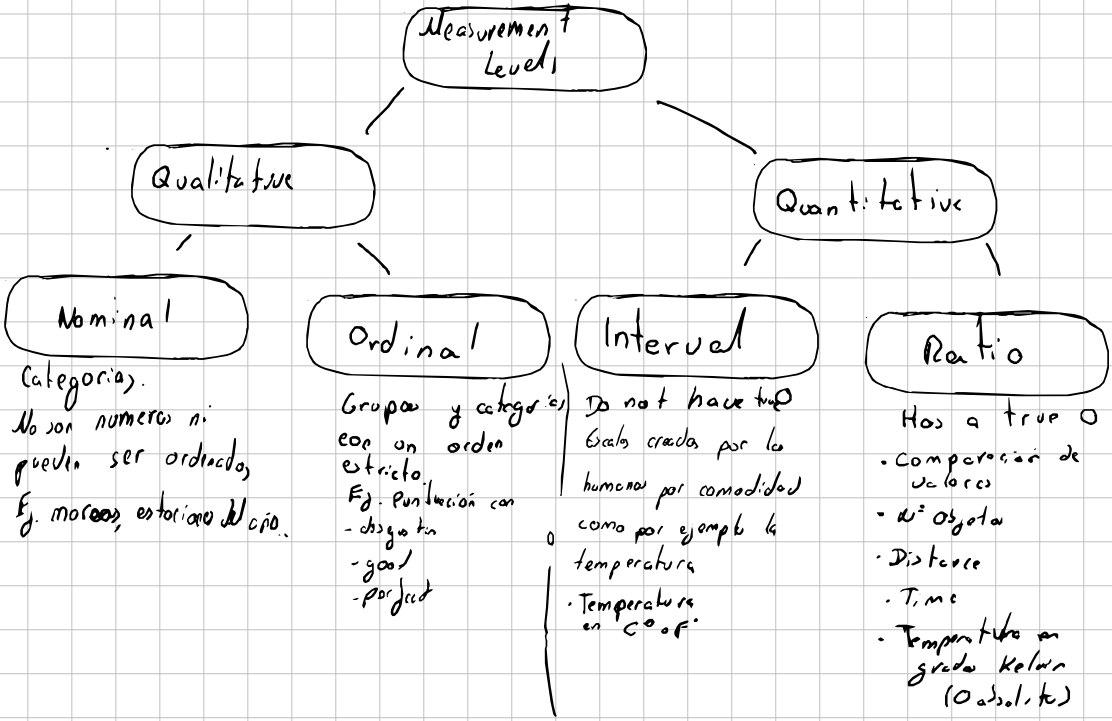
Estadística

Tipos de datos:

- Categorical data: Ej: marcas, true o false
- Numerical data: datos numéricos contables. Ej nº hijos
  - DISCRETE
  - Numerical data: la diferencia es la precisión, mientras que con el número de hijos puede tener 1, ..., n con por ejemplo el peso puede tener precisión infinita, una gota de sudor bajaría tu peso
  - CONTINUOUS

Ej discreto:
nº objetos
nº personas

Ej continuous
Tiempo
Distancia
Area
Altura


Measurement Level
Qualitative
Quantitative
Nominal
Ordinal
Interval
Ratio
Categorías.
No son numeros ni puedn ser ordenados
Ej. moreos, estaciones del año.
Grupos y categorías con un orden estricto.
Ej. Puntuación con
- disgusto
- good
- perfect
Do not have true 0
Escala creada por los humanos por comodidad como por ejemplo la temperatura
• Temperatura en C° o F°
Has a true 0
• Comparación de valores
• Nº objetos
• Distance
• Time
• Temperatura en grados Kelvin (0 absoluto)

TIPOS DE GRAFICOS

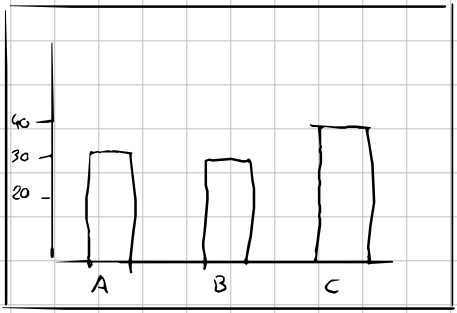
Representation of Categorical data

- Frequency distribution table

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

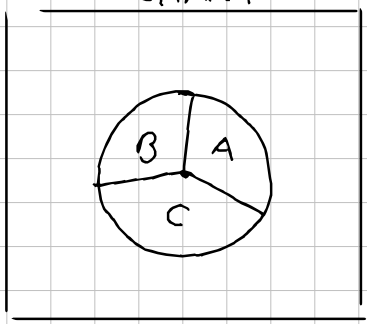
Puede mostrar % o fgos

Bar Chart



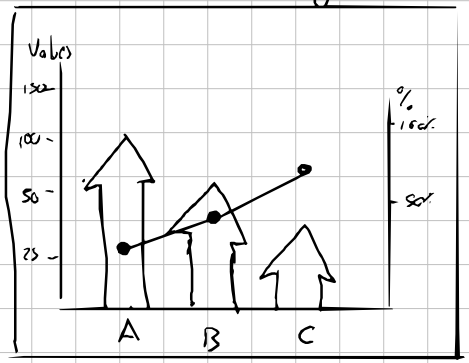
↖ Representación gráfica de una tabla de distribución

