Análisis e Interpretación de Datos

Mariana-Edith Miranda-Varela

15-sept-2024

Introducción a la estadística

¿Qué es la estadística?

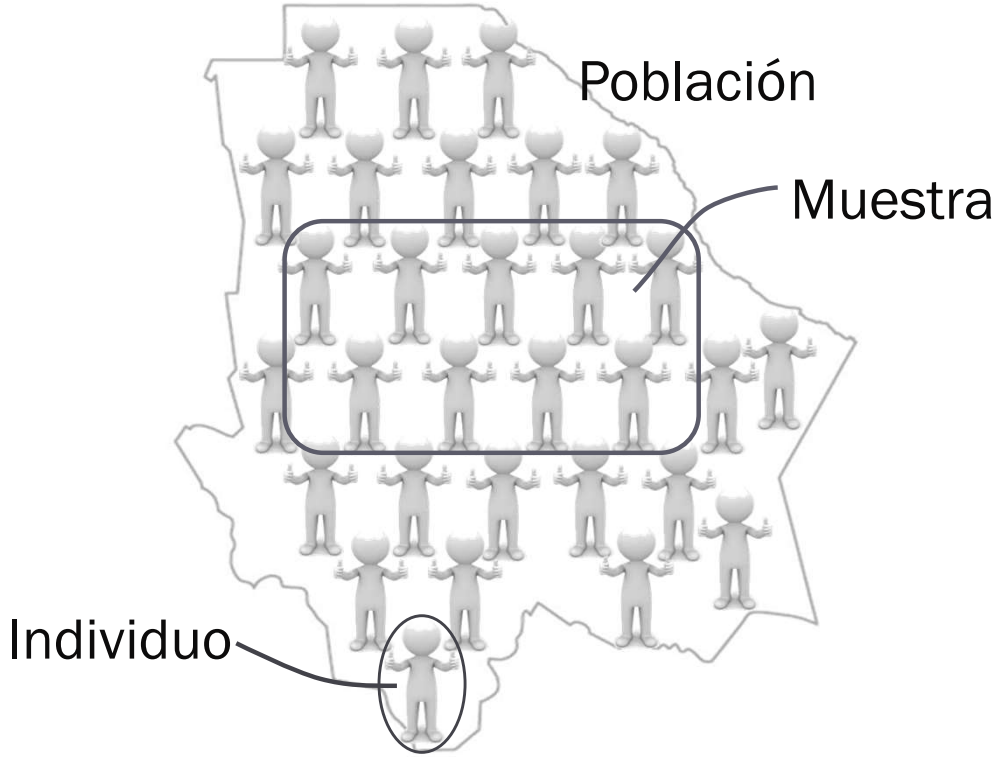
- Ciencia que maneja los datos a través de un proceso

- Ciencia que nos permite aprender de los datos

- Uso pretendido

Estadística descriptiva	Describe una población con base en la información recolectada de su muestra
Estadística inferencial	Pretende establecer conclusiones sobre la población

Población, muestra y muestreo



FUENTE: https://www.freepik.es/vector-premium/mapa-estado-chihuahua-division-administrativa-pais-mexico-ilustracion-vectorial_52098183.htm
https://www.istockphoto.com/es/foto/3-d-hombre-blanco-mostrando-pulgar-hacia-arriba-gm471353682-62728636

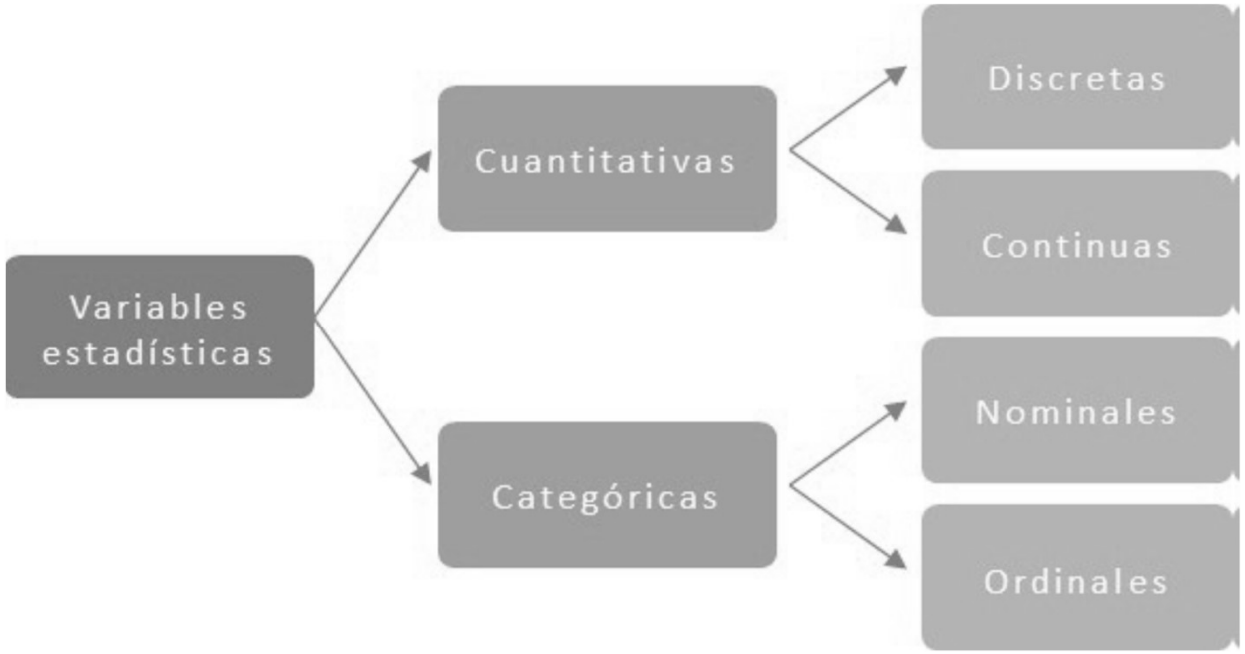
Población, muestra y muestreo

- Muestra representativa – debe contener diversidad, esto es, similar a la de la población.

- Inferencia estadística – es la extrapolación de características y propiedades de la muestra a las de la población

- Error de muestreo – se presenta al momento de inferir o extrapolar a partir de la muestra, el cómo es la realidad de la población

Tipos de variables estadísticas



Clasificación con base en su enfoque metodológico

- Variables dependientes – variable explicada o respuesta
- Variables independientes – variable explicativa o predictora

Tipos de variables estadísticas

Clasificación con base en su enfoque metodológico

- Variables dependientes – variable explicada o respuesta
- Variables independientes – variable explicativa o predictora

Variables intermedias y omitidas

- No están contempladas en el estudio o modelo
- Actúan como variables explicativas de la variable dependiente
- Se deben identificar para no establecer asociaciones y presuponer causalidades infundadas.

Variables dicotómicas

- Describen si ocurre algo o no
- 1 (ocurre), 0 (no ocurre)

Diseño de experimentos

- Observacional
  - Como su nombre lo indica sólo se observan los datos recolectados, es decir, no se intervienen ni alteran a los individuos de alguna manera.

- Experimentales
  - Se realiza algún tratamiento en los individuos para posteriormente observar sus efectos en los sujetos (unidades experimentales)

Razonamiento estadístico

¿Las conclusiones se extraen directa y naturalmente de los datos?

¿Los gráficos resumen de manera adecuada los datos?

¿Cuál fue el tipo de muestreo empleado para obtener los datos?

¿Quién es la fuente de datos?

¿Cuál es el objetivo de estudio?

Sesgo

FUENTE: https://www.istockphoto.com/pt/vetorial/3d-character-cant-choose-between-two-different-paths-gm689115930-126829355

Representando los datos: distribución de frecuencias

Frecuencias – es el conteo de datos, es decir, el número de veces que se repite un valor o categoría de una variable.

