

Estadística

Tipos de datos:

▷ Categorical data: Ej: morcas, true o false

▷ Numerical data: dato numerico contable, Ej: n° hijos

DISCRETE

Numerical data:

CONTINUOUS

La diferencia es la precisión, mientras que con el numero de hijos puedes tener 1, ..., n con por ejemplo el peso puedes tener precisión infinita, una gota de sudor bajaria tu peso

Ej discrete:
n° objetos
n° personas

Ej continuous
Tiempo
Distancia
Area
Altura

Measurement Levels

Qualitative

Nominal

Categorías.

No son números ni pueden ser ordenados.
Ej. marcas, estiramiento del pelo.

Ordinal

Grupos y categorías con un orden estricto.

Ej. Puntuación con

- disagree

- good

- perfect

Interval

Do not have true 0

Grados creados por la
humana por comodidad

como por ejemplo la
temperatura

• Temperature
in °C or °F

Ratio

Has a true 0

• Comparación de
valores

• d^2 objects

• Distance

• Time

• Temperature in
grados Kelvin
(absolute)

TIPOS DE GRAFICOS

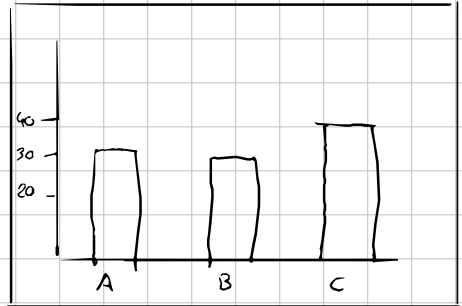
Representation of Categorical data

□ Frequency distribution table,

Frequency	
A	30%
B	30%
C	40%
total	100%

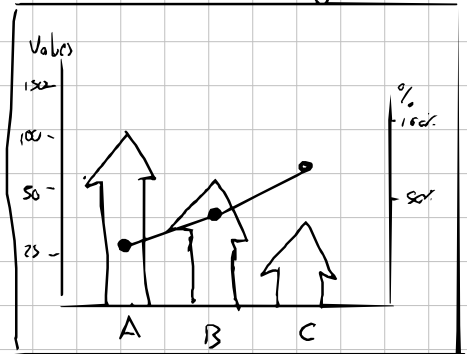
Puede mostrar % o $\frac{1}{100}$

Bar Chart



✗ Representación gráfica de una tabla de distribución

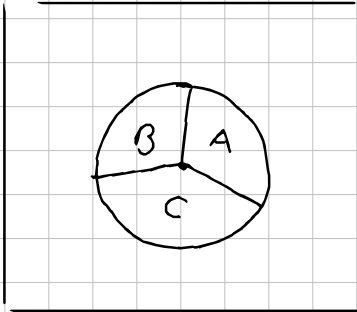
Pareto Diagram



Las categorías son expresadas en orden de frecuencia como barras y luego se pone una línea que va sumando los % a frecuencia de los dño mostrados.

• La barra izquierda es para el control de las barras y la de la derecha el % para la línea de frecuencia.

PIE CHART



Ver % del total (¿multiplicar)

Representación gráfica numerical Data

► Frequency distribution table

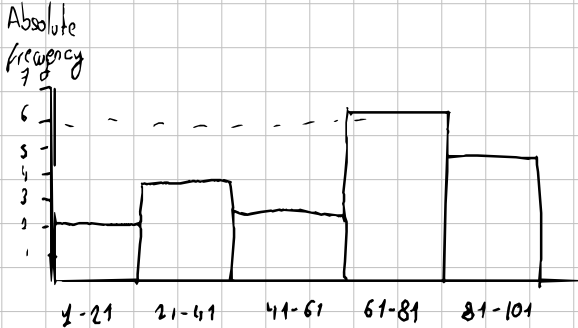
	freq	relative freq
1-21	2	0.1
21-41	4	0.2
41-61	3	0.15
61-81	6	0.3
81-100	5	0.25

Agrupamos los números en intervalos
(la cantidad y tamaño dependen de los datos)
para observar luego la frecuencia de los
mismos.

$$\text{Relative Freq} = \frac{\text{Freq}}{\text{total Freq}}$$

USADA PRINCIPALMENTE PARA RESUMIR Y FACILITAR
EL TRABAJO CON LOS DATOS EN LAS REPRESENTACIONES
GRÁFICAS.

