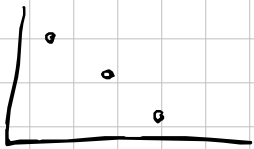
Tipos de Graficos

Qualetative data

PIE CHART

→ % total

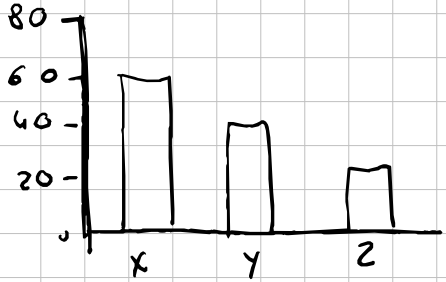
Dot plot



→ Categorias

Quantitative - Bar graph

→ Numbers
→ Distance
...



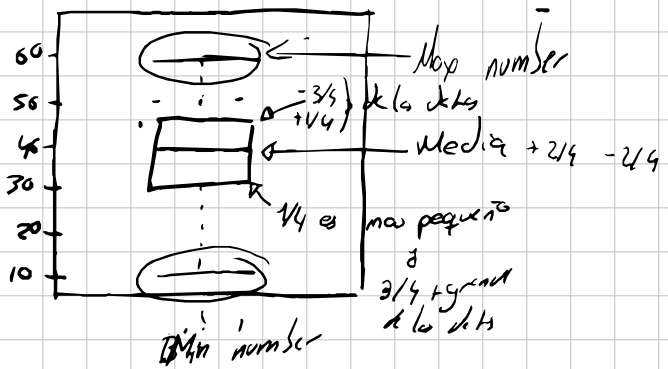
Histogram

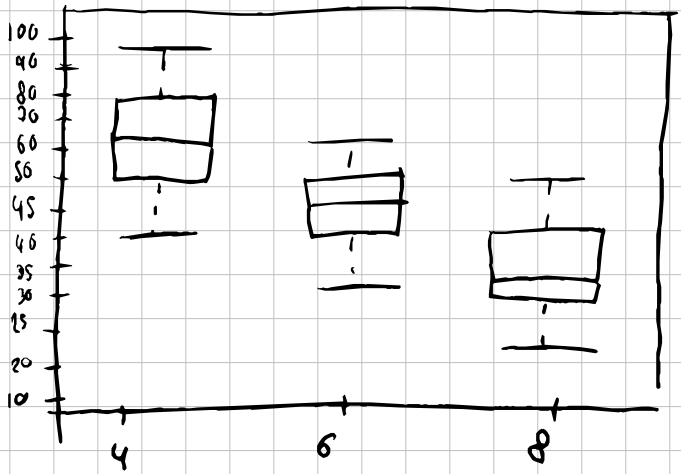
→ Permite adaptar el tamaño de los bloques a la frecuencia

→ Densidad: (altura de la barra)

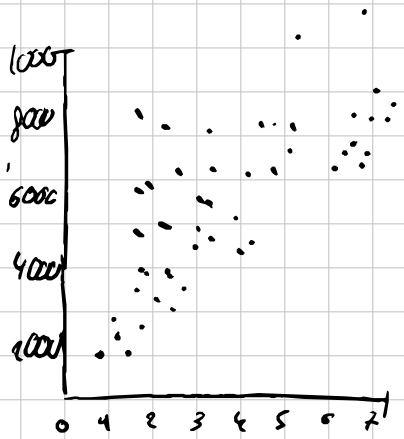
→ Porcentajes:

Box-Plot - Caja y bigote





Scatterplot - Dispersión



Usado principalmente para visualizar la relación entre dos valores

Como por ejemplo año de estudios y salarios

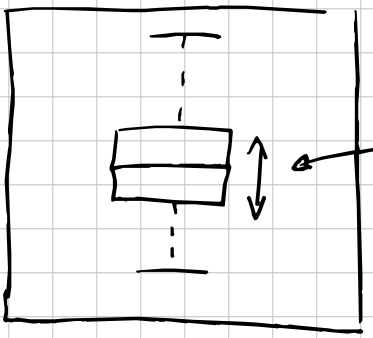
* Es importante dar todo el contexto necesario pero tampoco excesivo

Numerical Summary Measures

- Mean = average = media
- Median: punto medio, la mitad de los datos más altos y mitad más bajos.
  Mediana
- Moda: más repetido

- 1° Quartile o cuarto: 25% más bajo
- 2 Quartile: media
- 3 Quartile: 75% más bajo

Percentiles, the five Numbers Sumary, and Standard Deviation



Interquartile range

3rd quartile - 1st quartile

it measure how spread the data is

Standar Deviation: Scale with the units

$\bar{x}$ average of the numbers $x_0 \cdots x_n$

formula $$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Takes the squar root of the averag of each number minus the average

EXAMING OF HOW TO PRODUCE DATA

STATISTICAL INTERFERENCE

- Pick a random sample of the population to use it to calculate everything

- Population: whole dataset
- Parameter: quantity / value / data we are interested from the population
- Sample - muestra
- Statistic (estimate): parameter solo medido en el sample

Simple Random vs Stratified Random SAMPLING

- **Sample of convenience:**

  No suele ser una buena manera de sacar datos. Un ejemplo de este tipo sería coger solo clientes de amazon del barrio en lugar de toda España

  Bias. Sample que favorece un resultado concreto

  Selection Bias: sample of convenience hace más probable unos topicos sobre otros.