Modalidades	Frecuencias (absolutas)	Frecuencias relativas	Frecuencias absolutas acumuladas	Frecuencias relativas acumuladas
1	n_1	$f_1=n_1/N$	N_1	$F_1=N_i/N$
2	n_2	f_2	$N_2=n_1+n_2$	F_2
…	…	…	…	…
K	n_k	f_k	N	1
SUMA	N	1		

Frecuencias de una variable continua

- Se emplean intervalos
- Límite inferior y límite superior
- Marca de clase – valor promedio o representante de dicho intervalo

$$x_i = marca\ de\ clase = \frac{L_{i-1} + L_i}{2}$$

Gráficas básicas

Gráfico de barras

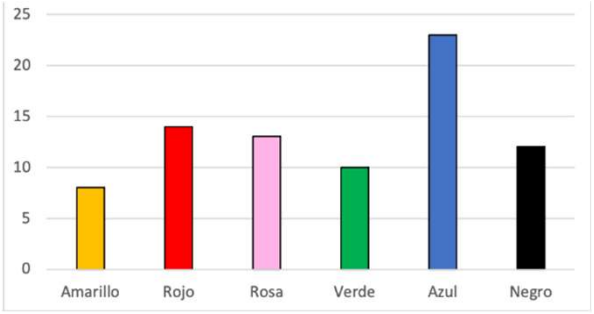
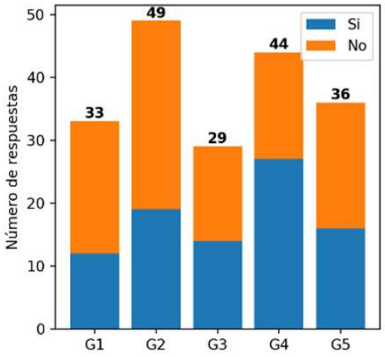


Gráfico de barras apiladas



Polígono de frecuencias *

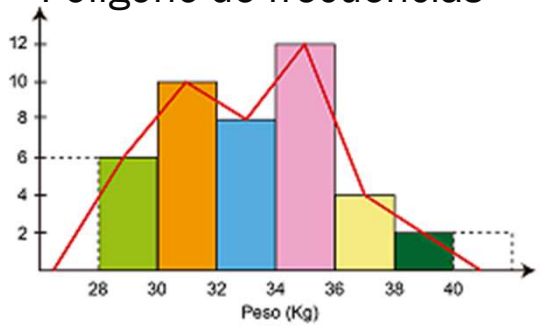
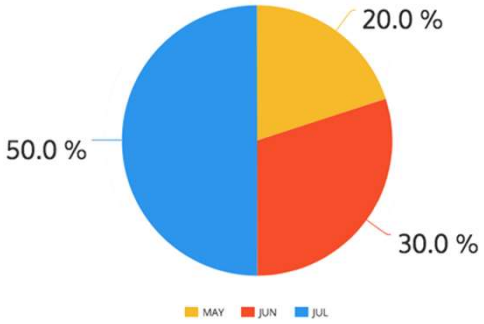
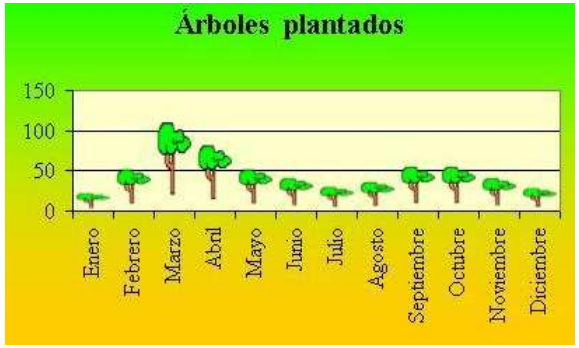


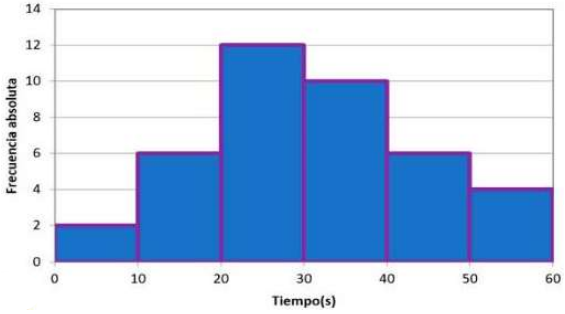
Gráfico de pastel



Pictograma



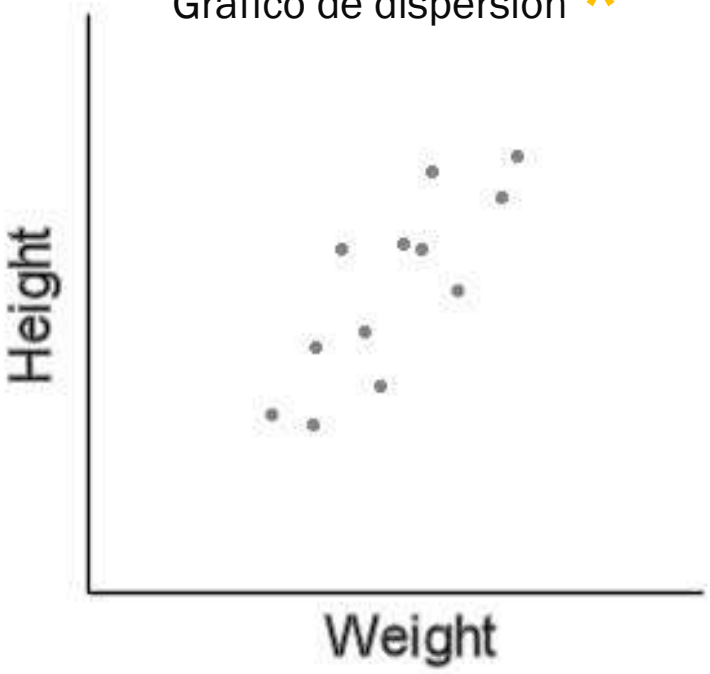
Histograma *



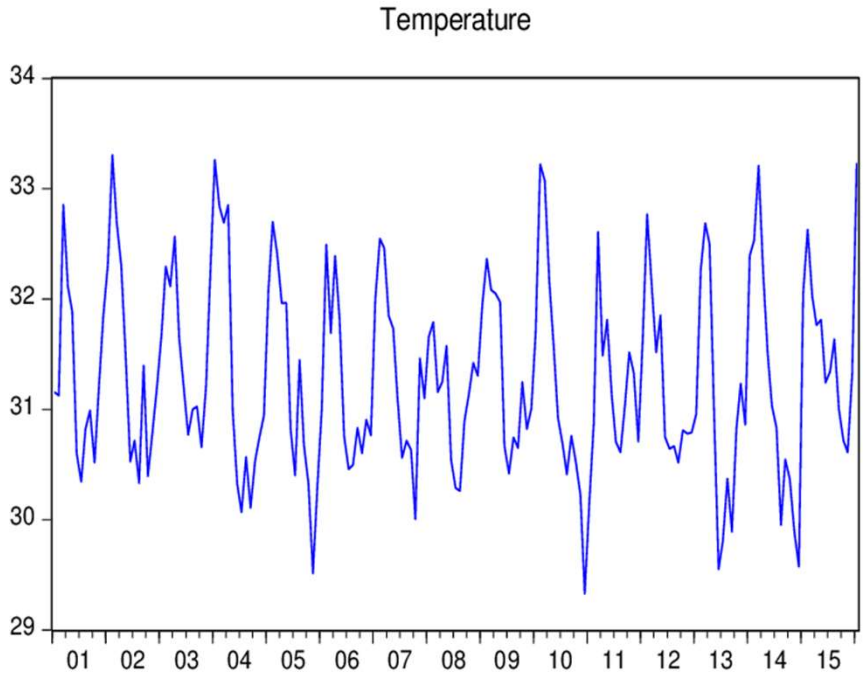
FUENTES:
https://matemovil.com/histogramas-ejemplos-y-ejercicios/
http://www.iered.org/archivos/Proyecto_coKREA/REAfinales2014/EstadisticaContexto_LuisaHernandez/pictograma.html
https://es.venngage.com/blog/grafica-de-pastel/
https://python-charts.com/es/parte-todo/grafico-barras-apiladas-matplotlib/
https://www.probabilidadyestadistica.net/diagrama-de-barras/

Gráficas básicas

Gráfico de dispersión *



Serie de tiempo *



FUENTE: http://astronomy.nmsu.edu/msussman/mbh-sig/04.html
https://www.researchgate.net/figure/Time-series-plot-of-the-temperature-data_fig1_312020809

Retos de la estadística en el Big Data

FUENTE: https://www.zdnet.com/article/who-really-owns-your-internet-of-things-data/

Problemas estadística clásica

1. Excesiva cantidad de información y datos
   - Códigos con métodos clásicos necesarios
   - Nuevos métodos para grandes cantidades de datos