► Histogram

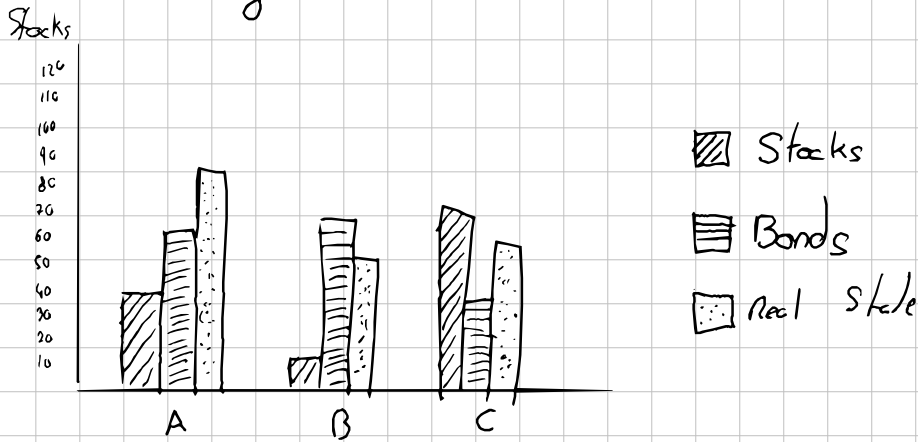


The bar touching is used to show continuity. Si los intervalos
son de diferentes tamaños se puede trabajar con anchuras de los
barras.

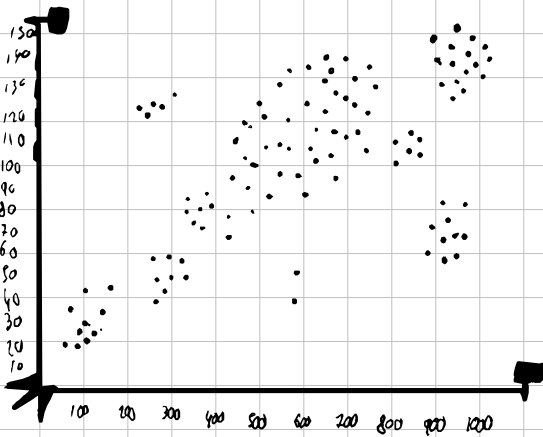
Relations of Two Variable,

Cross Tables

Side by Side bar chart



Scatter Plots



▷ facilita identificar clusters

▷ facilita la identificación de correlación entre 2 valores



Mean, Median and Mode

Mean / Average $\left[\begin{array}{l} \text{Population } \mu \\ \text{Sample } \bar{x} \end{array} \right]$

Median / Mediana $\left[\begin{array}{l} \text{Numero intermedio en el que el 50\%} \\ \text{de los datos son m\u00e1s altos y el} \\ \text{50\% m\u00e1s bajo.} \end{array} \right]$

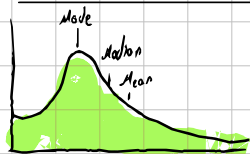
Mode / Moda $\left[\begin{array}{l} \text{Numero m\u00e1s repetido} \end{array} \right]$

Measures of Asmetry

Skewness
Asimetría
Sesgo

Indica cuando los datos, están concentrados en un lado o en el otro de el mean.
Para fijarnos debemos ver hacia donde está la cola

Positive Skewness

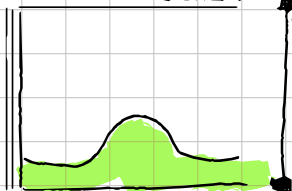


$$\text{Mean} > \text{median}$$

Los outliers estarán a la derecha

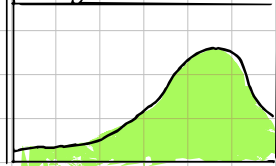
Mode: value with the higher representation

Zero Skewness



$$\text{Mean} = \text{median} = \text{mode}$$

Negative Skew



$$\text{Mean} < \text{Median}$$

Variance - Standar Deviation - (coefficient of Variation

Variance. Mide la dispersión de un conjunto de datos alrededor de su mean

$$\text{Population variance} = \sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

$$\text{Sample variance} = s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Basicamente suma de todos los valores menos el mean dividido por el total de valores.