PIE CHART



Ver % del total (% market share)

Pareto Diagram



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

- La barra izquierda es para el conteo de los datos y la de la derecha el % para la línea de frecuencias

Representación grafica numerical Data

Frequency distribution table

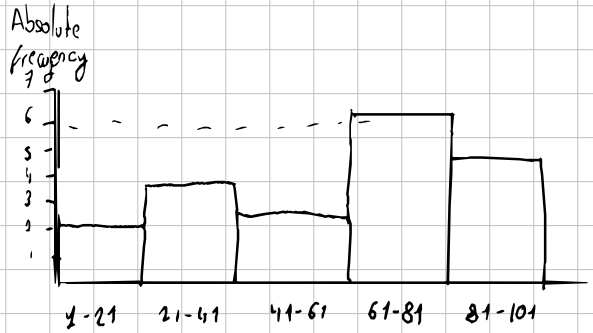
	Freq	Relative Freq
1 - 21	2	0.1
21 - 41	4	0.2
41 - 61	3	0.15
61 - 81	6	0.3
81 - 100	5	0.25

Agrupamos los numeros en intervalos (la cantidad y tamaño dependera 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 GRAFICAS.

Histogram

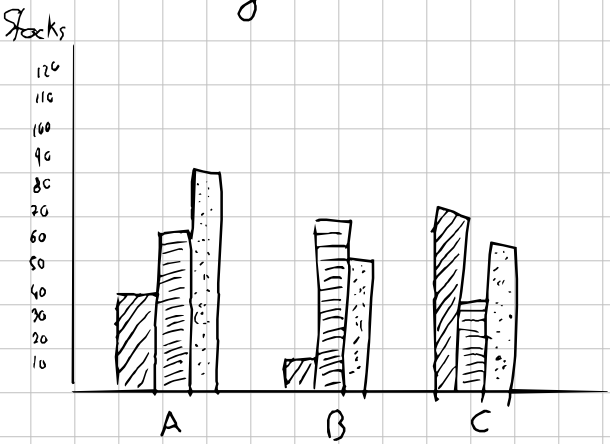


The bar touchin is used to show continuity. Si los intervalos son de diferentes tamaños se puede trabajar con anchuras de las barras.

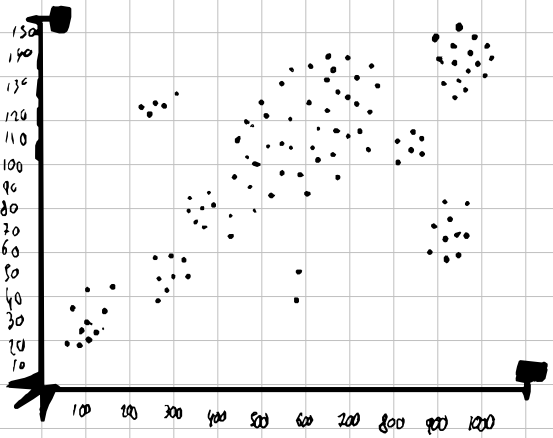
Relations of Two Variables

Cross Tables

Side by Side bar chart



Scatter Plots



- Facilita identificar clusters
- Facilita la identificación de conexión entre 2 valores

Mean, Median and Mode

Mean / Average
- Population μ
- Sample $\bar{x}$

Median / Mediana
- Numero intermedio en el que el 50% de los datos son más altos y el 50% más bajo.

Mode / Moda
- Numero más repetido

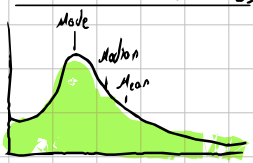
Measures of Asimetry

Skewness
Asimetria
Sesgo

Skewness

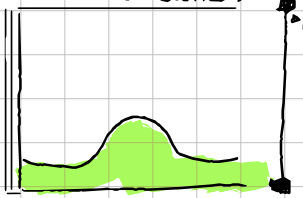
- Indica cuando los datos estan concentrados en un lado o en el otro de el mean.
- Para fijarnos debemos ver hacia donde esta la cola

Positive Skewness



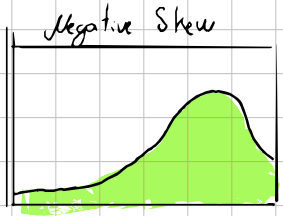
- Mean > median
- Los outliers estaran a la derecha
- Mode: value with the higher representation