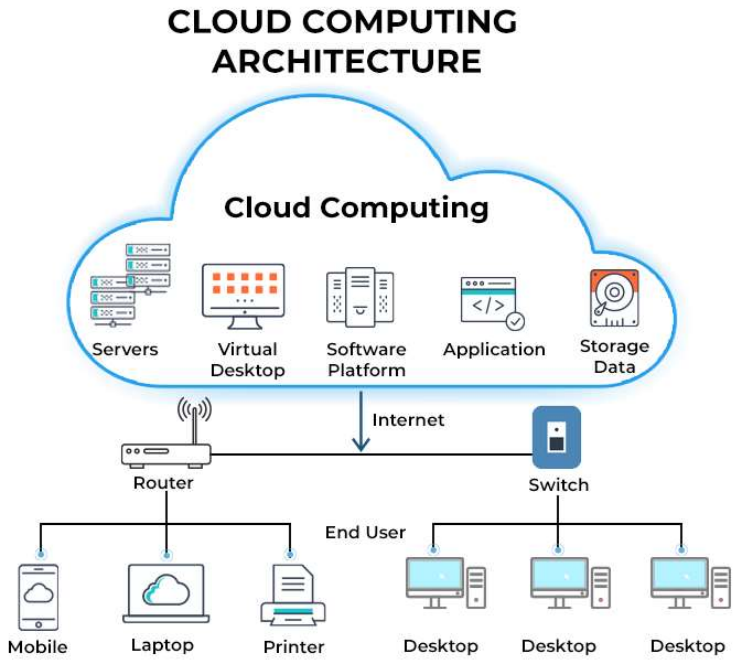
2. Datos outliers
   - Eliminar o suprimir

3. Complejidad
   - Procedencia
   - Transformación

Problemas estadística clásica

4. Infraestructura
   - Clústeres
   - Servicios de la nube
   - Métodos paralelizables

5. Políticas de privacidad
   - Autorización



6. Recolección de datos sin conocimiento del problema
   - Diseño del estudio -> Recolección de datos -> Aplicar encuestas

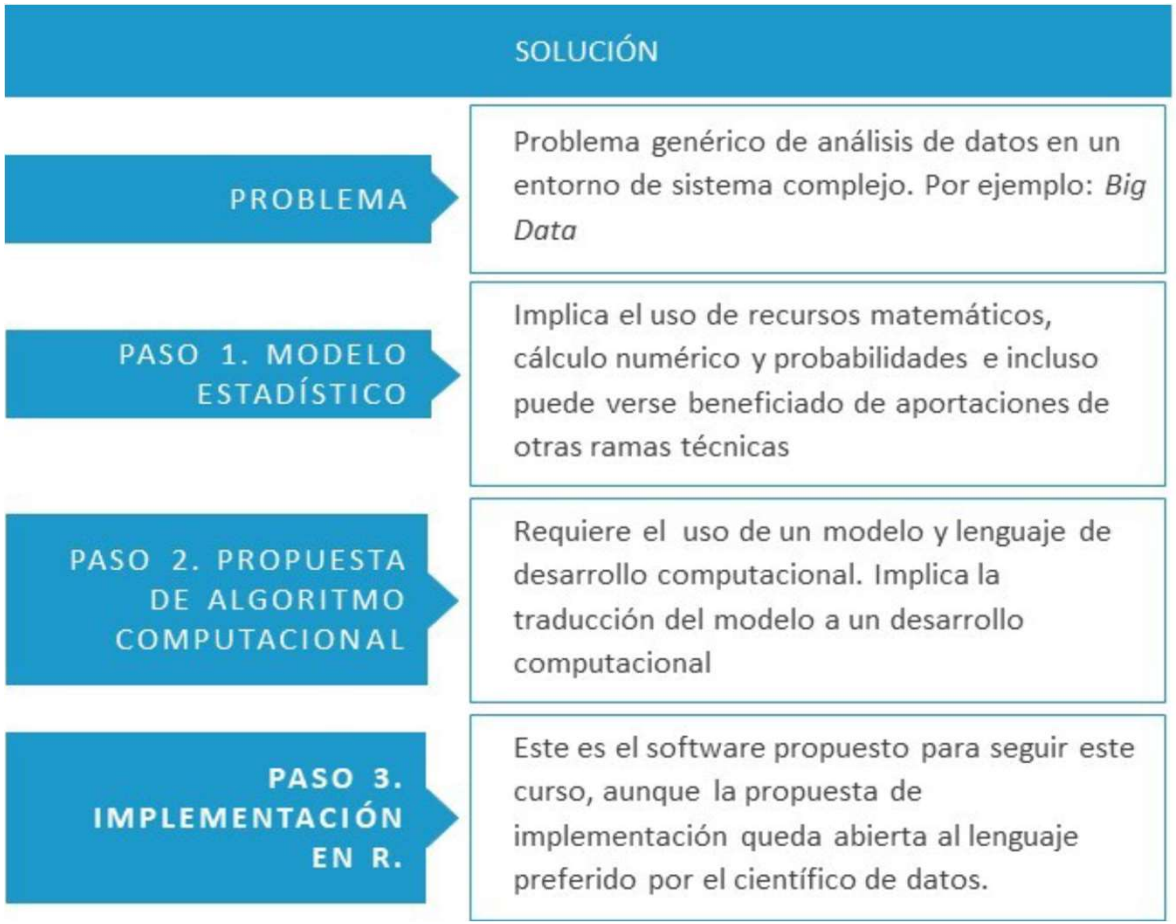
FUENTE:https://www.igmguru.com/blog/best-cloud-computing-services

Estadística computacional

Principios básicos

- La estadística computacional es un puente entre las ciencias clásicas y campos científicos
- Análisis estadístico es la ciencia de recopilar, discutir, visualizar y analizar datos
- Principios fundacionales de la estadística computacional subyacen en el conocimiento y control de tres áreas técnicas
    1. Programación
    2. Análisis numérico
    3. Estadística clásica – datos univariados

Principios básicos



Ámbitos de aplicación

- Estadística computacional en Biología, biestadística computacional
  - Emplean grandes volúmenes y diversos tipos de datos
  - Metodologías y herramientas eficientes de estadística computacional

- Big Data como base en el desarrollo de la medicina
  - Utilizar muchos datos para logar avances en estudios médicos

- Estadística computacional para facilitar trabajos de Ciberseguridad
  - Controlar el tráfico de datos y darle seguridad

Técnicas básicas de programación

- “Expresividad” del lenguaje
  - Seguir el código sin necesidad de ser especialista en desarrollo
- Seccionar el programa
  - Uso de funciones para estructurar el código
- “Modularidad”
  - Dividir el programa tanto como sea posible en “módulos”
  - Reutilizar el código

Software R

- Programa para el análisis, representación y visualización de datos. Es abierto (open source)
- Diferentes sistemas operativos
- Contiene implementaciones para el cálculo de “todas” las herramientas estadísticas
- Permite el acceso a otros programas de cálculo matemático
- Programación orientada a objetos

Software R

- Rutina de trabajo
    1. Identificar el problema
    2. Buscar qué hay ya implementado
    3. Acceder a ejemplos y mapearlos

- Características del lenguaje R
    - Flexible, solución numérica a cualquier problema estadístico
    - Reproducible, reutilizar el código
    - Código abierto, identificar errores e introducir sus mejoras
    - Interfaces controlables a través de línea de comandos

Software R

- Limitaciones
  - No tiene un método de trabajo intuitivo

- A nivel de hardware
  - Limitaciones de memoria
    - Acceso a memoria operativa
    - Guardar de manera periódica
  - Trabajo con extensas bases de datos
    - Paquetes auxiliares para dividir las BD

Medidas que resumen la información

Medias

Estas medidas indican la distribución de los datos a partir del medio o mitad del conjunto.

