

$$V(X) = V(X_1) + V(X_2) + \ldots + V(X_n) = n\pi(1 - \pi)$$

Tenga en cuenta que X no tendrá una distribución binomial si la probabilidad de éxito π no es constante de un ensayo a otro, o si los ensayos no son completamente independientes (es decir, un éxito o fracaso en un ensayo altera la probabilidad de éxito en otro ensayo).

$$\text{If} \quad \text{Bin}(n_1, \pi) \text{ and } X_2 \sim \text{Bin}(n_2, \pi), \text{ then } X_1 + X_2 \sim \text{Bin}(n_1 + ., \pi)$$

A medida que n aumenta, para fijo π, la distribución binomial se aproxima a la distribución normal:

$$N(n\pi, n\pi(1 - \pi))$$

Por ejemplo, si se toman muestras sin reemplazo de una población finita, entonces la distribución hipergeométrica es apropiada.

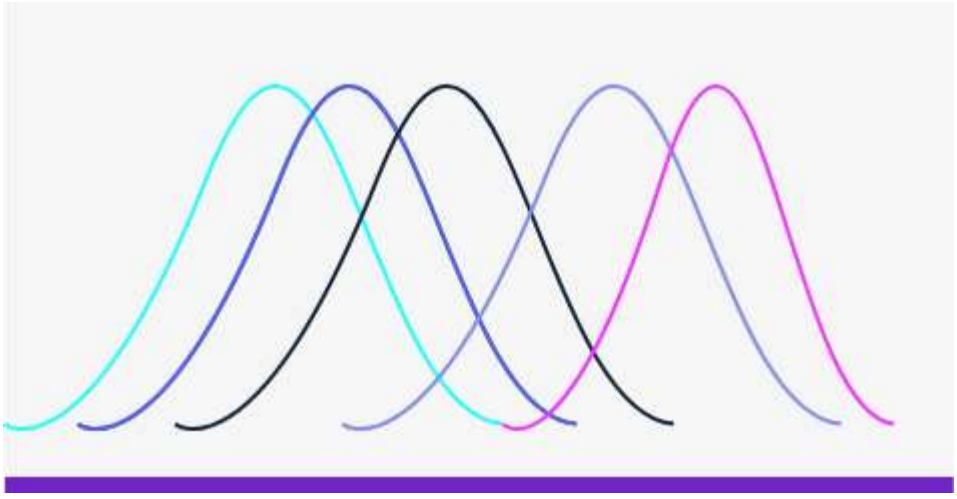
Para profundizar aún más en este ítem, diríjase al:

Anexo 2 - Modelos de probabilidad discretos.

Distribuciones continuas

Estas pueden tomar cualquier valor dentro de un rango definido. Este rango puede ser infinito (por ejemplo, para la distribución normal) en cuyo caso se habla de una distribución limitada o si es finita (por ejemplo, la distribución uniforme) es una distribución delimitada.

Figura 3. Distribuciones continuas



En el Anexo 3 - Distribuciones continuas

Encontrarán una tabla que ofrece una descripción general de varias distribuciones continuas que se usan comúnmente en el modelado de análisis de riesgos, de modo que pueda enfocarse más fácilmente en cuáles podrían ser las más apropiadas para sus

necesidades de modelado. Siga los enlaces para obtener una explicación detallada de cada uno. Se ha utilizado el nombre más común para cada distribución.

Una distribución continua se utiliza para representar una variable que puede tomar cualquier valor dentro de un rango definido (dominio). Por ejemplo, la estatura de un inglés adulto escogido al azar tendrá una distribución continua, porque la estatura de una persona es esencialmente infinitamente divisible. Se podría medir su estatura en centímetro, milímetro, décimo de milímetro, etc.; la escala se puede dividir repetidamente generando cada vez más valores posibles.

Propiedades como el tiempo, la masa y la distancia, que son infinitamente divisibles, se modelan utilizando distribuciones continuas. En la práctica, también se usan distribuciones continuas para modelar variables que son, en verdad, discretas, pero donde la brecha entre los valores permitidos es insignificante: por ejemplo, el costo del proyecto (que es discreto con pasos de un centavo, un centavo, etc.), tipo de cambio (que solo se cotiza a unas pocas cifras significativas), número de empleados en una gran organización, etc.

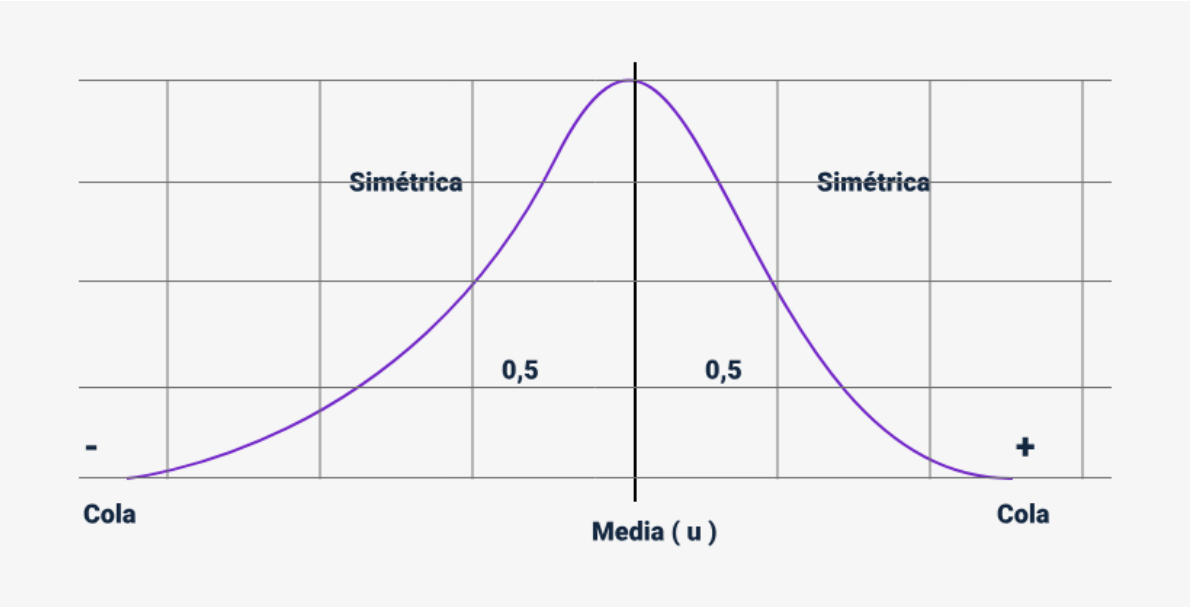
Distribución normal

En la teoría de la probabilidad, la distribución normal (o gaussiana o Gauss o Laplace-Gauss) es un tipo de distribución de probabilidad continua para un valor real- variable aleatoria. La forma general de su función de densidad de probabilidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

El parámetro μ es la media o expectativa de la distribución (y también su mediana y moda), mientras que el parámetro es su desviación estándar σ. La varianza de la distribución es σ2. Se dice que una variable aleatoria con una distribución gaussiana tiene una distribución normal y se denomina desviación normal.

Figura 4. Distribución normal



Cualquier combinación lineal de una colección fija de desviaciones normales es una desviación normal. Muchos resultados y métodos, como la propagación de la incertidumbre y el ajuste de parámetros por mínimos cuadrados, pueden derivarse analíticamente de forma explícita cuando las variables relevantes se distribuyen normalmente. Una distribución normal a veces se denomina informalmente curva de campana. Sin embargo, muchas otras distribuciones tienen forma de campana (como las distribuciones de Cauchy, de Student y logística).

1.4 Muestreo, distribución de medias muéstrales

Para este tema se invita a revisar de qué trata el muestreo y la distribución de medias muestrales.

El muestreo es un proceso para seleccionar una muestra representativa de una población. Se utiliza en situaciones donde no es factible obtener información de todos los miembros de la población, como en análisis biológicos, control de calidad o encuestas sociales. El método básico es el muestreo aleatorio simple, donde cada elemento tiene la misma probabilidad de ser seleccionado. El muestreo y la inferencia estadística son herramientas importantes en la investigación científica y el análisis de datos.

En una muestra aleatoria de una clase de 50 estudiantes, por ejemplo, cada estudiante tiene la misma probabilidad, 1/50, de ser seleccionado. Cada combinación de elementos extraídos de la población también tiene la misma probabilidad de ser seleccionados. El muestreo basado en la teoría de la probabilidad permite al investigador determinar la probabilidad que los hallazgos estadísticos son el resultado del azar.

Los métodos más utilizados, refinamientos de esta idea básica, son el muestreo estratificado (en el que la población se divide en clases y se extraen muestras aleatorias simples de cada clase), el muestreo por conglomerados (en el que la unidad de la muestra es un grupo, como un hogar) y muestreo sistemático (muestras tomadas por cualquier sistema que no sea una elección aleatoria, como cada décimo nombre en una lista).

Los métodos más utilizados, refinamientos de esta idea básica, son el muestreo estratificado (en el que la población se divide en clases y se extraen muestras aleatorias simples de cada clase), el muestreo por conglomerados (en el que la unidad de la muestra es

un grupo, como un hogar) y muestreo sistemático (muestras tomadas por cualquier sistema que no sea una elección aleatoria, como cada décimo nombre en una lista).