▷ Standard Deviation:

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

Most common measure of variability

Coefficient of Variation

Standar deviation
mean

$$C_v = \frac{\sigma}{\mu}$$

Esto se usa para comparar $\hat{C}_v = \frac{s}{\bar{x}}$

datasets. No tiene una unidad de medida por ser

Por ejemplo si comparamos un dataset en dolares y otro en pesos el "s" de ambos sera extremadamente diferente, pero usando \hat{C}_v nos dan datos con los que poder comparar

Covariance

▷ Population Covariance

$$\sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x) \cdot (y_i - \mu_y)}{N}$$

▷ Sample Covariance

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

Si las 2 variables se mueven en la misma dirección la covarianza será positiva. Si se mueven en opuesta negativa y si son independientes será 0

Correlation Coefficient

$$\frac{\text{Covariance}(x, y)}{\text{Standard deviation}(x) \cdot \text{Standard deviation}(y)}$$

$$\frac{s_{xy}}{s_x s_y}$$

$$\frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Esto nos permite tener un dato numérico interpretable con el que trabajar.

$$-1 \leq \text{Correlation Coefficient} \leq 1$$

Siendo 1 el máximo posible de relación entre ellos a esto lo llamamos "PERFECT POSITIVE CORRELATION"

Basicamente que una variable es totalmente explicada por otra

Correlación de 0 significa independencia total

Correlación de -1 significa que son opuestos aunque es extra-ordinario ver un -1

* Ej. compañía de helados (verano) y de abrigo (invierno)

$$- \text{Corr}(x, y) = \text{Corr}(y, x)$$

Correlación not imply Causation

Central Limit Theorem

Sampling distribution of any distribution will approximate to a normal distribution.

$$N\left(\mu, \frac{\sigma^2}{n}\right)$$

Standard Error

Standard deviation of the sampling distribution

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

It is the variability of the different samples we extracted

It is used to verify how well you approximated the true mean

A más sample (n) menor es, indicando que

no aproximamos más

Estimators and Estimates

Estimator: Aproximación dependiente únicamente en la información del sample

Estimate: valor específico

Point Estimate - single number that is the middle of the confidence interval

Confidence Interval:

Ej. \bar{x} (mean sample) es un point estimate de μ (population mean)
same for s^2 and σ^2

Bias: variabilidad. Ej. unbiased estimator $\bar{x} = \mu$

Es algo así como un modificador que añadimos a la estimación que favorece a un resultado concreto

Efficiency: A más eficiente menos bias y variancia tendría el estimator

Statistic	Estimator
Broad Term	Type of Statistic

CONFIDENCE INTERVAL

Representación más precisa y realista de la realidad

* Nunca se puede tener una confianza del 100% excepto si se hace con el population

Confidence Level $(1 - \alpha)$

$0 \leq \alpha \leq 1$ Ej. if confidence = 95% (0.95) $\alpha = 0.05$

FORMULA CALCULO INTERVALOS

$$\left[\begin{array}{l} \text{Point Estimate} - \text{Reliability Factor} \cdot \text{Standard Error} , \text{Point Estimate} + \text{Reliability Factor} \cdot \text{Standard Error} \end{array} \right]$$



$$\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} , \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

\bar{X} = sample mean

σ = Standard deviation = $\sqrt{\text{variance}} = \sqrt{\sigma^2}$

n = number of values in the sample

z = Standard normal distribution = $\frac{Y - \mu}{\sigma}$

α = 1 - confidence level = $1 - (1 - \alpha)$ | Generalmente se usan valores de confianza de 90%, 95% y 99 por lo que $\alpha = 0.1, 0.05$ y 0.01

Para identificar luego el valor de z se buscan en la z table el valor más aproximado y sumar el valor de su fila y columna.

Cálculo umbrales de confianza con T. Con variance unknown

1. Cálculo standard error: $\frac{s}{\sqrt{n}}$

2. Fórmula cálculo intervalo sin conocimiento de variancia

$$= \bar{x} \pm t_{n-1, \alpha/2} \cdot \frac{s}{\sqrt{n}}$$

2.1 Si suponemos el población variancia sería

$$= \bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

n : sample size

α : confidence level

Margen de Error

variance known

$$Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

variance unknown

$$t_{n-1, \alpha/2} \cdot \frac{s}{\sqrt{n}}$$

A estas expresiones tambien las llamamos margen de error

o ME

Por lo que definimos el umbral de confianza como:

$$\bar{x} \pm ME$$

Confidence Intervals with more than 1 population

TYPES OF SAMPLES - DEPENDANT / INDEPENDANT

► Dependant.

used for example when we are researching the same subject over time

Ej. • weight loss and blood sample

- habits on each member of a couple

- cause and effect. Note de corte y aceptación a universidad

► Independent Samples.