  Non-response bias: la gente que responde puede ser diferente a la que no responde

  voluntary response: generalmente los reviews solo vienen de clientes con o muy buenas o muy malas experiencias.

- **Simple random sample**
  - Select subject without replacement

- **Stratified Random Sample**

  Dividir el population en grupos de topicos similares "Strata"

Bias and Chance Error

Estimate = parameter + bias + chance error

(bias: desviación para todos los datos del sample)

Observation Studies

Son los resultados de comparación de datos obtenidos de una variable aunque no tiene porque ser ocasionada por la que obtuvo los datos para ser perjudicante que esa variable esté ligada a otras asociaciones

Ej +red meat =? +cancer

Pero realmente no es la carne roja per se es que la gente que come carne roja suele hacer menos ejercicio y beber más alcohol

A este alcohol o ejercicio se le llama lurking variable

Para poder asegurar un vínculo es necesario un experimento para asegurar su efecto.

Similar a los estudios médicos y los grupos medicados y de control con el placebo

Un experimento debe ser double-blind ni los pacientes ni experimentadores deben saber en qué grupo está el paciente

Nondominization

- $\mathcal{N}$

Probability

Addition A, B dependent $P(A \text{ or } B) = P(A) + P(B)$

Independent A & B

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

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

$$P(A \& B) = P(A) \cdot P(A \mid B)$$

- Useful in independent P

Bayes rule

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

$$P(D|+) = \frac{P(+|D)\,P(D)}{P(+)} \Rightarrow$$

$$\frac{P(+|D)\,P(D)}{P(+|D)P(D) + P(+|no\ D)\,P(no\ D)}$$

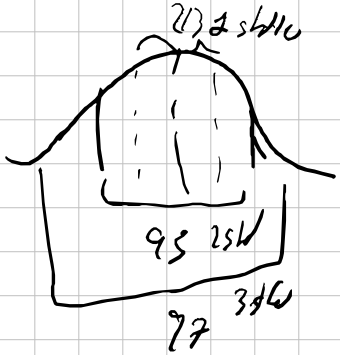
Normal curve - bell shape

Empirical rule

1 - 2/3 data in 1 Std dv

95% in 2 Standard

99.7 in 3 standard



Standard data

$$Z = \frac{data - \bar{x}}{s} = Z\text{-score}$$ (No Un.)

~~t, kg, kc~~

mean 0

stddv 1

Normal Approximation

Calculate % of height between

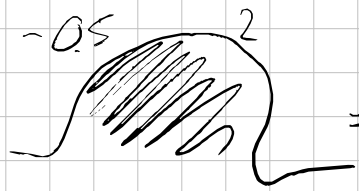
67.4 & 71.9

mean = 68.3

s = 1.8

$$\frac{67.4 - 68.3}{1.8} = -0.5$$

$$\frac{71.9 - 68.3}{1.8} = 2$$

Area 

= 

- 

= 66.8%

Calculos % normal data

$$Z = \frac{\text{valor} - \bar{x}}{s}$$

$$\text{valor} = \bar{x} - Zs$$

Binomial Setting and Coefficient

2 in 3 in a 49% setting

$P(2 \text{ in } 3) = P(AAB \text{ or } ABA \text{ or } BAA) =$

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

$P(B) \cdot P(A) \cdot P(A) =$

$= \underline{3} \times (0.49)(0.49)(0.51) =$

Independent repetitions $= m$

for 2 in 5

$5 \cdot (0.49)(0.49)(0.51)(0.51)(.51)$

Binomial Coefficient

$$m = \frac{n!}{k!(n-k)!}$$

Binomial Formula

$$P(k \text{ success in } n \text{ experiments}) = \frac{n!}{k! \cdot (n-k)!} p^k (1-p)^{n-1}$$

Expected value or SE for the sum

S_n = sum n draws $\quad S_n = n\bar{x}_n$

$E(S_n) = n\mu \quad SE(S_n) = \sqrt{n}\sigma$

The more n the more the SE goes up

- SE
  more n $\left[\begin{array}{l} +SE(S_n) \\ -SE(n) \end{array}\right.$

less n $\left[\begin{array}{l} +SE(n) \\ -SE(S_n) \end{array}\right.$

Simulating Values: X has k outcomes

$$\mu = \sum_{i=1}^{k} x_i P(X = x_i)$$

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 P(X = x_i)$$

X has ∞ outcomes, density f

(E.g. which follow normal curve) * We will not use this:

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx \qquad \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

Square root Law