Una alternativa al muestreo probabilístico es muestreo de juicio, en el que la selección se basa en el juicio del investigador y existe una probabilidad desconocida de inclusión en la muestra para cualquier caso dado. Por lo general, se prefieren los métodos de probabilidad porque evitan el sesgo de selección y permiten estimar el error de muestreo (la diferencia entre la medida obtenida de la muestra y la de toda la población de la que se extrajo la muestra).

Algunas características de muestreo son:

Los Contadores Públicos Certificados utilizan el muestreo durante las auditorías para determinar la precisión y la integridad de los saldos de las cuentas.

Los tipos de muestreo incluyen muestreo aleatorio, muestreo por bloques, muestreo por juicio y muestreo sistemático.

Las empresas utilizan el muestreo como una herramienta de marketing para identificar las necesidades y deseos de su mercado objetivo.

En auditorías financieras, los contadores públicos certificados utilizan el muestreo de auditoría para evaluar la precisión de los saldos de cuentas en los estados financieros. Este método se utiliza cuando la población de transacciones es grande. Además, los gerentes empresariales emplean el muestreo de clientes para evaluar la demanda de nuevos productos y el éxito del marketing. El muestreo es una herramienta útil en la toma de decisiones y evaluación de datos.

La muestra elegida debe ser una representación justa de toda la población. Al tomar una muestra de una población más grande, es importante considerar cómo se elige la muestra. Para obtener una muestra representativa, se debe extraer al azar y abarcar a toda la población. Por ejemplo, se podría usar un sistema de lotería para determinar la edad promedio de los estudiantes en una universidad tomando una muestra del 10% del cuerpo estudiantil.

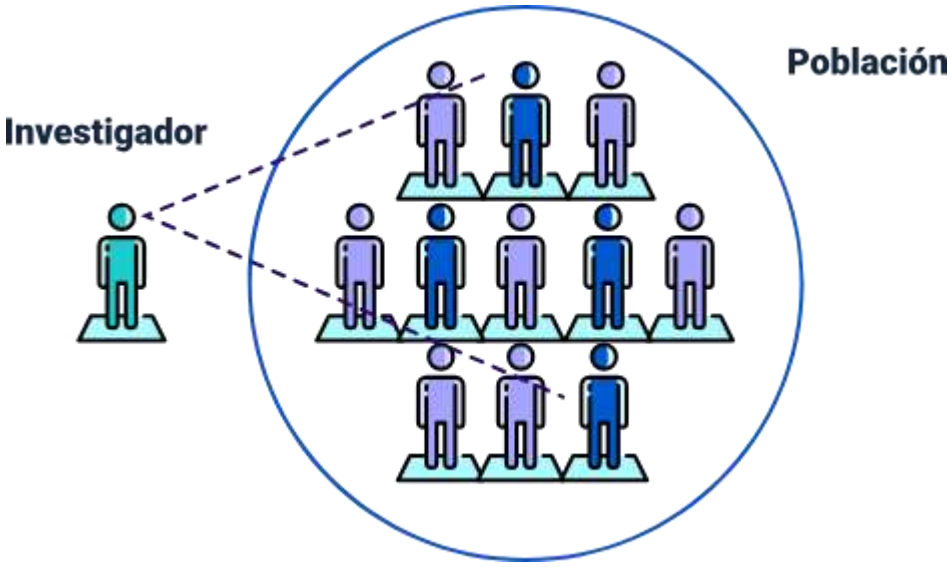
Y ahora se sugiere complementar la información con la siguiente presentación.

a. Muestreo aleatorio

Con el muestreo aleatorio, cada elemento de una población tiene la misma probabilidad de ser elegido. Es el más alejado de cualquier sesgo potencial porque no hay un juicio humano involucrado en la selección de la muestra. Por ejemplo, una muestra aleatoria puede incluir elegir los nombres de 25 empleados de un sombrero en una empresa de 250 empleados. La población es de 250 empleados y la muestra es aleatoria porque cada empleado tiene la misma probabilidad de ser elegido.

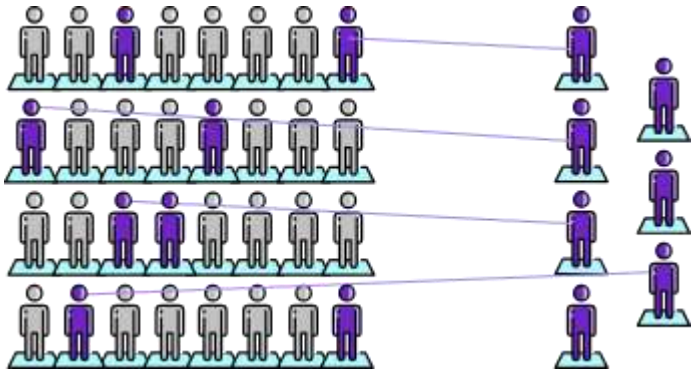
b. Muestreo discrecional o de juicio

Se puede utilizar el juicio del auditor para seleccionar la muestra de la población completa. Un auditor solo puede estar preocupado por transacciones de naturaleza material. Por ejemplo, suponga que el auditor establece el umbral de importancia relativa para las transacciones de cuentas por pagar en $ 10,000. Si el cliente proporciona una lista completa de 15 transacciones de más de $ 10,000, el auditor puede optar por revisar todas las transacciones debido al pequeño tamaño de la población.



c. Muestreo en bloque

El muestreo por bloques toma una serie consecutiva de elementos dentro de la población para usar como muestra. Por ejemplo, una lista de todas las transacciones de ventas en un período contable podría clasificarse de varias formas, incluso por fecha o por monto en dólares.



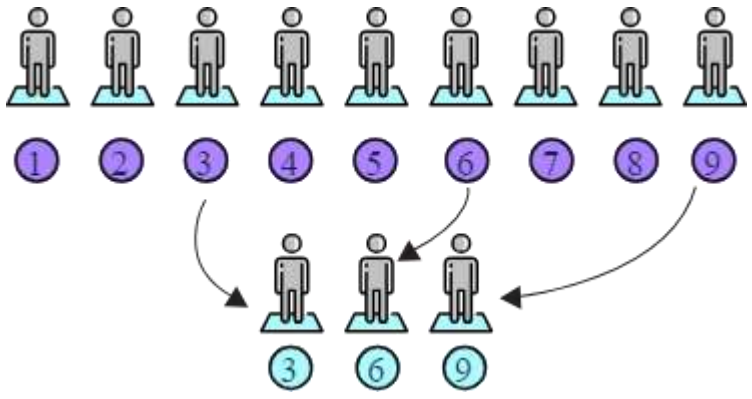
Un auditor puede solicitar que el contador de la empresa proporcione la lista en un formato u otro para seleccionar una muestra de un segmento específico de la lista. Este método requiere muy poca modificación por parte del auditor, pero es probable que un bloque de transacciones no sea representativo de la población completa.

Alternativamente, un auditor puede identificar todas las cuentas del libro mayor con una variación mayor al 10% del período anterior. En este caso, el auditor está limitando la población de la que se deriva la selección de la muestra. Desafortunadamente, el juicio humano utilizado en el muestreo siempre conlleva la posibilidad de sesgo, ya sea explícito o implícito.

d. Muestreo sistemático

Los investigadores utilizan el método de muestreo sistemático para elegir los miembros de la muestra de una población a intervalos regulares. Requiere la selección de un punto de partida para la muestra y el tamaño de la muestra que pueda repetirse a intervalos regulares.

Este tipo de método de muestreo tiene un rango predefinido y, por lo tanto, esta técnica de muestreo es la que requiere menos tiempo.



Suponga que un auditor está revisando los controles internos relacionados con la cuenta de efectivo de una empresa y desea probar la política de la empresa que estipula que los cheques que superen los $ 10,000 deben estar firmados por dos personas.

La población consiste en cada cheque de la compañía que exceda los $ 10,000 durante el año fiscal, que, en este ejemplo, fue de 300. El auditor usa estadísticas de probabilidad y determina que el tamaño de la muestra debe ser el 20% de la población o 60 cheques. El intervalo de muestreo es 5 (300 controles / 60 controles de muestra).

Por lo tanto, el auditor selecciona una de cada cinco verificaciones para probarlas. Suponiendo que no se encuentran errores en el trabajo de prueba de muestreo, el análisis estadístico le da al auditor una tasa de confianza del 95% de que el procedimiento de verificación se realizó correctamente. El auditor prueba la muestra de 60 cheques y no encuentra errores, por lo que concluye que el control interno sobre el efectivo está funcionando correctamente.

e. Ejemplo de muestreo de marketing

Las empresas tienen como objetivo vender sus productos y / o servicios a los mercados objetivo. Antes de presentar productos al mercado, las empresas generalmente identifican las necesidades y deseos de su público objetivo. Para hacerlo, pueden emplear un muestreo de la población del mercado objetivo para obtener una mejor comprensión de esas necesidades para luego crear un producto y / o servicio que satisfaga esas necesidades. En este caso, recoger las opiniones de la muestra ayuda a identificar las necesidades del conjunto.