- Pop variance known

- Pop variance unknown assumed different

- Pop variance known assumed same

Dependant Sample

Un ejemplo de uso es en la medicina comparando los datos de los pacientes antes y después de tomar una medicación

* Cuando trabajamos con estadísticas en Biología, tendemos a asumir distribuciones normales por lo extraordinariamente comunes que son.

Ej: calcio magnesio en sangre antes y después medicina

Paciente	Before	After	Dif
1	2	1.7	-0.3
2	1.4	1.7	0.3
3	1.3	1.8	0.5
4	1.1	1.3	0.2
5	1.8	1.7	-0.1
6	1.6	1.5	-0.1
7	1.5	1.6	0.1
8	0.7	1.7	1
9	0.9	1.7	0.8
10	1.5	2.4	0.9

1. Calculamos datos para la diferencia
usaremos un α de 95% de confianza

$$\alpha = 0.05$$

$$n = 10$$

$$\text{mean} = \bar{d} = 0.33$$

2. Calculo standard deviation

np. std (vals, df=n-1) → calcula la std de la muestra
 $S = 0.455$

3. Una vez tenemos los datos de cálculo calculamos el umbral de confianza

sample diff →

es estadística T

* $\bar{d} \pm t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}$

$$\bar{d} \pm t_{n-1, \alpha/2} \cdot \frac{S}{\sqrt{n}} = 0.33 \pm t_{9, 0.025} \cdot \frac{0.455}{\sqrt{10}}$$

$$0.33 \pm 2.262 \cdot 0.14388$$

$$0.33 \pm 0.32545656$$

$$[0.004, 0.655]$$

4. Es rango positivo por lo que interpretariamos que la medicina si ha ayudado a aumentar el nivel de magnesio

Independent Samples known VARIANCE

1°. Type: Known population variance

Department / Grades	Eng	Manag
size	100	70
sample mean	58	65
pop std	10	5

• Son datos totalmente independientes tanto en relación, tamaño y datos.

• Queremos encontrar con una confianza del 95% la diferencia de medias

diff $\bar{x} - \bar{y}$

size ? (population)
mean -7
pop std

1. Calculamos la variancia de la diferencia

$$\sigma_{\bar{x} - \bar{y}}^2 = \frac{\sigma_e^2}{n_e} + \frac{\sigma_m^2}{n_m} = \frac{10^2}{100} + \frac{5^2}{70} = 1.36$$

2. Calculamos el umbral

$$\text{means} \rightarrow (\bar{x} - \bar{y}) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} = (-9.28, -4.72)$$

95% confidence

DISCLAIMER

- Se han comparados los datos de Ingeniería con los de management. Si se comprobasen los de management con los de ingeniería el resultado cambiaría a (4.72, 9.28) es decir EL ORDEN DE COMPARACIÓN IMPORTA
- (murmurando el resultado de una comparación o otra es simétrico respecto a 0)

Independent Samples - Same unknown variance

Ejemplo: comparación de precios de monedas en LA y NY

NY	LA
3.86	3.02
3.76	3.22
3.87	3.24
3.99	3.02
4.02	3.08
4.25	3.15
4.13	3.82
3.98	3.44
3.99	
3.62	

	NY	LA
Mean	3.94	3.25
Std dev	0.18	0.27
Sample Size	10	8

1º Paso, estimar el population variance

Para esto calculamos el unbiased estimator llamado "Pooled sample variance"

$$S_P^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2} = \frac{(10-1) \cdot 0.18^2 + (8-1) \cdot 0.27^2}{10+8-2}$$

$$S_P^2 = 0.05$$

$$S = 0.22$$

$$= 0.05$$

2º Paso, uso de Student's T para el resto del cálculo con normalidad usando el pooled sample variance como nuestro population variance usando la formula modificada

$$(\bar{x} - \bar{y}) \pm t_{n_x + n_y - 2, \alpha/2} \sqrt{\frac{S_P^2}{n_x} + \frac{S_P^2}{n_y}}$$

$$(3.94 - 3.25) \pm 2.12 \sqrt{\frac{0.05}{10} + \frac{0.05}{8}}$$

$$CI_{95\%} = (0.47, 0.92)$$

INDEPENDENT SAMPLE - unknown assumed different variance

USO DE LA FORMULA

$$(\bar{x} - \bar{y}) \pm t_{v, \alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

v = calculo de los grados de libertad

NO TRATAR DE MEMORIZAR ESTO

$$v = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\left(\frac{s_x^2}{n_x}\right)^2 / (n_x - 1) + \left(\frac{s_y^2}{n_y}\right)^2 / (n_y - 1)}$$