Zero Skewness



- Mean = median = mode

Negative Skew



- Mean < 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:

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

Most common measure of variability

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

Coefficient of Variation

$$\frac{\text{Standar deviation}}{\text{mean}} \qquad c_v = \frac{\sigma}{\mu}$$

Esto se usa para comparar datasets. No tiene una unidad de medida para

$$\hat{c}_v = \frac{s}{\bar{x}}$$

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

$$\frac{\text{Covariance}(x,y)}{\text{Standard deviation}(x) \cdot \text{Standard deviation}(y)} \qquad \frac{S_{xy}}{S_x S_y} \qquad \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

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

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

Siendo 1 el maximo posible de relación entre ellos a esto lo llamamos "PERFECT POSITIVE CORRELATION" Basicamente que una variable esta totalmente explicada por otra

Correlacion de 0 significa independencia total

Correlacion de -1 significa que son opuestos aunque es extremadamente raro ver un -1

* Ej compañía de helados (verano) y de abrigos (invierno)

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

Correlacion not imply Causation

Central Limit Theorem

Sampling distribution of any distribution will aproximate 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's the variability of the different samples we extracted

It's used to verify how well you aproximated the true mean

A más sample (n) menor es, indicandonos que no aproximamos más

Estimators and Estimates

Estimator. Aproximación dependiente unicamente 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 creamos a la estimación que favorece a un resultado concreto

Efficiency: A mas efficiente menos bias y variance tendrá el estimator

Statistic	Estimator
Broad Term	Type of Statistic

CONFIDANCE INTERVAL

Representacion 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

$$\left[\text{Point Estimate} - \text{Reliability Factor} \circ \text{Standard Error} \;,\; \text{Point Estimate} + \text{Reliability Factor} \circ \text{Standard Error}\right]$$

Point Estimate

$$\bar{X} - Z_{\alpha/2} \circ \frac{\sigma}{\sqrt{n}} \;,\; \bar{X} + Z_{\alpha/2} \circ \frac{\sigma}{\sqrt{n}}$$

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

$\bar{X}$ = sample mean

σ = Standar 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 = $1 - (1-\alpha)$ | General mente se usan valores de confianza de 90%, 95 o 99 por lo que $\alpha = 0.1$, 0.05 o 0.01

Calculo umbrales de confianza con T. Con variance unknown

1. Calculo standard error: $\frac{s}{\sqrt{n}}$ (sample mean)

2. Formula calculo intervalo sin conocimiento de varianza

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

2.1 Si supiesemos el population variance seria

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

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 les llamamos margen de error o ME

Por lo que definimos el intervalo 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

  Eg. • weight loss and blood sample

  • habits on each member of a couple

  • cause and effect. Nota de corte y aceptación a universidad

- **Independent Samples.**

  • Pop variance known

  • Pop variance unknown assumed different

  • Pop variance known assumed same

Dependant Samples

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

* Cuando trata-os con estadísticas en biología, tendemos a asumir distribuciones normales por lo extra-común que son.

Ej: calcio magnesio en sangre antes y después medicina

Patient	Before	After	Diff
4	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

$\alpha = 0.05$ $\quad n = 10$

mean = $\bar{d} = 0.33$

2. Calculo standar deviation

np.std(vals, ddof=1) → calculo sd for sample

$S = 0.455$

3. Una vez tenemos los dato de calculo calculamos el umbral de confianza

sample diff → Es student's T

$\bar{d} \pm ME$ Sample of normal $\quad \bar{d} \pm t_{n-1, \alpha/2} \cdot \frac{S}{\sqrt{n}} = 0.33 \pm t_{9, 0.025} \cdot \frac{0.455}{3.1622}$

$0.33 \pm 2.262 \cdot 0.14388$

0.33 ± 0.32545656

$[0.004, 0.655]$

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

Independent Samples known VARIANCE

1°. Type: known population variance

Departament 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 notas

diff x-y	
size	? (population)
mean	-7
pop std	

1. Calculo la varianza de la diferencia

$$\sigma^2_{diff} = \frac{\sigma_e^2}{n_e} + \frac{\sigma_m^2}{n_m} = \frac{10^2}{100} + \frac{5^2}{70} = 1.36$$

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 comparado los datos de Ingeniería con los de managment. Si se comprobasen los de managment con los de ingeniería el resultado cambiaría a $(4.72, 9.28)$ es decir **EL ORDEN DE COMPARACIÓN IMPORTA** (matemáticamente el resultado de una comparación a otra es simétrico respecto a 0)

Independent Samples - Same unknown variance

Ejemplo comparación de precios de manzanas 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	3.81
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

1° Paso, estimar el population variance

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

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} = \frac{(10-1)\cdot 0.18^2 + (8-1)\cdot 0.27^2}{10+8-2} =$$

$$= 0.05$$

$s_p^2 = 0.05$

$s = 0.22$

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

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

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

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

INDEPENDET 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

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