Distribución de medias muestrales

La media muestral de un grupo de observaciones es una estimación de la media poblacional μ. Dada una muestra de tamaño n, considere n variables aleatorias independientes X1, X2, …, Xn, cada una de las cuales corresponde a una observación seleccionada al azar. Cada una de estas variables tiene la distribución de la población, con media μ y desviación estándar σ. La media muestral se define como:

$$\bar{x} = \frac{1}{n}(X_1 + x_2 + \ldots + X_n)$$

Por las propiedades de las medias y las varianzas de las variables aleatorias, la media y la varianza de la media muestral son las siguientes:

$$U_{\bar{x}} = u$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

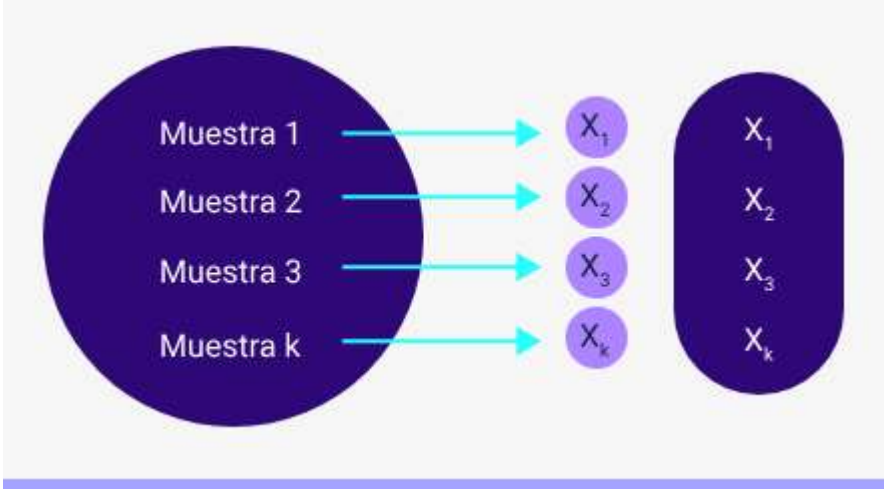
Aunque la media de la distribución de □ es idéntica a la media de la distribución de la población, la varianza es mucho menor para tamaños de muestra grandes.

Por ejemplo, suponga que la variable aleatoria X registra el puntaje de un estudiante seleccionado al azar en una prueba nacional, donde la distribución de la población para el puntaje es normal con una media de 70 y una desviación estándar de 5 (N (70,5)). Dada una muestra aleatoria simple (SRS) de 200 estudiantes, la distribución de la puntuación media de

la muestra tiene una media de 70 y una desviación estándar de 5 / sqrt (200) = 5 / 14,14 = 0,35.

La distribución de medias muestrales se define como el conjunto de medias de todas las posibles muestras aleatorias de un tamaño específico (n) seleccionadas de una población específica. Esta distribución tiene características bien definidas (y predecibles) que se especifican en el teorema del Límite central.

Figura 5. Distribución de medias muestrales



Cuando la distribución de la población es normal, la distribución de la media muestral también es normal. Para una distribución de población normal con media μ y desviación estándar σ, la distribución de la media de la muestra es normal, con media μ y desviación estándar.

$$U_{\bar{x}} = u$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Este resultado se deriva del hecho de que cualquier combinación lineal de variables aleatorias normales independientes también se distribuye normalmente. Esto significa que para dos variables aleatorias normales independientes X, Y y cualquier constante a y b, Ax + By serán distribuidos normalmente. En el caso de la media muestral, la combinación lineal es = (1 / n) * (X1 + X2 + ... Xn).

Teorema del Límite central

El resultado más importante sobre las medias muestrales es el teorema del Límite central. En pocas palabras, este teorema dice que para un tamaño de muestra n suficientemente grande, la distribución de la media muestral X se acercará a una distribución normal. Esto es cierto para una muestra de variables aleatorias independientes de cualquier distribución de población, siempre que la población tenga una desviación estándar finita σ. Una declaración formal del teorema del límite central es la siguiente:

Si X es la media de una muestra aleatoria X1, X2, ..., Xn de tamaño n de una distribución con una media finita μ y una varianza positiva finita σ², entonces la distribución de W =

$$\frac{\overline{X} - u}{o/\sqrt{n}}$$

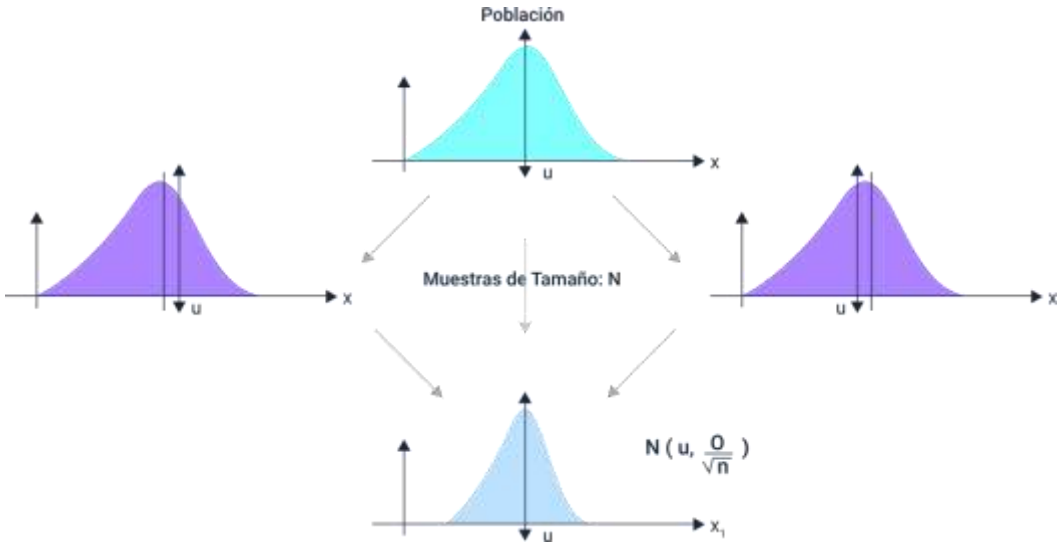
es N (0,1) en el límite cuando n se acerca al infinito.

Esto significa que la variable □ se distribuye

$$N\left(\mu, \frac{o}{\sqrt{n}}\right)$$

Una aplicación bien conocida de este teorema es la aproximación normal a la distribución binomial.

Figura 6. Teorema del Límite central



El concepto de distribución de medias muestrales y sus características deben ser intuitivamente razonables:

Es de notar que las medias muestrales son variables. Si se seleccionan dos (o más) muestras de la misma población, es probable que las dos muestras tienen medias diferentes.

Aunque las muestras tendrán medias diferentes, se debe esperar que las medias muestrales estén cerca de la media de la población. Es decir, las medias de la muestra deberían “acumularse” alrededor de μ. Por tanto, la distribución de las medias muestrales tiende a adoptar una forma normal con un valor esperado de μ.

Debe darse cuenta de que la media de una muestra individual probablemente no será idéntica a la media de su población; es decir, habrá algún “error” entre X y μ. Algunas medias de muestra estarán relativamente cerca de μ y otras relativamente lejos. El error estándar proporciona una medida de la distancia estándar entre X y μ.

Puntuaciones Z y ubicación dentro de la distribución de medias muestrales

Dentro de la distribución de las medias muestrales, la ubicación de cada media muestral se puede especificar mediante una puntuación Z,

$$z = \frac{\bar{X} - u}{\left(\frac{O}{\sqrt{n}}\right)}$$

Como siempre, una puntuación Z positiva indica una media muestral mayor que μ y una puntuación Z negativa corresponde a una media muestral menor que μ.

El valor numérico de la puntuación Z indica la distancia entre □ y μ medida en términos del error estándar.

1.5 Estimación y prueba de hipótesis

Un área de preocupación en la estadística inferencial es la estimación del parámetro de población a partir de la estadística de la muestra; aquí es importante darse cuenta del orden. La estadística de muestra se calcula a partir de los datos de muestra y el parámetro de población se infiere (o estima) a partir de esta estadística de muestra:

Sssssssssssssssssssssssssss

Otra área de la estadística inferencial es la determinación del tamaño de la muestra. Es decir, qué tamaño de muestra debe tomarse para hacer una estimación precisa. En estos casos, las estadísticas no se pueden utilizar, ya que aún no se ha tomado la muestra.

a. Estimaciones puntuales

Hay dos tipos de estimaciones que encontraremos: estimaciones puntuales y estimaciones de intervalo. La estimación puntual es el mejor valor individual. Un buen estimador debe satisfacer tres condiciones:

Insesgado: el valor esperado del estimador debe ser igual a la media del parámetro.

Consistente: el valor del estimador se acerca al valor del parámetro a medida que aumenta el tamaño de la muestra.

Relativamente eficiente: el estimador tiene la varianza más pequeña de todos los estimadores que podrían usarse.

b. Intervalos de confianza

La estimación puntual va a ser diferente del parámetro de población porque debido al error de muestreo, y no hay forma de saber quién se acerca al parámetro real. Por esta razón, a los estadísticos les gusta dar una estimación de intervalo que es un rango de valores utilizados para estimar el parámetro. Un intervalo de confianza es una estimación de intervalo con un nivel de confianza específico. Un nivel de confianza es la probabilidad de que la estimación del intervalo contenga el parámetro. El nivel de confianza es 1-alfa y esta se encuentra dentro del intervalo de confianza.