- Media aritmética

$$\bar{x} = \frac{\sum x_i}{n}$$

- Es la medida más usada
- Es sensible a los valores atípicos

- Media aritmética para datos con frecuencia

$$\bar{x} = \frac{\sum n_i \, x_i}{\sum n_i}$$

- Media ponderada

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

- Se emplea cuando unos datos tienen más importancia que otros

Medias

- Media armónica
  - Promedia variaciones con respecto a la misma variable

$$H = \frac{n}{\sum \frac{1}{x_i}}$$

- Media geométrica
  - Promedio que se usa para calcular el rendimiento medio de porcentajes o tasas de crecimiento
  - Se usa cuando los datos cambian de forma multiplicativa y no aditiva

- Media cuadrática
  - Raíz de la media cuadrática
  - Se emplea para dar importancia a valores pequeños de una variable

Mediana (*Me*)

- Es más robusta que la media cuando se tienen valores *outliers*
- Información más fiable en grandes cantidades de datos
- Observación central del conjunto de datos
  - Número impar – valor central
  - Número par – promedio de las observaciones centrales

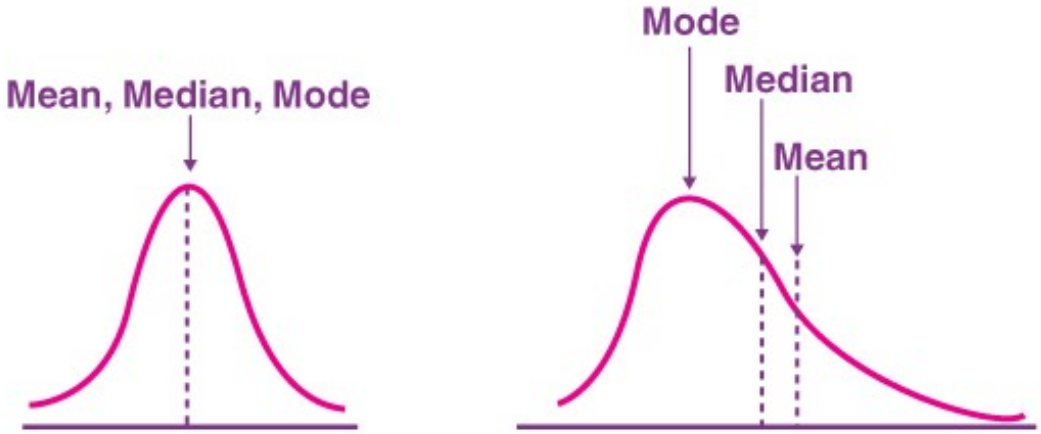
- Distribución de datos se encuentra en intervalos

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i$$

- Se puede emplear en variables cualitativas

Moda

- Es el valor más frecuente de los datos.
- Si se tienen 2 datos con la misma frecuencia es bimodal

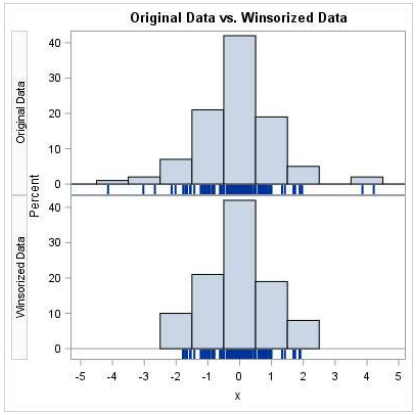


FUENTE: https://byjus.com/maths/mean-median-mode/

Medidas de tendencia central robustas

- Media recortada
  - Calcula la media aritmética a un subconjunto central del conjunto de datos
  - Media recortada al y%
  - Media recortada al 0% = media aritmética
  - Media recortada al 25% se conoce como “centrimedia”

~~1 5~~ 32 35 38 43 45 49 52 56 60 63 ~~80 98~~



- Media winsorizada
  - La diferencia con la media recortada consiste en que se sustituyen el menor y mayor valor (tras el proceso de eliminación) en las posiciones eliminadas.

FUENTE: https://blogs.sas.com/content/iml/2017/02/08/winsorization-good-bad-and-ugly.html

Medidas de dispersión

Estas medidas indican cuanto se desvían los datos (distribución)

- Rango

$$rango = x_{max} - x_{min}$$

- Varianza

$$Varianza = s^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

  - Promedio de las desviaciones de datos con respecto a su media
  - Varianza poblacional σ

- Desviación estándar (o típica)

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$

  - Depende de la escala de los datos, por lo que no se puede usar para comparar datos

Medidas de dispersión robustas

- Varianza winsorizada

$$s_W^2 = \frac{\sum(W_i - \bar{x}_\alpha^W)^2}{n}$$

- Cuasivarianza winsorizada

$$s_W^2 = \frac{\sum(W_i - \bar{x}_\alpha^W)^2}{n-1}$$

Medidas de posición y forma

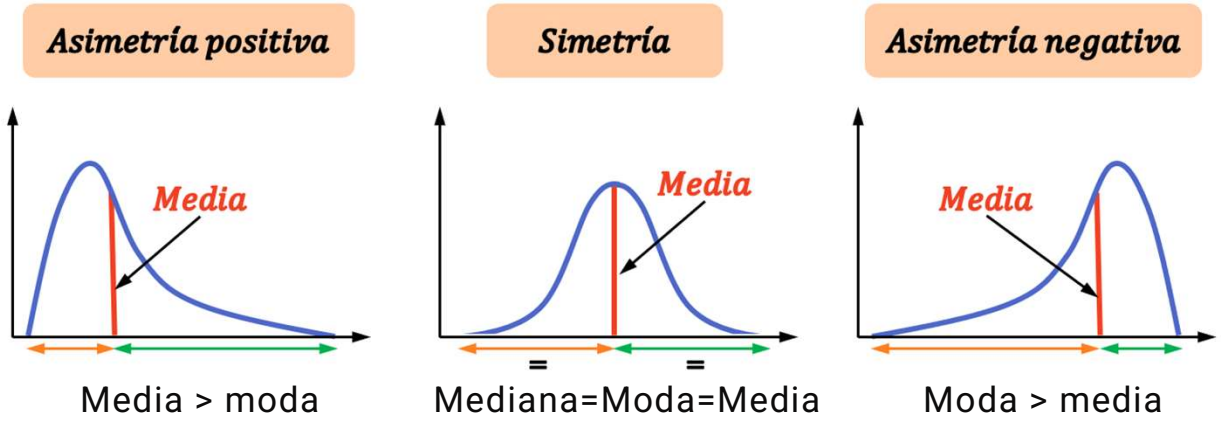
Sirven para saber si un valor esta alejado de su media

- Cuartil
  - Dividir en cuartos iguales los datos
    - Q1 – 25% de los datos son inferiores a él
    - Q2 – 50% de las observaciones, coincide con la mediana
    - Q3 – 75% de las observaciones a la izquierda
  - Percentiles – Dejan un % de datos a su izquierda

$$Q_k = L_{i-1} + \frac{\frac{kN}{4} - N_{i-1}}{n_i} a_i$$

Distribución de datos

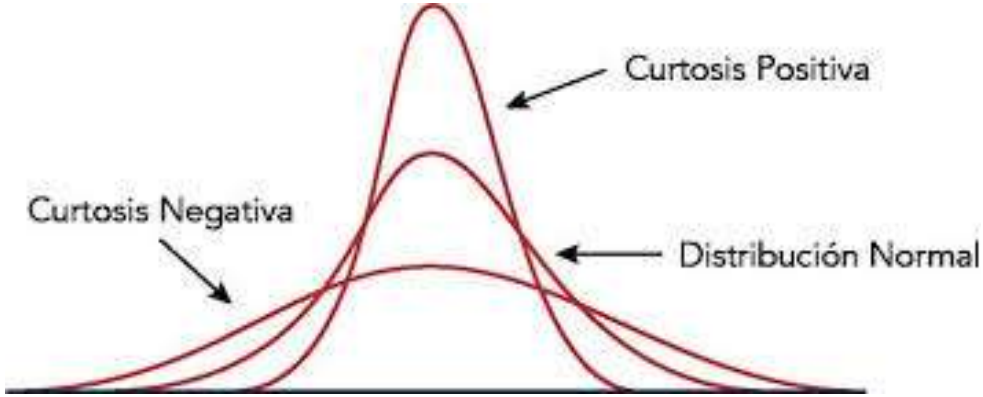
- Engloban la simetría y apuntamiento
- Distribución
  - Indican sesgos



FUENTE: https://www.probabilidadyestadistica.net/tipos-de-asimetria/

Distribución de datos

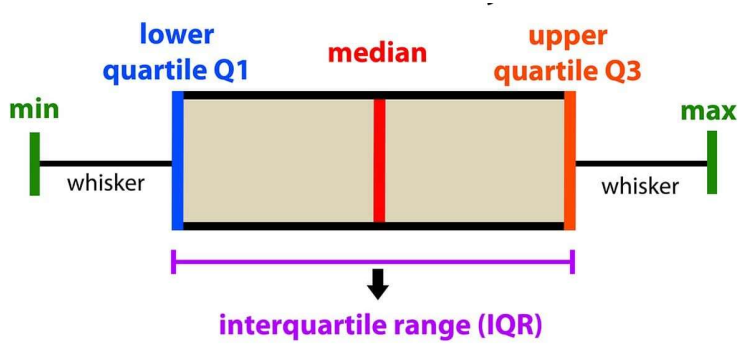
- Apuntalamiento
  - Grado de concentración de los datos alrededor de la media y zona central de la distribución