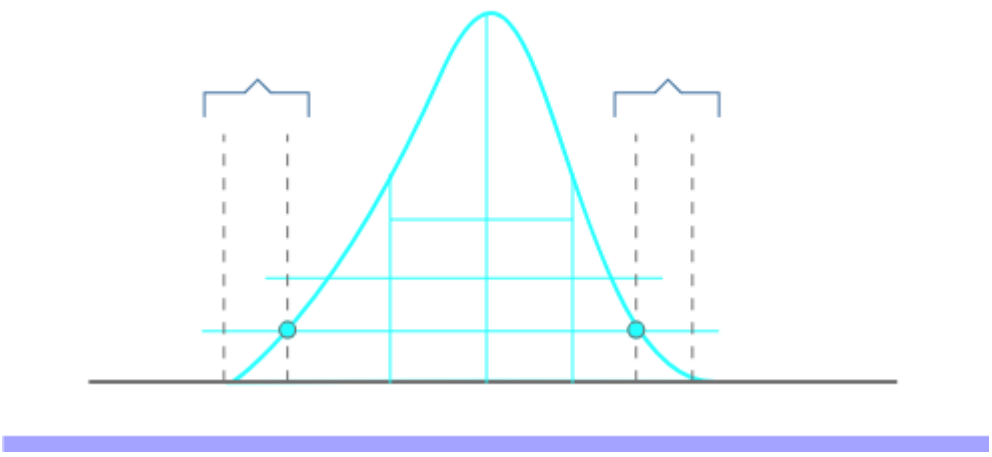
c. Error máximo de la estimación

El error máximo de la estimación se indica con E y es la mitad del ancho del intervalo de confianza. El intervalo de confianza básico para una distribución simétrica se establece para que sea la estimación puntual menos el error máximo de la estimación es menor que el parámetro de población real que es menor que la estimación puntual más el error máximo de

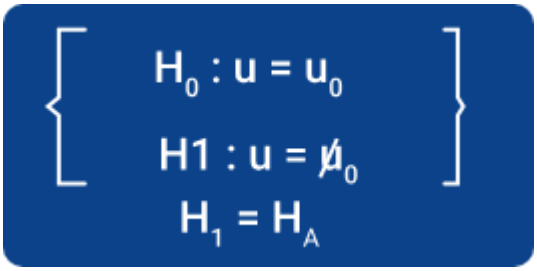
la estimación. Esta fórmula funcionará para medias y proporciones, porque usarán las distribuciones Z o T que son simétricas.

d. Prueba de hipótesis

Cuando evalúas una hipótesis, debes tener en cuenta tanto la variabilidad de tu muestra como el tamaño de la muestra. La prueba de hipótesis se usa generalmente cuando se comparan dos o más grupos. Por ejemplo, puede implementar protocolos para realizar la intubación en pacientes pediátricos en el entorno prehospitalario. Para evaluar si estos protocolos tuvieron éxito en mejorar las tasas de intubación, podría medir la tasa de intubación a lo largo del tiempo en un grupo asignado aleatoriamente a la capacitación en los nuevos protocolos y compararla con la tasa de intubación a lo largo del tiempo en otro grupo de control que no recibió capacitación en los nuevos protocolos.



Cuando evalúas una hipótesis, debes tener en cuenta tanto la variabilidad de tu muestra como el tamaño de la muestra. Con base en esta información, le gustaría hacer una evaluación de si las diferencias que ve son significativas o si es probable que se deba a la casualidad. Esto se hace formalmente a través de un proceso llamado prueba de hipótesis.



Cinco pasos en la prueba de hipótesis:

Especificar la hipótesis nula.

Especificar la hipótesis alternativa.

Establecer el nivel de significación (α).

Calcule la estadística de prueba y el valor P correspondiente.

Esbozando una conclusión.

Para ampliar la información sobre estos importantes pasos en la estimación y la prueba de hipótesis, se invita a revisar:

Especifique la hipótesis nula

La hipótesis nula (H0) es una declaración de ningún efecto, relación o diferencia entre dos o más grupos o factores. En los estudios de investigación, un investigador suele estar interesado en refutar la hipótesis nula.

Ejemplos

- No hay diferencia en las tasas de intubación entre las edades de 0 a 5 años.

- Los grupos de intervención y control tienen la misma tasa de supervivencia (o la intervención no mejora la tasa de supervivencia).
- No existe asociación entre el tipo de lesión y si el paciente recibió o no una vía intravenosa en el entorno prehospitalario.

Especifique la hipótesis alternativa

La hipótesis alternativa (HA) es la afirmación de que existe un efecto o diferencia. Esta suele ser la hipótesis que el investigador está interesado en probar. La hipótesis alternativa puede ser unilateral (solo proporciona una dirección, por ejemplo, inferior) o bilateral. A menudo, utilizamos pruebas de dos caras incluso cuando nuestra verdadera hipótesis es unilateral, porque requiere más evidencia en contra de la hipótesis nula para aceptar la hipótesis alternativa.

Ejemplos

- La tasa de éxito de la intubación difiere según la edad del paciente tratado (bilateral).
- El tiempo hasta la reanimación de un paro cardíaco es menor para el grupo de intervención que para el control (unilateral).
- Existe una asociación entre el tipo de lesión y si el paciente recibió o no una vía intravenosa en el entorno prehospitalario (bilateral).

Establecer el nivel de significancia (α)

El nivel de significancia (denotado por la letra griega alfa α) generalmente se establece en 0,05. Esto significa que hay 5% de probabilidad de que acepte su hipótesis alternativa cuando su hipótesis nula sea realmente cierta. Cuanto menor sea el nivel de significancia, mayor será la carga de la prueba necesaria para rechazar la hipótesis nula o, en otras palabras, para respaldar la hipótesis alternativa.

Calcule la estadística de prueba y el valor P correspondiente

En otra sección presentamos algunas estadísticas de prueba básicas para evaluar una hipótesis. La prueba de hipótesis generalmente utiliza una estadística de prueba que compara grupos o examina asociaciones entre variables. Cuando se describe una sola muestra sin establecer relaciones entre variables, se suele utilizar un intervalo de confianza.

El valor p describe la probabilidad de obtener un estadístico de muestra tan o más extremo por casualidad solo si su hipótesis nula es cierta. Este valor p se determina en función del resultado de su estadística de prueba. Sus conclusiones sobre la hipótesis se basan en su valor p y su nivel de significancia.

Ejemplos

Valor p = 0,01. Esto sucederá 1 de cada 100 veces por pura casualidad si su hipótesis nula es cierta. No es probable que suceda estrictamente por casualidad.

Valor p = 0,75. Esto sucederá 75 de cada 100 veces por pura casualidad si su hipótesis nula es cierta. Es muy probable que ocurra estrictamente por casualidad.

Precauciones sobre los valores p

El tamaño de la muestra afecta directamente su valor p. Los tamaños de muestra grandes producen valores p pequeños incluso cuando las diferencias entre los grupos no son significativas. Siempre debe verificar la relevancia práctica de sus resultados. Por otro lado, un tamaño de muestra demasiado pequeño puede provocar que no se identifique una diferencia cuando realmente existe.

Se debe planificar el tamaño de la muestra con anticipación para tener suficiente información de su muestra para mostrar una relación o diferencia significativa, si existe. Consulte el cálculo de un tamaño de muestra para obtener más información.

Ejemplo 1

Las edades medias fueron significativamente diferentes entre los dos grupos (16,2 años frente a 16,7 años; p = 0,01; n = 1.000). ¿Es esta una diferencia importante? Probablemente no, pero el gran tamaño de la muestra ha dado como resultado un valor p pequeño.

Ejemplo 2

Las edades medias no fueron significativamente diferentes entre los dos grupos (10,4 años frente a 16,7 años; p = 0,40, n = 10). ¿Es esta una diferencia importante? Podría ser, pero debido a que el tamaño de la muestra es pequeño, no se puede determinar con certeza si se trata de una verdadera diferencia o simplemente sucedió debido a la variabilidad natural en la edad dentro de estos dos grupos.

Si realiza una gran cantidad de pruebas para evaluar una hipótesis (lo que se denomina prueba múltiple), debe controlar esto en su designación del nivel de significancia o en el cálculo del valor p. Por ejemplo, si tres resultados miden la efectividad de un fármaco u otra intervención, tendrá que ajustar estos tres análisis.

Sacar una conclusión

Valor p < = nivel de significancia (a) = >, rechaza su hipótesis nula a favor de su hipótesis alternativa. El resultado es estadísticamente significativo.

Valor p > nivel de significancia (a) = > no rechaza su hipótesis nula. El resultado no es estadísticamente significativo.

La prueba de hipótesis no está configurada para que pueda probar absolutamente una hipótesis nula. Por lo tanto, cuando no se encuentra evidencia en contra de la hipótesis nula, no se puede rechazar la hipótesis nula. Cuando se encuentra evidencia suficientemente fuerte en contra de la hipótesis nula, se rechaza la hipótesis nula. Sus conclusiones también se traducen en una declaración sobre su hipótesis alternativa. Cuando presente los resultados de una prueba de hipótesis, incluya también las estadísticas descriptivas en sus conclusiones. Informe valores p exactos en lugar de un rango determinado. Por ejemplo, “La tasa de intubación difirió significativamente según la edad del paciente y los pacientes más jóvenes tienen una tasa más baja de intubación exitosa (p = 0,02)”. Aquí hay dos ejemplos más con la conclusión expresada de varias formas diferentes.