FUENTE: https://issuu.com/qxmedic/docs/ebook_-_estad_stica_v3_pi2023/s/26195109

Gráficos de caja

- Resumen la información de los datos
- Medidas que se representan
  - Mínimo
  - Q1 (límite de la caja)
  - Mediana (divide la caja en 2)
  - Q3 (límite de la caja)
  - Máximo
- Rango intercuartílico



FUENTE: https://www.simplypsychology.org/boxplots.html

Datos atípicos y análisis exploratorio de datos

- Valores atípicos o extremos (outliers)
  - Distan de la mayoría de los datos
  - Afectan la media
  - Afectan la dispersión media que mide la desviación típica
  - Altera la distribución de datos, histogramas
  - Se identifican en el análisis exploratorio de los datos

Regresión y correlación

Introducción

- Frecuencias marginales
  - Frecuencias de la distribución unidimensional de la variable que es la que resulta de no tener en cuenta la otra variable

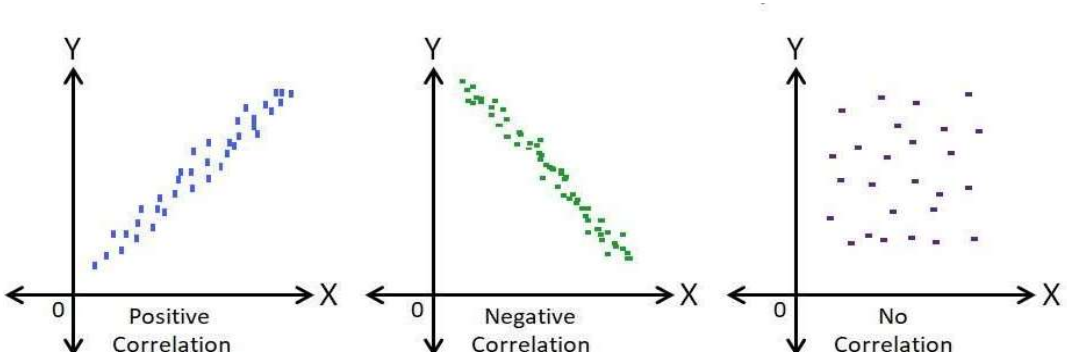
- Covarianza
  - Mide lo que covarían las dos variables, es decir, es la media aritmética de los productos de las desviaciones de cada variable con respecto a su media

$$S_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

  - Mide la fuerza de la relación entre las variables aunque a través del coeficiente de correlación

Introducción

- Herramienta gráfica



- Examinar un gráfico de dispersión
  - Patrones que se manifiestan
  - Forma, dirección y fuerza del patrón
  - Valores atípicos
  - Identificar grupos aislados

FUENTE: https://trailhead.salesforce.com/es/content/learn/modules/data-analysis-in-tableau-desktop/compare-measures-using-a-scatter-plot

Introducción

- Asociación entre variables
  - Valores de una de las variables es más propenso a los valores que toma la otra

- Estudio de la asociación entre variables
  - Se estudia para describirla
  - Predecir o explicar una variable a partir de otra

- Tipos de variables
  - Variable explicativa (predictora), variable independiente
  - Variable respuesta, variable dependiente

Correlación

- Coeficiente de correlación de Pearson
  - Medida de la fuerza de la relación lineal entre dos variables cuantitativas

$$r_{xy} = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{S_{xy}}{S_x S_y}$$

  - Propiedades
    - Está limitado entre -1 y 1
    - No importa el orden en que se calcula (x, y) o (y, x)
    - Cuando vale 0, indica que no existe asociación lineal
    - Cuando vale 1 o -1, la relación es máxima y se considera “perfecta”
  - No dice nada de otras relaciones no lineales que puedan existir
  - La correlación es sensible a los outliers

Regresión lineal

- Método de los mínimos cuadrados (MMC)
  - Procedimiento matemático que se emplea para calcular las ecuaciones de regresión y sus componentes
- La recta de regresión siempre pasa por el punto de las medias (x, y)
- La pendiente de la recta está presente en el coeficiente de correlación
- Bondad de ajuste r^2
  - Coeficiente de determinación
  - Mide el porcentaje de varianza de y explicada por x
- Valoración de la aproximación lineal
  - Revisar los valores de la correlación obtenidos

Gráfico de residuos

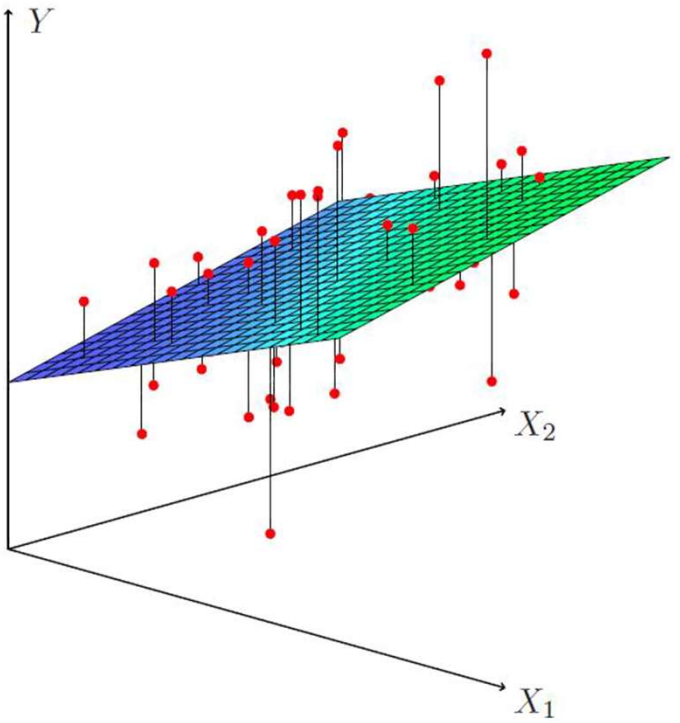
- Diagrama de dispersión con la variable explicativa en X y los residuos en Y
- La media de los residuos es 0
- No se observa algún patrón si la recta se ajusta bien

- Heterocedasticidad
  - Diferente varianza a lo largo de la variable explicativa
  - Errores crecen o decrecen según la magnitud de la variable explicativa

- Valores atípicos
  - Perjudican el ajuste correcto de una recta de regresión
  - Alteran su pendiente y punto de corte

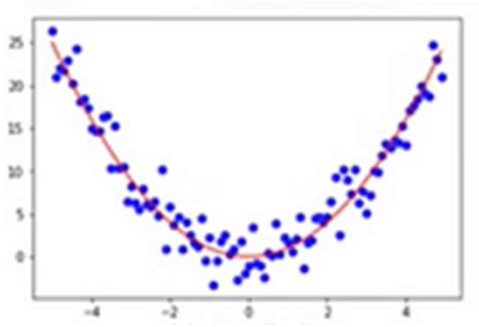
Regresión lineal multivariable

- Generación de un plano que contenga todos los puntos
- Relación que existe entre una variable con un conjunto de variables
- Error asociado a cada predicción
  - Residuo

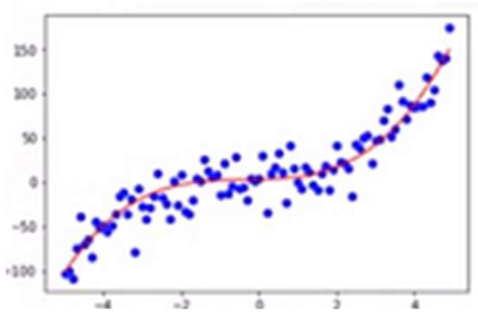


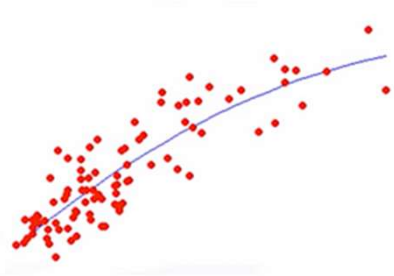
FUENTE:https://medium.com/analytics-vidhya/multiple-linear-regression-an-intuitive-approach-f874f7a6a7f9

Regresión no lineal



- Transformación en las variables
- Las asociaciones no lineales no se representan en el valor de correlación
- Grafico de residuos o de dispersión
  - Sospecha de una relación no lineal





- Relaciones logarítmicas o exponenciales

FUENTE: https://medium.com/@toprak.mhmt/non-linear-regression-4de80afca347

LTS (Least Trimmed Squares)

- Regresión de mínimos cuadrados recortados
- Variación del método de regresión por mínimos cuadrados que reduce la influencia de outliers.
- Trabaja con subconjuntos de puntos a los que se aplica el MMC
- Devuelve la versión que minimice los datos
- Pasos
  1. Selección de número de puntos
  2. Formación de subconjuntos
  3. Aplicación de MMC
  4. Selección del la opción con menor error

Probabilidad condicional y variables aleatorias

Probabilidad

Medida que se asocia a la ocurrencia de un suceso aleatorio, donde este es un evento de incertidumbre

$\Omega = \{\text{cara}, \text{cruz}\}$

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

Función que asocia la realización de un experimento aleatorio a un resultado.