Ejemplo

H0: no hay diferencia en la supervivencia entre el grupo de intervención y el de control.

HA: existe una diferencia en la supervivencia entre el grupo de intervención y el de control.

α = 0,05: aumento del 20% en la supervivencia para el grupo de intervención; valor p = 0,002

Conclusión

Rechace la hipótesis nula a favor de la hipótesis alternativa.

La diferencia en la supervivencia entre el grupo de intervención y el de control fue estadísticamente significativa. Hubo un aumento del 20% en la supervivencia para el grupo de intervención en comparación con el control (p = 0,001)

Ejemplo

H0: no hay diferencia en la supervivencia entre el grupo de intervención y el de control.

HA: existe una diferencia en la supervivencia entre el grupo de intervención y el de control.

α = 0,05: aumento del 5% en la supervivencia entre el grupo de intervención y el de control; valor p = 0,20.

Conclusión

No rechace la hipótesis nula.

La diferencia en la supervivencia entre el grupo de intervención y el de control no fue estadísticamente significativa.

No hubo un aumento significativo en la supervivencia para el grupo de intervención en comparación con el control (p = 0,20).

1.6 Formulario de muestreo y estimación

Las fórmulas que se usan para calcular la media muestral y todas las demás estadísticas muestrales son ejemplos de fórmulas de estimación o estimadores. El valor particular que calculamos a partir de observaciones de muestra utilizando un estimador se denomina estimación. Por ejemplo, el valor calculado de la media muestral en una muestra determinada se denomina estimación puntual de la media poblacional. Las tres propiedades deseables de un estimador son:

Insesgado

Tu valor esperado es igual al parámetro que se está estimando.

Eficiencia

Tiene la varianza más baja en comparación con otros estimadores insesgados del mismo parámetro.

Coherencia

A medida que aumenta el tamaño de la muestra, el error muestral disminuye y las estimaciones se acercan al valor real.

En el Anexo 4 - Formulario de muestreo, se puede encontrar una clasificación por tablas de fórmulas más utilizadas y necesarias para llevar a cabo todo lo relacionado a procesos de muestreo de datos.

1.7 Probabilidad de sucesos compatibles e incompatibles

Dos o más eventos son compatibles, si pueden cumplirse simultáneamente; es decir, si tienen al menos un resultado común. Por el contrario, son incompatibles o mutuamente excluyentes y su intersección es el conjunto vacío Ø si se analiza el siguiente experimento de lanzar un dado.

Comencemos con el siguiente experimento: tiramos un dado de seis caras y vemos cuál es el resultado. Consideremos los siguientes eventos A= {2, 3}, B= {1,2}, C= {5}.

Observamos que, si extraemos 2, luego A está satisfecho, así como B. Decimos que los eventos son compatibles, esto significa que pueden ocurrir simultáneamente. Por el contrario, los eventos B y C son incompatibles, ya que los dos no pueden suceder simultáneamente.

Para ver cuando dos eventos son compatibles o no, podemos observar que A y B tienen un elemento común: 2, por lo tanto, serán compatibles. De lo contrario, A y C no tienen ningún elemento común y, por tanto, son incompatibles. Expresamos esto diciendo que dos eventos A y B son compatibles si:

$$A \cap B \neq \emptyset$$

y, por el contrario, son incompatibles si:

$$A \cap B = \emptyset$$

Si tenemos tres o más eventos, decimos que son incompatibles de dos en dos si dos eventos son incompatibles (de manera similar, son compatibles de dos en dos si dos eventos son compatibles). En nuestro caso A, B y C no son incompatibles de dos en dos, ya que, aunque A y C, al igual que B y C son incompatibles, A y B son compatibles.

Probabilidad de sucesos compatibles e incompatibles

Dos o más eventos son compatibles, si pueden cumplirse simultáneamente; es decir, si tienen al menos un resultado común. Por el contrario, son incompatibles o mutuamente excluyentes y su intersección es el conjunto vacío Ø. Si volvemos a centrarnos en el experimento de lanzar un dado.

Eventos

A = {2,3} B = {1,2} C = {4,5} cumplen con lo siguiente:

A y B son compatibles, y B y C son incompatibles, B ∩ C = Ø

1.8 Cálculo de probabilidades y probabilidad condicionada

Una probabilidad es un modelo representado matemáticamente de un fenómeno aleatorio. Las probabilidades pueden ser marginales, conjuntas o condicionales.

Comprender sus diferencias y cómo manipularlas es clave para tener éxito en la comprensión de los fundamentos de las estadísticas.

Cálculo de probabilidades de eventos compatibles e incompatibles

Se explicará cómo calcular la probabilidad de un evento aislado con la ley de Laplace y calcular la probabilidad de la unión de dos eventos cuando son compatibles y cuándo son incompatibles.

Ley de Laplace:

No es más que la fórmula para calcular la probabilidad de que ocurra un evento aislado. Cuando en un experimento aleatorio, todos los eventos tienen la misma probabilidad de ocurrir, la probabilidad de que ocurra un evento A es:

$$\frac{N^0 \text{ casos favorables a A}}{N^0 \text{ casos posible}}$$

En el numerador colocamos el número de casos favorables para que ocurra el evento A y en el denominador colocamos el número de eventos posibles. Por ejemplo, en un dado, ¿cuál es la probabilidad de obtener un 2?

En este caso, el número de casos favorables es 1, ya que el dado tiene solo un 2. El número de casos posibles es 6, que son los números que tiene un dado. Por lo tanto, la probabilidad de que salga un 2 se puede escribir como P(2) y es igual a:

$$P(2) = \frac{1}{6} = 0{,}16$$

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

Siendo 1 la probabilidad del evento seguro (siempre ocurrirá) y 0 es la probabilidad del evento imposible (nunca ocurrirá).

Cómo calcular la probabilidad de ocurrencia de A o B, si los eventos son incompatibles: Si A y B son dos eventos incompatibles, es decir, no pueden ocurrir al mismo tiempo, la probabilidad de que ocurra A o B será la suma de las probabilidades de que cada evento ocurra por separado.

$$P(A \cup B) = P(A) + P(B)$$

La probabilidad de obtener una bola blanca es:

$$P(B) = \frac{1}{4}$$

La probabilidad de obtener una bola negra es:

$$P(N) = \frac{1}{4}$$

Por lo tanto, la probabilidad de obtener una bola blanca o una bola negra es:

$$P(BUN) = P(B) + P(N) =$$

$$= \frac{1}{4} + \frac{1}{4} = \frac{2}{4} = 0{,}5$$

Cómo calcular la probabilidad de que ocurra A o B, si los eventos son compatibles: si A y B son dos eventos compatibles, es decir, pueden ocurrir al mismo tiempo, entonces la probabilidad de que ocurran A o B será:

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

P (A ∩ B) dice: la probabilidad de que ocurran A y B.

Esta vez, la suma de las probabilidades de que cada evento ocurra por separado debe restarse de la probabilidad de que los dos eventos ocurran al mismo tiempo. Por ejemplo, calcular la probabilidad de que al lanzar un dado el número obtenido sea par o que sea un 4. En este caso, el evento “obtener un número par” y el evento “obtener un 4” son compatibles, porque si obtenemos un 4 están sucediendo ambos eventos a la vez. Por lo tanto, la probabilidad de obtener 4 o un número par se calculará con la fórmula:

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

La probabilidad de obtener un número par es:

$$P\,(par) = \frac{3}{6} = 0{,}5$$

La probabilidad de obtener un 4 es:

$$P\,(4) = \frac{1}{6} = 0{,}16$$

La probabilidad de obtener un número par y un 4 es:

$$P\,(par \cap 4\,) = \frac{1}{6} = 0{,}16$$

Solo hay una posibilidad entre 6 ya que 4 es el único número que cumple ambos eventos al mismo tiempo. Finalmente, la probabilidad de obtener un número par o 4 es:

$$P\,(par\ U\,4\,) = P\,(par) + P(4) - P\,(par \cap\ 4) =$$

Que la sustitución de cada término por su valor nos queda a nosotros:

$$= 0{,}5 + 0{,}16 - 0{,}16 = 0{,}5$$

Probabilidad condicionada

En la teoría de la probabilidad, la probabilidad condicionada o condicional es una medida de la probabilidad de que ocurra un evento, dado que ya ha ocurrido otro evento (por suposición, presunción, afirmación o evidencia). Veamos:

Si el evento de interés es A y se sabe o se supone que ocurrió el evento B, “la probabilidad condicional de A dado B “, o “la probabilidad de A bajo la condición B “, generalmente se escribe como P (A | B) y ocasionalmente PB(A). Esto también se puede entender como la fracción de probabilidad B que se cruza con A:

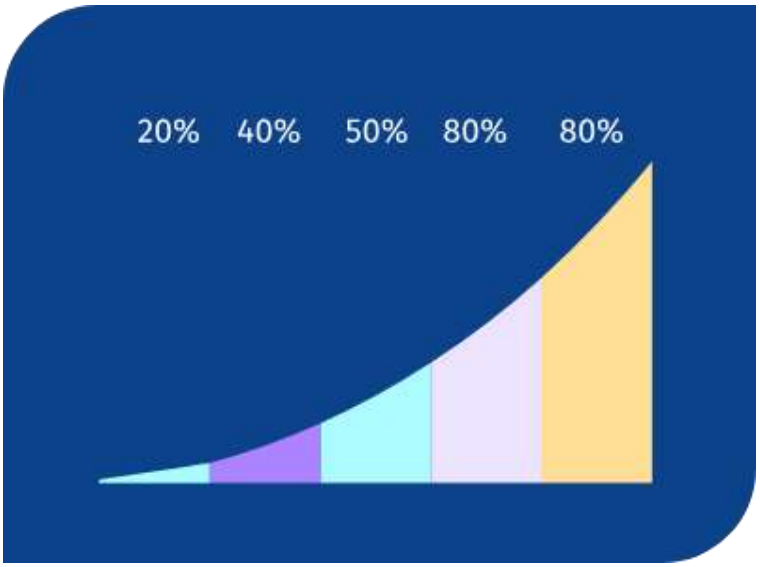
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Por ejemplo…

La probabilidad de que una persona determinada tenga tos en un día determinado puede ser solo del 5%. Pero si sabemos o asumimos que la persona está enferma, es mucho más probable que esté tosiendo. Por ejemplo, la probabilidad condicional de que alguien no esté bien tosiendo podría ser del 75%, en cuyo caso tendríamos que P (Tos) = 5% y P (Tos | Enfermo) = 75%. Aunque, no tiene que haber relación o dependencia entre A y B, y no tienen que ocurrir simultáneamente.

Teniendo en cuenta que…

P (A | B) puede o no ser igual a P (A) (la probabilidad incondicional de A). Si P (A | B) = P (A), entonces se dice que los eventos A y B son independientes: en tal caso, el conocimiento sobre cualquiera de los eventos no altera la probabilidad de los demás. P (A | B) (la probabilidad condicional de A dado B) típicamente difiere de P (B | A).



Por ejemplo…

Si una persona tiene fiebre del dengue, podría tener 90% de probabilidades de dar positivo en la prueba de la enfermedad. En este caso, lo que se está midiendo es que si ha ocurrido el evento B (tener dengue), la probabilidad de A (resultado positivo) dado que B ocurrió es del 90%: P (A | B) = 90%. Alternativamente, si una persona da positivo en la prueba del dengue, es posible que solo tenga 15% de probabilidades de tener esta rara enfermedad debido a las altas tasas de falsos positivos.

En este caso…

La probabilidad del evento B (tener dengue) dado que el evento A (resultado positivo) que ha ocurrido es 15%: P (B | A) = 15% Ahora debería ser evidente que igualar falsamente las dos probabilidades puede conducir a varios errores de razonamiento, lo que comúnmente se ve a través de falacias de tasa base.

Si bien las probabilidades condicionales pueden proporcionar información extremadamente útil…

A menudo se proporciona o se dispone de información limitada. Por lo tanto, puede ser útil revertir o convertir una probabilidad de condición usando el teorema de Bayes:

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

Otra opción es mostrar probabilidades condicionales en una tabla de probabilidad condicional para iluminar la relación entre eventos.

1.9 Combinatoria: variaciones, permutaciones y combinaciones

La combinatoria se ocupa de conjuntos finitos y sus objetos básicos son permutaciones, variaciones y combinaciones (y se refieren tanto a las operaciones como a los resultados de estas operaciones). Dado que siempre es posible identificar m y n elementos conjuntos con los conjuntos {1, 2, …, m} y {1, 2, …, n}, es suficiente considerar estos conjuntos, y esta situación es una a uno estándar. Recordamos estas tres nociones (en sus dos apariencias: cuándo se permiten las repeticiones y cuándo no).

Esto se hace comenzando con variaciones sin repeticiones (validando esta decisión por una relación directa con la cardinalidad de todos los mapas finitos), aunque se podría comenzar con combinaciones o con permutaciones. Al cumplir con la permutación, hay que prestar una especial atención a las operaciones con factorial.

En el Anexo 5 - Combinatoria - Variaciones, permutaciones y combinaciones, se encuentran las respectivas fórmulas para un caso determinado y su respectivo ejemplo.

2. Métodos para procesar, graficar y analizar datos

Después de la recolección de datos, limpieza inicial y clasificación de los mismos se sigue un conjunto de procedimientos para obtener los datos / información deseados del trabajo de campo que se desea ejecutar, para procesar, graficar y analizar los hechos de manera lógica y científica.

2.1 Métodos de investigación

Estos métodos son procesos que se utilizan para recopilar datos. Puede utilizar estos datos para analizar métodos o procedimientos actuales y encontrar información adicional sobre un tema. Los profesionales utilizan métodos de investigación mientras estudian medicina, comportamiento humano y otros temas académicos. Hay dos categorías principales de métodos de investigación: métodos de investigación cualitativa y métodos de investigación cuantitativa; ahora se ejemplifica la mirada cuantitativa:

Los métodos de investigación cuantitativa…

Implican el uso de números para medir datos. Los investigadores pueden utilizar el análisis estadístico para encontrar conexiones y significado en los datos. Los métodos de investigación cualitativa implican explorar información y datos no numéricos. Estos métodos de investigación también examinan cómo las personas pueden conectar el significado con sus experiencias y emociones.

Tipos de métodos de investigación:

Método sintético

Busca la reconstrucción de componentes dispersos de un objeto o evento para estudiarlos en profundidad y crear un resumen de cada detalle. El proceso de este método se desarrolla de lo abstracto a lo concreto para aglutinar cada segmento que conforma una unidad y poder comprenderlo.

Método analítico

Se encarga de desagregar los apartados que componen el conjunto del caso a estudiar, estableciendo las relaciones de causa, efecto y naturaleza. A partir de los análisis realizados, se pueden generar analogías y nuevas teorías para comprender los comportamientos.

Se desarrolla en la comprensión de lo concreto a lo abstracto, descomponiendo los elementos que constituyen la teoría general para estudiar con mayor profundidad cada elemento por separado y, de esta manera, conocer la naturaleza del fenómeno de estudio para revelar su esencia. Mediante el razonamiento y la síntesis se enfatiza la profundidad del análisis de forma metódica y concisa para obtener un conocimiento profundo de cada parte y particularidad de lo estudiado.

Método inductivo

Con este método se pueden analizar situaciones particulares a través de un estudio individual de los hechos que formule conclusiones generales, que ayudan al descubrimiento de temas y teorías generalizadas que parten de la observación sistemática de la realidad. Es decir, se refiere a la formulación de hipótesis a partir de lo vivido y observado de los elementos de estudio para definir leyes generales. Consiste en la recogida de datos ordenados en variables en busca de regularidades.

Método deductivo

Se refiere a un método desde lo general para enfocarse en lo específico a través de razonamientos lógicos e hipótesis que pueden sustentar conclusiones. Este proceso se basa en los análisis anteriores, leyes y principios validados y probados para ser aplicados a casos particulares.

En este método, todo el esfuerzo de investigación se basa en las teorías recopiladas, no en lo observado o experimentado; forma parte de una premisa para delinear y concluir la situación de estudio, deduciendo el camino a seguir para implementar las soluciones.

Método comparativo

Es una búsqueda de similitudes y comparaciones sistemáticas que sirve para la verificación de hipótesis para encontrar parentesco y se basa en la documentación de múltiples casos para realizar análisis comparativos. Básicamente consiste en poner dos o más elementos uno al lado del otro para encontrar diferencias y relaciones y así definir un caso o problema y poder actuar en el futuro.

El uso de la comparación es útil para comprender un tema, ya que puede conducir a nuevas hipótesis o teorías de crecimiento y mejora. Tiene varias etapas en las que destaca la observación, descripción, clasificación, comparación en sí y su conclusión.

2.2 Métodos de procesamiento de datos más conocidos

La recopilación, manipulación y procesamiento de datos recopilados para el uso requerido se conoce como procesamiento de datos, es una técnica realizada normalmente por una computadora; el proceso incluye la recuperación, transformación o clasificación de

información. Sin embargo, el procesamiento de datos depende en gran medida de lo siguiente:

El volumen de datos que deben procesarse.

La complejidad de las operaciones de procesamiento de datos.

Capacidad y tecnología incorporada del sistema informático respectivo.

Habilidades técnicas.

Limitaciones de tiempo.

Para completar la temática, se invita a realizar un recorrido por los diferentes métodos de procesamiento de datos:

a. Programación de usuario único

Generalmente lo hace una sola persona para su uso personal. Esta técnica es adecuada, incluso, para oficinas pequeñas.

b. Programación múltiple

Esta técnica proporciona la posibilidad de almacenar y ejecutar más de un programa en la Unidad Central de Procesamiento (CPU) simultáneamente. Además, la técnica de programación múltiple aumenta la eficiencia de trabajo global del ordenador respectivo.

c. Procesamiento en tiempo real

Esta técnica facilita que el usuario tenga un contacto directo con el sistema informático, y el procesamiento de datos. Esta técnica también se conoce como el modo directo o la técnica del modo interactivo y se desarrolla exclusivamente para realizar una tarea. Es una especie de procesamiento en línea, que siempre permanece en ejecución.

d. Procesamiento en línea

Esta técnica facilita la entrada y ejecución de datos directamente; por lo tanto, no se almacena ni se acumula primero y luego se procesa. La técnica está desarrollada de tal manera que reduce los errores de entrada de datos, ya que valida los datos en varios puntos y también asegura que solo se ingresen los datos corregidos. Esta técnica se usa ampliamente para aplicaciones en línea.

e. Procesamiento de tiempo compartido

Esta es otra forma de procesamiento de datos en línea que facilita a varios usuarios compartir los recursos de un sistema informático en línea. Esta técnica se adopta cuando se necesitan resultados rápidamente. Además, como sugiere el nombre, este sistema se basa en el tiempo.

A continuación, se presentan algunas de las principales ventajas del procesamiento de tiempo compartido:

Se pueden atender varios usuarios simultáneamente.

Todos los usuarios tienen casi la misma cantidad de tiempo de procesamiento.

Existe la posibilidad de interacción con los programas en ejecución.

f. Procesamiento distribuido

Esta es una técnica de procesamiento de datos especializada en la que varias computadoras (que están ubicadas de forma remota) permanecen interconectadas con una sola computadora host que forma una red de computadoras.

Todos estos sistemas informáticos permanecen interconectados con una red de comunicaciones de alta velocidad. Esto facilita la comunicación entre computadoras, sin embargo, el sistema informático central mantiene la base de datos maestra y supervisa en consecuencia.

2.3 Tipos de gráficas para el análisis de datos

Es necesario conocer las diferentes gráficas para realizar un análisis de datos; a saber:

Tipos de gráficas para el análisis de datos

Las investigaciones muestran que creamos 2,5 trillones de bytes de datos todos los días. ¿Qué tipos de visualización de datos utiliza para digerir correctamente todos esos datos?

Si bien esta es una cifra asombrosa,

Solo aumentará a medida que evoluciona Internet de las cosas. De hecho, el 90% de los datos del mundo se generaron solo en los últimos dos años. Con tanta información accesible al alcance de la mano, es importante comprender cómo organizarla en información analizable y procesable.

Sin embargo,

Si administra varios activos de contenido con múltiples fuentes de datos, puede resultar difícil determinar cómo dar forma a su estrategia de análisis. Aquí es donde ayuda conocerlos mejores tipos de visualización de datos.

La visualización de datos…

Es el proceso de convertir sus datos en representaciones gráficas que comunican relaciones lógicas y conducen a una toma de decisiones más informada.

En resumen,

La visualización de datos es la representación de datos en un formato gráfico o pictórico. Permite a los tomadores de decisiones clave ver análisis complejos en un diseño visual, para que puedan identificar nuevos patrones o comprender conceptos desafiantes.

Desde las métricas del sitio web y el desempeño del equipo de ventas hasta los resultados de las campañas de marketing y las tasas de adopción de productos, existe una variedad de puntos de datos que su organización necesita rastrear. Cuando se tienen las manos ocupadas haciendo malabares con varios proyectos a la vez, se necesita un método de informes rápido y eficaz que le permita expresar un punto claro. ¿Se sabe qué tipo de método de visualización de datos utilizar?

Algunos de los tipos más comunes de gráficos y gráficos de visualización de datos incluyen:

- Gráfico de columnas.
- Gráfico de barras.
- Gráfico de barras apiladas.
- Gráfico de columnas apiladas.
- Gráfico de área.
- Gráfico de doble eje.
- Gráfico de líneas.
- Gráfico de Mekko.
- Gráfico circular.

- Gráfico de cascada.
- Gráfico de burbujas.
- Gráfico de diagrama de dispersión.
- Gráfico de viñetas.
- Gráfico de embudo.
- Mapa de calor.

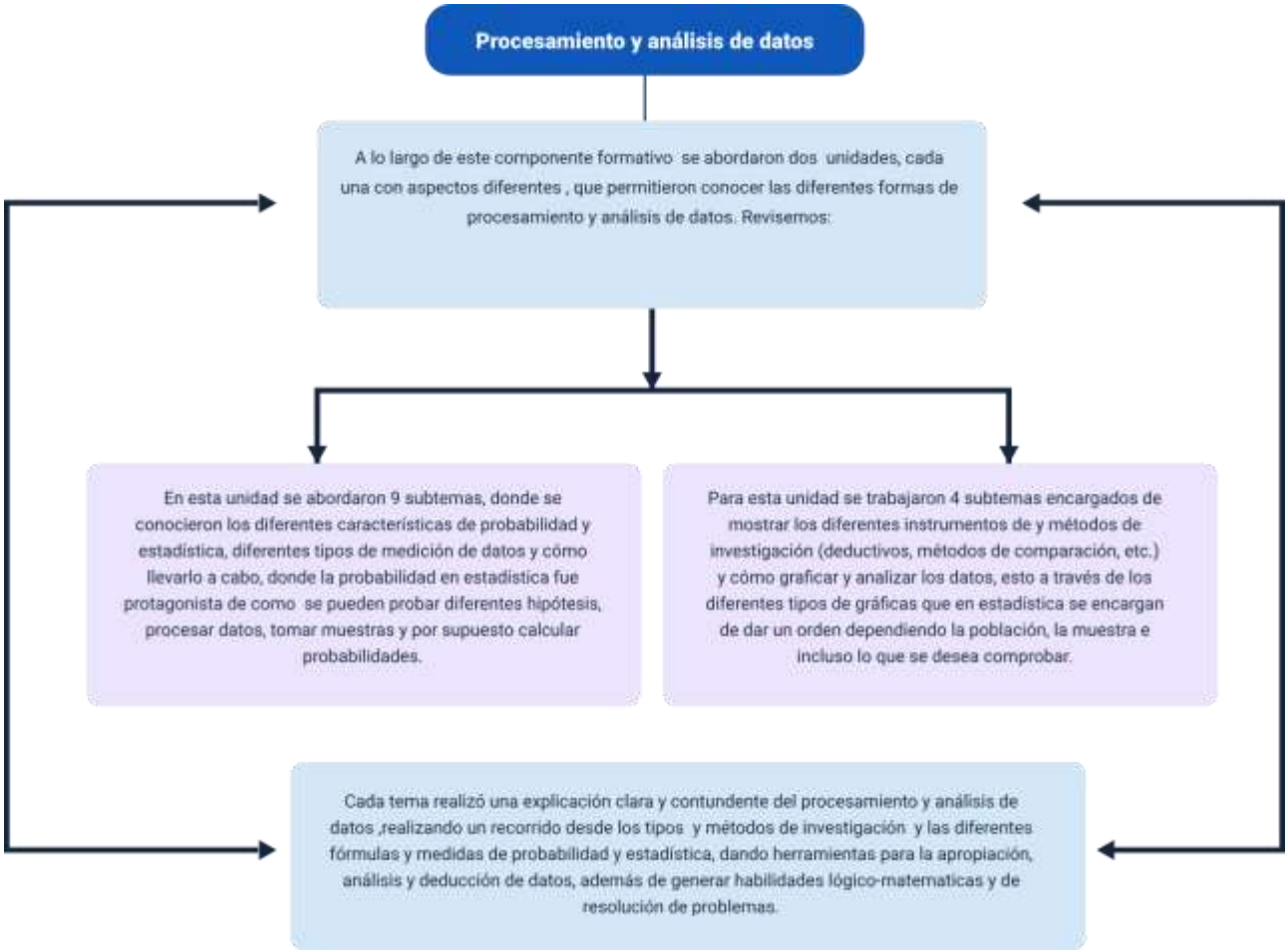
Si bien todos sirven para agilizar y mejorar la interpretación de datos, no todos son apropiados para el mismo trabajo. Elegir la ayuda visual adecuada es la clave para evitar la confusión del usuario y asegurarse de que su análisis sea preciso. Ahora se analizarán 10 de estos 15 tipos de tablas y gráficos.

Explicación de 10 tipos de visualización de datos

Existen innumerables tipos diferentes de tablas, gráficos y otras técnicas de visualización que pueden ayudar a los analistas a representar y transmitir datos importantes. Echemos un vistazo a 10 de los más comunes

Síntesis

El procesamiento y análisis de datos son elementos clave en la toma de decisiones basada en información cuantitativa. Dentro de este campo, se abordan diversos subtemas, como la probabilidad y estadística, que incluyen el estudio de variables continuas, distribuciones bidimensionales, discretas y continuas, así como muestreo, estimación, prueba de hipótesis y combinatoria. Además, se exploran métodos para procesar, graficar y analizar datos, como métodos de investigación, tipos de gráficas y herramientas software ampliamente utilizadas en esta área. Estas habilidades permiten comprender mejor los datos, identificar patrones y tomar decisiones informadas en diversos campos.