Probabilidad

Concepción frecuentista

- La probabilidad de un suceso A es su frecuencia relativa

$$\lim_{n \to \alpha} \text{frec. relativa}\,(A) = \frac{n_A}{n} = \text{Prob}(A)$$

- Regla de Laplace

$$p(A) = \frac{\text{casos favorables}}{\text{casos posibles}}$$

Principios de la teoría de probabilidad

1. P(Ω) = 1, la probabilidad de un suceso que ocurre siempre (**suceso seguro**)

2. $0 \leq p(A) \leq 1$, la probabilidad será 0 cuando el suceso no ocurre nunca (**suceso imposible**)

3. Si A y B son dos sucesos disjuntos, es decir $A \cap B = \emptyset$, entonces

$$\mathrm{P}(\mathrm{A} \cup B) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$$

4. Si el espacio muestral está conformado por infinitos sucesos disjuntos A_i, entonces

$$\mathrm{P}(A_1 \cup A_2 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$$

Principios de la teoría de probabilidad

Otros resultados

1. $P(\phi) = 0$

2. Si $\bar{A}$ es el complemento de A, entonces $P(\bar{A}) = 1 - P(A)$

3. Si $A \subset B$ entonces $P(A) \leq P(B)$

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

Probabilidad condicional e independencia

- Probabilidad condicionada

$$P(A \text{ sabiendo que ha ocurrido } B) = P(A|B)$$

- Si ha ocurrido uno de los eventos y no modifica la probabilidad del otro se dice que son independientes

Variable aleatoria

- Dos clases
  - Discreta – número finito
  - Continua – Valores dentro de un intervalo

- Dos tipos de modelos de probabilidad
  - Discretos
  - Continuos

Modelos de probabilidad discretos

- Probabilidad mayor o igual a cero para cada valor de la variable X
- La suma de probabilidades es 1 y cada probabilidad esta entre 0 y 1
- A la función que asigna la probabilidad se le conoce como función de probabilidad

Modelos de probabilidad discretos

- Función de distribución indica la acumulación de probabilidad en un rango de valores discretos hasta uno dado

$$F(x_i) = P(X \leq x_i) = P(X = x_1) + P(X = x_2) + \cdots + P(X = x_i)$$

- Valor esperado o esperanza matemática
  - Resume la información de una variable aleatoria
  - Da una “idea” del resultado “promedio” de un conjunto de resultados posibles

$$E(X) = \sum x_i p_i$$

Distribución binomial

- Experimento con un número de éxitos
- Variable dicotómica
- Parámetros n (número de realizaciones) y p (probabilidad de éxito)

$$X \sim Bi(n, p)$$

- Probabilidad del fracaso 1-p
- Probabilidad de que una variable aleatoria binomial (n, p) tome un valor

$$P(X = k) = \binom{n}{k} p^k q^{n-k}$$

Modelos continuos

- Función de densidad *f(x)*
- Probabilidades nulas en puntos concretos
- Área que encierra la función de densidad entre ella y el eje X

- Distribución normal
  - Se caracteriza por su media y su desviación típica

$$X \sim N(\mu, \sigma)$$

  - La normal es simétrica respecto a su media (mediana y moda)
  - La mayor parte de la masa de probabilidad se acumula en torno a la media y cuanto más se aleja es más improbable

Modelos continuos

- Tipificación
  - Se emplean las reglas de media y varianza
  - Transformación de una N(μ, σ) en una N(0, 1)

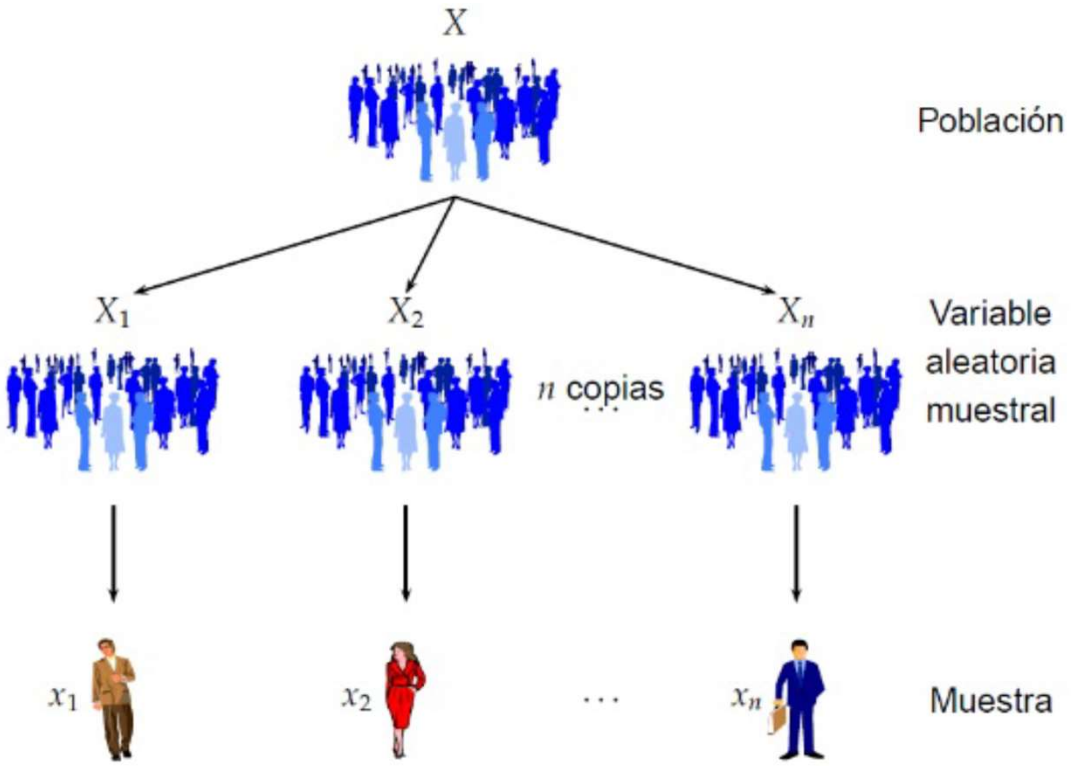
$$\frac{X-\mu}{\sigma} = z \leftrightarrow E\left(\frac{X-\mu}{\sigma}\right) = \frac{E(X)-\mu}{\sigma} = 0; V\left(\frac{X-\mu}{\sigma}\right) = \frac{V(X)}{\sigma^2} = 1$$

Distribución en el muestreo

Distribución en el muestreo del conteo y la proporción muestral

- Técnica inferencial
  - Estadístico, función de la muestra obtenida de una población
  - Distribución de muestreo
  - Requisitos
    1. Cada observación de la muestra sigue la misma distribución
    2. Todas las observaciones son independientes entre sí

Distribución en el muestreo del conteo y la proporción muestral



Teorema Central del Límite y distribución de la media muestral

- Pilar para la estadística inferencial

"El Teorema Central del Límite (TCL de aquí en adelante) afirma que cuando tenemos n variables independientes X_1, X_2, .., X_n (con n suficientemente grande) su suma $X_1+X_2+...+X_n$ es una variable aleatoria que se distribuye aproximadamente como una normal. Esta aproximación será mejor cuanto mayor sea n."

Aplicación del TCL en ámbitos Big Data

- El TCL asume que la información sigue una distribución normal

- Datos outliers
  - Conjunto de datos de tamaño considerable
  - Subconjunto en la población con propiedades diferentes a las de la media
  - Eliminarlos, perdida de información

- Estadística robusta
  - Análisis global y análisis exhaustivo

Estimulación puntual vs estimulación por intervalos

- Realizar estimaciones
  - Estimador puntual da un valor como estimador único
    - Se comete un error, no se logra acertar de manera absoluta con el parámetro.
    - Sesgo en la estimación del parámetro
  - Estimador insesgado
    - Produce estimaciones sin sesgo para un parámetro

Estimulación puntual vs estimulación por intervalos

- Estimación por intervalos a través de un intervalo de confianza
- Estimador de varianza mínima
- Existe un estimado de varianza único.