Material complementario

Tema	Referencia APA del Material	Tipo de material	Enlace del Recurso o Archivo del documento material
2.2. Métodos de procesamiento de datos más conocidos.	Tesis de Cero a 100 – TV. (2019). *Guía básica para análisis estadístico de datos | Parte 1*. [Video]. YouTube.	**Video**	**https://www.youtube.com/watch?v=Wn9gQjWNdiY**
2.2. Métodos de procesamiento de datos más conocidos.	Tesis de Cero a 100 – TV (2019). *Guía básica análisis estadístico de datos | Parte 2*. [Video]. YouTube.	**Video**	**https://www.youtube.com/watch?v=CnxMorGipGw**
2.2. Métodos de procesamiento de datos más conocidos.	Comunicación numérica. (2020). *Fundamentos del análisis de datos para toma de decisiones*. [Video]. YouTube.	**Video**	**https://www.youtube.com/watch?v=qvZxvMWMvDo**
2.3. Tipos de gráficas para el análisis de datos.	A2 Capacitación: Excel. (2019). *Minicurso de business intelligence en Excel - tablas dinámicas, gráficas y dashboards - Parte 1*. [Video]. YouTube.	**Video**	**https://www.youtube.com/watch?v=TgcGa0cnlpg**
2.3. Tipos de gráficas para el análisis de datos.	A2 Capacitación: Excel. (2019). *Minicurso de business intelligence en Excel - tablas dinámicas, gráficas y dashboards - - Parte 2*. [Video]. YouTube.	**Video**	**https://www.youtube.com/watch?v=ucJM-wpuoKg**

Herramientas *software* más conocidas para procesar, graficar y analizar datos.	Las mates fáciles. (2020). *Probabilidad: sucesos compatibles e incompatibles – Explicación.* [Video]. YouTube.	**Video**	**https://www.youtube.com/watch?v=njjrzLk5RB0**

Glosario

Amplitud o rango: diferencia entre el valor máximo y mínimo de los valores de una variable se encuentran comprendidos el 100% de los valores muestrales.

Área bajo la curva (entre dos puntos): si la curva viene dada por una función de densidad teórica, representa la probabilidad de que la variable aleatoria tome un valor dentro del intervalo determinado por esos dos puntos.

Características: propiedades de las unidades o elementos que componen las muestras. Se miden mediante variables. Se asume que los individuos presentan diferentes características.

Causalidad: relación entre causa y efecto. Generalmente identificados como variables. No hay que confundir causalidad con correlación. La correlación mide la similitud estructural numérica entre dos variables. Normalmente la existencia de correlación es condición necesaria para la causalidad.

Coeficiente de correlación: estadístico que cuantifica la correlación. Sus valores están comprendidos entre 1 y 1

Coeficiente de variación: medida de dispersión relativa. No tiene unidades y se calcula dividiendo la cuasi-desviación típica entre la media muestral. Se suele expresar en tanto por ciento.

Contraste bilateral: contraste de hipótesis en la que la hipótesis alternativa da opción a igualdad o superioridad.

Contraste de hipótesis: proceso estadístico que se sigue para la toma de decisiones a partir de la información de la muestra. Comparando el valor del estadístico experimental con el valor teórico rechazamos o no la hipótesis nula.

Correlación: concordancia entre dos variables según el sentido de la relación de estas en términos de aumento o disminución.

Datos censurados: en análisis de supervivencia son datos donde no se conoce el tiempo total hasta la aparición del fracaso / éxito bien porque el individuo se retiró del estudio, bien porque se acabó el estudio (datos con censura administrativa). Existen datos censurados por la izquierda y por la derecha.

Distribución de datos: en la realización de un experimento, corresponde a la recogida de los datos experimentales para cada individuo y cada variable.

Escala: la distribución de datos puede recogerse en distintas escalas: nominal, dicotómica, discreta o continua.

Estimación: técnicas estadísticas que a partir de la información de la estadística descriptiva pretenden conocer cómo es la población en global. Existen técnicas de estimación puntuales y por intervalos de confianza.

Factor de clasificación: variable que se usa para clasificar los datos experimentales en grupos. Los factores de clasificación son variables nominales. Cada factor de clasificación se compone de niveles. Así la variable “fumador” codificada como “nunca”, “exfumador”, “fumador actual” es un factor de clasificación con tres niveles.

Hipótesis: cualquier teoría que formule posibles líneas de trabajo experimental. Ver hipótesis nula y alternativa.

Imprecisión: error que se comete en la predicción.

Modelo: intento matemático / estadístico para explicar una variable respuesta por medio de una o más variables explicativas o factores.

Muestras: subgrupos de observaciones de la población de estudio.

Observación: sinónimo de caso, registro e indivíduo.

Parámetros: son valores desconocidos de características de una distribución teórica. El objetivo de la estadística es estimarlos bien dando un valor concreto, bien dado un intervalo confidencial.

Rango: diferencia entre el valor máximo y mínimo de una muestra o población. Solo es válido en variables continuas.

Sesgo: la diferencia entre el valor del parámetro y su valor esperado. También se utiliza en contraposición de aleatorio, así una muestra sesgada es no aleatoria.

Simetría: medida que refleja si los valores muéstrales se extienden o no de igual forma a ambos lados de la media.

Transformaciones: cambios de escala con el propósito de conseguir linealidad, normalidad en los datos.

Valores numéricos: resultados de las variables para cada individuo en la muestra de estudio. Su naturaleza puede ser nominal, dicotómica, ordinal o continua.

Variable: objeto matemático que puede tomar diferentes valores. Generalmente asociado a propiedades o características de las unidades de la muestra. Lo contrario de variable es constante.

Variable aleatoria: variable cuyo resultado varía según la muestra según una distribución de probabilidad.

Variable continua: aquella que puede tomar una infinidad de valores, de forma que dados dos valores cualquiera, también pueda tomar cualquier valor entre dichos valores.

Variable discreta: variable que toma un número finito o infinito de valores, de forma que no cubre todos los posibles valores numéricos entre dos dados, en contraposición de las continuas.

Variables: describen características en las observaciones realizadas.

Referencias bibliográficas

Devore, J. (2008). Probabilidad y estadística para ingeniería y ciencias. Cengage Learning. https://intranetua.uantof.cl/facultades/csbasicas/matematicas/academicos/jreyes/DOCENCIA/APUNTES/APUNTES%20PDF/Probabilidad%20y%20Estadistica%20para%20Ingenieria%20y%20Ciencias%20-%20Jay%20Devore%20-%20Septima%20Edicion.pdf

Orellana, L., D., y Sánchez, G., M. (2006). Técnicas de recolección de datos en entornos virtuales más usadas en la investigación cualitativa. RIE - Revista de Investigación Educativa Asociación Interuniversitaria de Investigación Pedagógica Murcia, 24(1), 205-222. https://www.redalyc.org/pdf/2833/283321886011.pdf

Rustom, J., A. (2012). Estadística descriptiva, probabilidad e inferencia. Una visión conceptual y aplicada. Universidad de Chile. https://repositorio.uchile.cl/bitstream/handle/2250/120284/Rustom_Antonio_Estadistica_descriptiva.pdf

Tanvi, B., Gautam, S., & Vijay, M. (2020). Determining Sufficient Volume of Data for Analysis with Statistical Framework. Springer https://www.springerprofessional.de/en/determining-sufficient-volume-of-data-for-analysis-with-statisti/18346230

Torres, I., M., Paz, K., G., y Salazar, F. (s. f.). Métodos de recolección de datos para una investigación. Boletín Electrónico No. 03. https://fgsalazar.net/LANDIVAR/ING-PRIMERO/boletin03/URL_03_BAS01.pdf

Créditos

Nombre	Cargo	Regional y Centro de Formación
Claudia Patricia Aristizábal	Responsable del Equipo	Dirección General
Norma Constanza Morales Cruz	Responsable de línea de producción	Regional Tolima Centro de Comercio y Servicios
Jaime Mauricio Peñaloza Trespalacios	Experto Técnico	Regional Distrito Capital - Centro Electricidad Electrónica y Telecomunicaciones
Leidy Carolina Arias Aguirre	Diseñadora instruccional	Regional Distrito Capital- Centro de Diseño y Metrología
Carolina Coca Salazar	Revisora Metodológica y Pedagógica	Regional Distrito Capital - Centro de Diseño y Metrología
José Gabriel Ortiz Abella	Corrector de estilo	Regional Distrito Capital - Centro de Diseño y Metrología
Juan Gilberto Giraldo Cortés	Diseñador instruccional	Regional Tolima – Centro de Comercio y Servicios
María Inés Machado López	Metodóloga	Regional Tolima – Centro de Comercio y Servicios
José Yobani Penagos Mora	Diseñador Web	Regional Tolima - Centro de Comercio y Servicios
Oscar Daniel Espitia Marín	Desarrollador Fullstack	Regional Tolima - Centro de Comercio y Servicios
Gilberto Junior Rodríguez Rodríguez	Storyboard e Ilustración	Regional Tolima - Centro de Comercio y Servicios
Nelson Iván Vera Briceño	Producción audiovisual	Regional Tolima - Centro de Comercio y Servicios

Oleg Litvin	Animador	Regional Tolima - Centro de Comercio y Servicios
Francisco Javier Vásquez Suarez	Actividad Didáctica	Regional Tolima - Centro de Comercio y Servicios
Jorge Bustos Gómez	Validación y vinculación en plataforma LMS	Regional Tolima - Centro de Comercio y Servicios
Gilberto Naranjo Farfán	Validación de contenidos accesibles	Regional Tolima - Centro de Comercio y Servicios