Intervalos de confianza

Introducción a los intervalos de confianza

- No es posible saber si una estimación puntual es buena o mala.

- Intervalo de confianza
  - Margen para situar el parámetro con cierta seguridad
  - Su amplitud depende de la confianza que se desea tener

- Nivel de confianza
  - Nivel de seguridad relacionado con la probabilidad de que el parámetro esta contenido en el IC

Nivel de significación

- La cantidad de probabilidad complementaria al nivel de confianza se conoce como nivel de significación
- Nivel de error que se está dispuesto a asumir

$$P[a \leq \theta \leq b] = 1 - \alpha$$

Nivel de significación

- La significancia es fundamental al momento de contrastar hipótesis.
- Se presenta cuando los estadísticos que se emplean toman valores a partir de los cuales se rechaza H_0.
- Se representa por α e indica la máxima probabilidad de cometer un error tipo I, esto es, rechazar H_0 siendo verdadera.
- Generalmente son 0.01, 0.05 o 0.1

Calculando el tamaño de la muestra

- Margen de error deseado

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

- Muestra piloto
  - Muestra que se recoge previamente
  - “Tantear” características de la población

Contraste de hipótesis

Distribución en el muestreo del conteo y la proporción muestral

Estadística inferencial

Consta de dos partes

1. Estimación de los parámetros de la población de estudio		2. Contraste de hipótesis

Procedimiento formal estadístico para decidir si una afirmación acerca de una población es verdadera o no a partir de los datos.

Hipótesis

- **Hipótesis nula** (H_0) es una afirmación que se va a probar si es aceptada o rechazada. Esta hipótesis representa lo conocido o lo que ya está establecido.

- **Hipótesis alternativa** (H_1) expresa lo novedoso o lo que contradice a lo establecido o un punto de vista conservador.

Prueba de una hipótesis estadística

Pruebas unilaterales (B y C)

La hipótesis alternativa indica un cambio en una dirección, > o <, con respecto a la hipótesis nula.

Pruebas bilaterales (A y D)

La hipótesis alternativa no indica una dirección para el cambio.

A	$H_0: \theta = \theta_0$ $H_1: \theta \neq \theta_0$
B	$H_0: \theta \geq \theta_0$ $H_1: \theta < \theta_0$
C	$H_0: \theta \leq \theta_0$ $H_1: \theta > \theta_0$
D	$H_0: \theta_1 \leq \theta \leq \theta_2$ $H_1: \theta < \theta_1 \; o \; \theta > \theta_2$

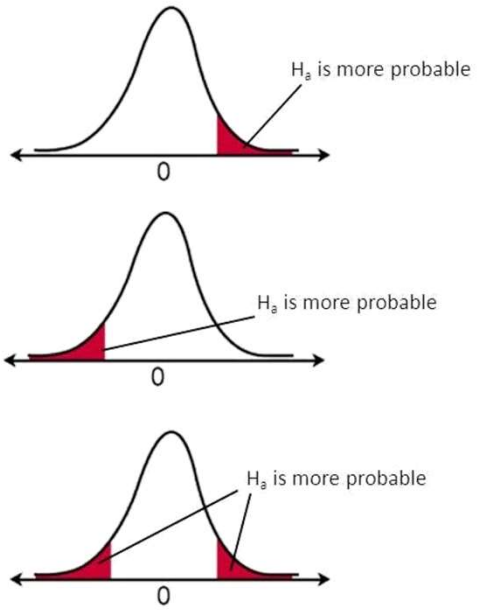
Contraste de hipótesis

- Significancia estadística
  - Se da cuando los estadísticos que se emplean para el contraste de hipótesis toman valores a partir de los cuales se rechaza H_0.

- Estadístico de contraste
  - Es cualquier número basado en la muestra de datos y que ayuda a tomar una decisión sobre H_0 y H_1.

- Distribución de probabilidad
  - Es la que resulta de suponer que H es verdadera

Región de aceptación y rechazo



Right-tail test

$H_a: \mu > \text{value}$

Left-tail test

$H_a: \mu < \text{value}$

Two-tail test

$H_a: \mu \neq \text{value}$

FUENTE: https://www.geo.fu-berlin.de/en/v/soga-py/Basics-of-statistics/Hypothesis-Tests/Introduction-to-Hypothesis-Testing/Critical-Value-and-the-p-Value-Approach/index.html

Dos tipos de error en la significancia estadística

Decisión	Estado real	
	H_0 es verdadera	H_0 es falsa
Rechazar H_0	**Error de tipo I** α	Decisión correcta $1-\beta$
No rechazar H_0	Decisión correcta $1-\alpha$	**Error de tipo II** β

Potencia del test o contraste

- Reducir al máximo las probabilidades de cometer errores de tipo I y II
- La probabilidad de cometer estos errores es inversamente proporcional.

$$P[\text{rechazar } H_0 | H_0 \text{ es verdadera}] = \alpha$$
$$P[\text{no rechazar } H_0 | H_0 \text{ es falsa}] = \beta$$

- La disminución se logra al aumentar el tamaño de la muestra

Dos tipos de error en la significancia estadística

p-value

Probabilidad de obtener un valor del estadístico de prueba que sea al menos tan extremo como el obtenido a partir de los datos muestrales

Regla del *p-value*

$$Si\ \ p\ valor\ \leq \alpha \rightarrow \mathbf{Rechazamos\ H_0}$$

$$Si\ \ p\ valor > \ \alpha \rightarrow \mathbf{Aceptamos\ H_0}$$

Pasos a seguir en un contraste de hipótesis

1. Sintetizar la hipótesis que se desea probar
2. Delimitar H_1 en base a H_0
3. Fijar un error α
4. Elegir el estadístico de prueba para contrastar las hipótesis
5. Recoger una muestra aleatoria y calcular el estadístico
    - Región de aceptación / rechazo
    - *p-value*

Contrastes paramétricos para dos muestras

- Dos muestras provienen de dos poblaciones
  - Datos apareados
  - Muestras apareadas

- Objetivo
  - Reducir las fuentes de variabilidad entre las muestras

- Muestras dependientes o relacionadas
  - Relación

Regresión

Modelo de regresión simple

$$E(y/x) = \alpha + \beta x$$

donde:

α	Constante del modelo (β_0)
β	Coeficiente de regresión o pendiente (β_1)

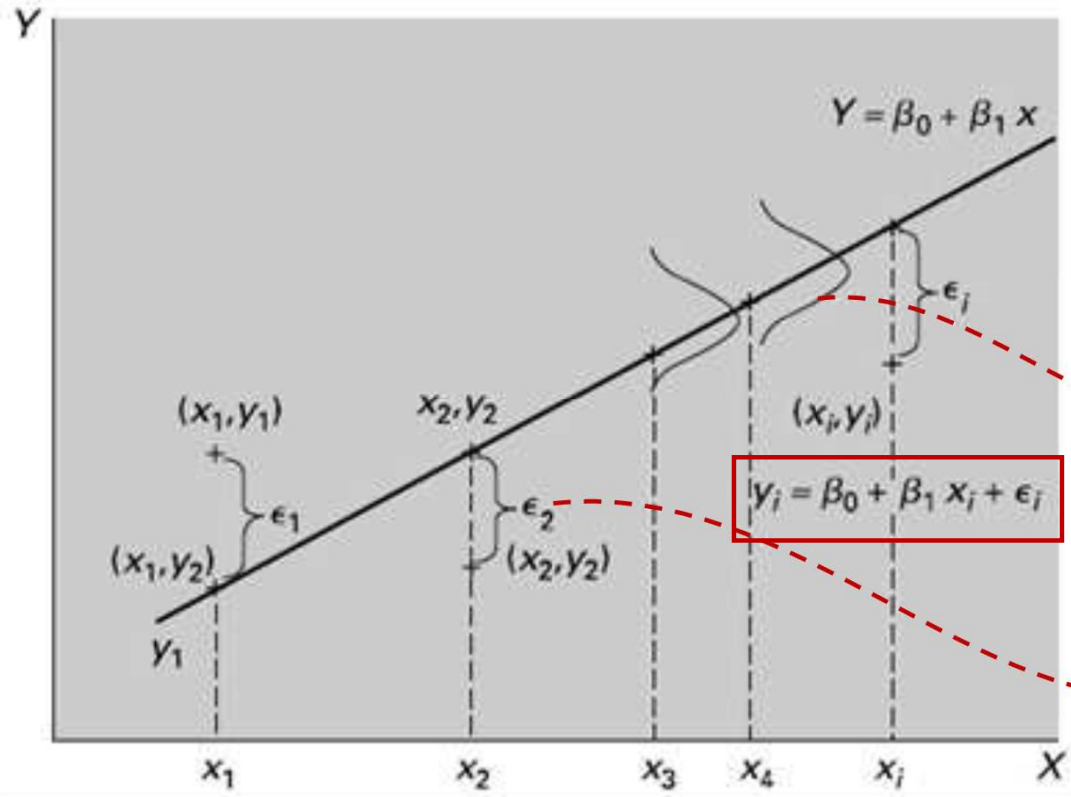
$$y \sim N(\alpha + \beta x;\ \sigma)$$

donde:

y	Se desviaría σ

Errores de la recta de regresión lineal

$$e_i = y_i - \hat{y}_i$$



Curvas normales

Modelo

Desviaciones a la recta

Modelo de regresión simple

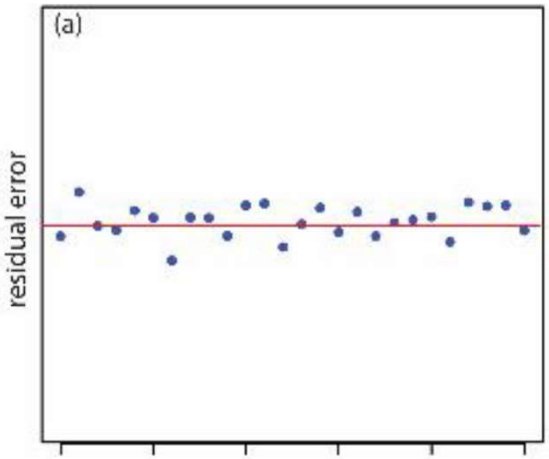
Requisitos

1. Observaciones **aleatorias**
   - No deben haber patrones

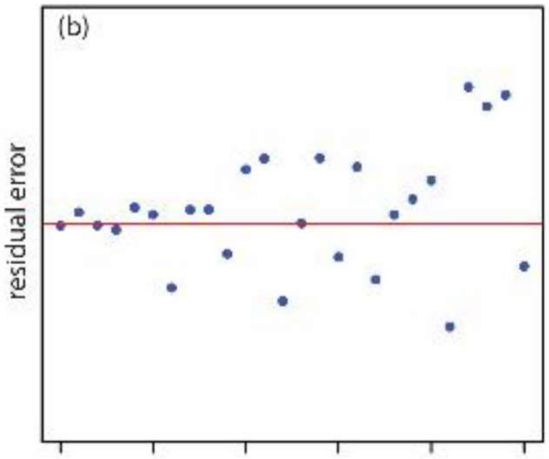
2. Relación **lineal**

3. **Varianza homogénea** a lo largo de x

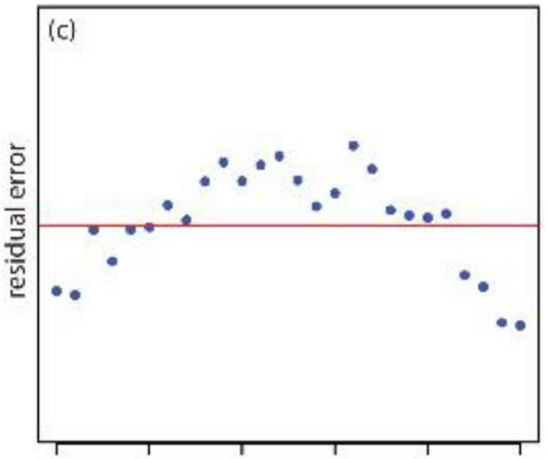
Faltas al modelo



Caso deseable



Varianza no constante (heterocedasticidad)



Falta de linealidad

Contrastando la regresión

Regresión es significativa

- Contrastar sobre la pendiente

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

- Una inferencia sobre la pendiente de la recta

$$b \sim N\left(\beta; \frac{\sigma}{\sqrt{n} * s_x}\right)$$

Contrastando la regresión

- La desviación total del modelo

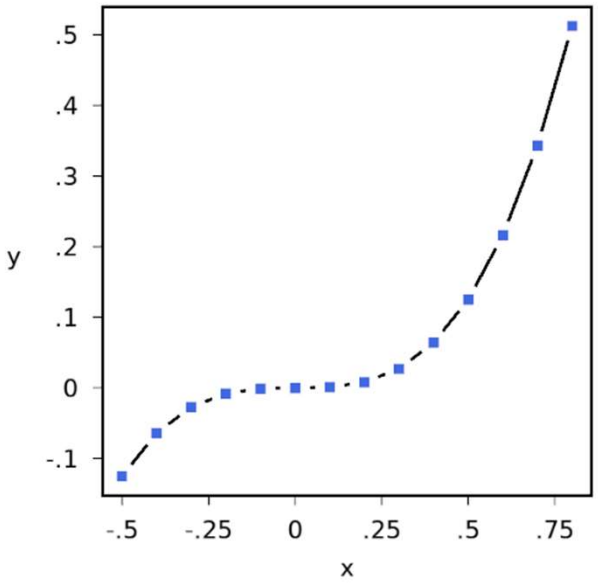
$$SCT = \sum (y_i - \bar{y})^2$$

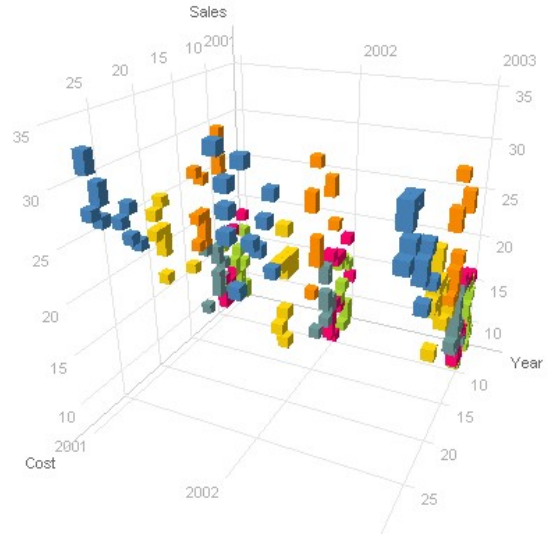
- Modelo debe contener parte de estas desviaciones
- Suma de cuadrados del modelo (SCM)

$$SCM = \sum (\hat{y}_i - \bar{y})^2$$

Análisis de componentes principales

Motivación





FUENTE: https://support.minitab.com/en-us/minitab/help-and-how-to/graphs/3d-scatterplot/before-you-start/example/
https://www.codeproject.com/Articles/5338541/How-to-Create-High-Performance-and-Publication-Qua

Motivación

- Entornos Big Data
  - Datos con alta dimensionalidad
    - Representación coherente de los datos
    - Transformación
  - Número total de datos
    - Número de individuos por la dimensión o número de características de cada individuo
    - Transformar datos para reducir la dimensión del conjunto de datos sacrificando la menor cantidad de precisión e información
  - Técnicas de análisis de componentes principales

Definición

"*PCA consiste en encontrar transformaciones ortogonales de las variables originales para conseguir un nuevo conjunto de variables incorreladas, denominadas Componentes Principales, que se obtienen en orden decreciente de importancia*"

- Crea nuevas variables de dimensión como combinación de las variables ya existentes

Definición

- Beneficios
  - Variable formada a partir de las variables del modelo, valor resumen
  - Variable que mejor recoge la variabilidad del modelo
    - Mejor describe la variabilidad del modelo de datos
    - Minimiza la perdida de información
  - Número reducido de dimensiones que permiten una correcta representación de la información

- Vector de componentes
  - Vector de números que se multiplican por los valores de cada individuo

Pasos

1. Calculo de componentes principales
   - Matrices de varianzas y covarianzas

2. Selección del número de componentes a incluir en el modelo
   - Mayor precisión a más componentes

3. Análisis de resultados y graficas del modelo

Aplicaciones

- Reducción del volumen de datos
- Reducción de ruido en imágenes
- Detección de cambios en los datos
  - Datos de una población en momentos diferentes
  - Variables inamovibles – variables que participan en CP

muchas gracias.