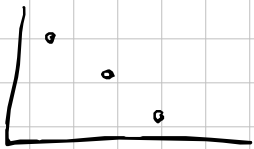
Tipos de Graficos

Qualetative data

PIE CHART

→ % total

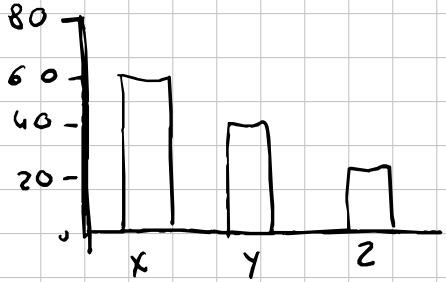
Dot plot



→ Categorias

Quantitative - Bar graph

→ Numbers
→ Distance
...



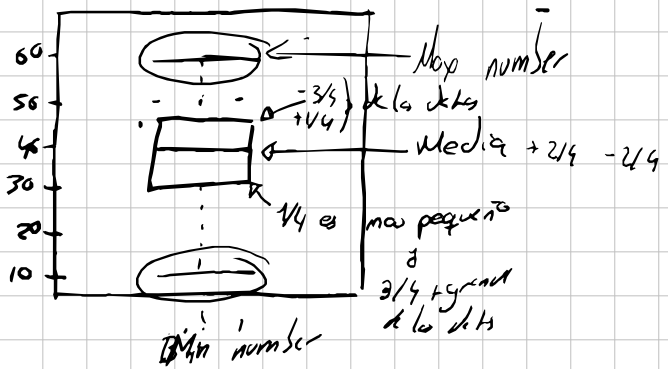
Histogram

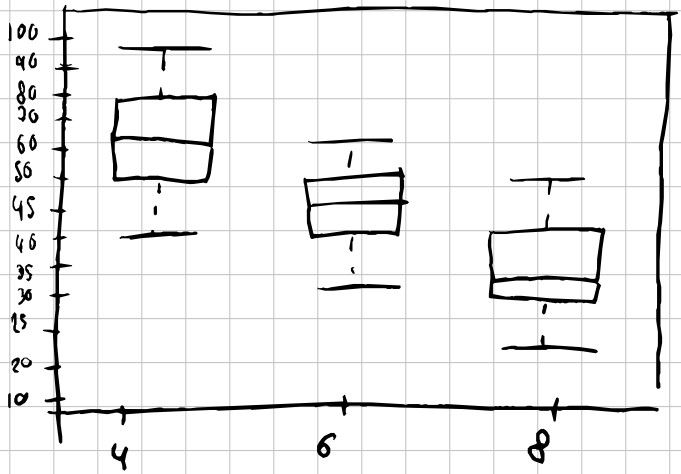
→ Permite adaptar el tamaño de los bloques a la frecuencia

→ Densidad: (altura de la barra)

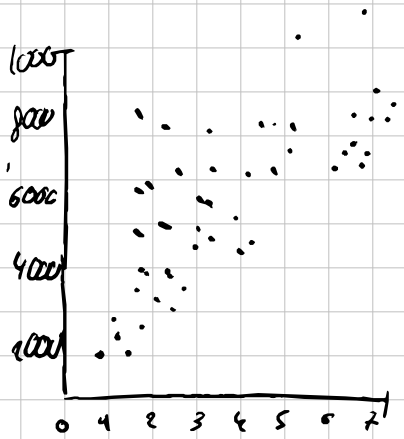
→ Porcentajes:

Box-Plot - Caja y bigote





Scatterplot - Dispersión



Usado principalmente para visualizar la relación entre dos valores

Como por ejemplo año de estudios y salarios

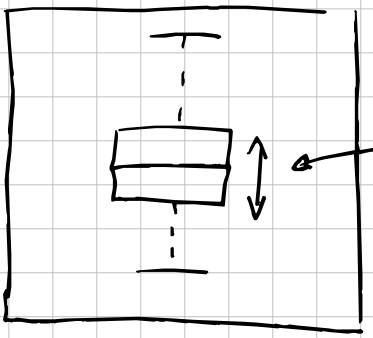
* Es importante dar todo el contexto necesario pero tampoco excesivo

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: $P(A \text{ or } B) = PA + PB$

Independent A & B

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

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

$$P(A \& B) = PA \cdot P(A|B)$$

- Useful in independent P

Bayes rule

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

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

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

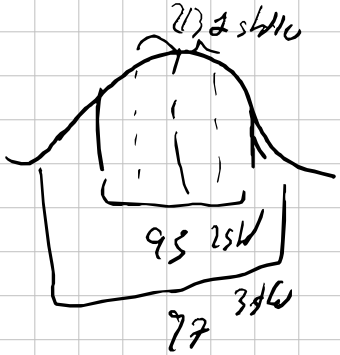
Normal curve - bell shape

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

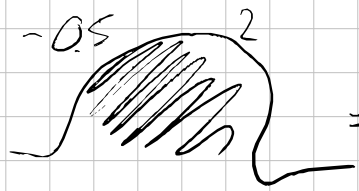
67.4 & 71.9

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

$SE(\bar{x}_n)$ goes to zero as the sample size increase.

more large closer to μ

Law of large numbers

this only apply for averages and % it do not apply for sums.

Since for sums more large the size more SE will have

CENTRAL LIMIT THEOREM

As n grow larger more similer to the normal curve will be.

* In large sample draws WITH REPLACEMENT

$\mu = p$

$\sigma = \sqrt{p(1-p)}$

$SE(X) = \sqrt{np(1-p)}$

Use it when:

- We sample with replacement
- or
- we simulate independent random variables

- Statistic we are looking must be a Sum (average and % are sums in disguise)
- Sample size must be large enought (40+)

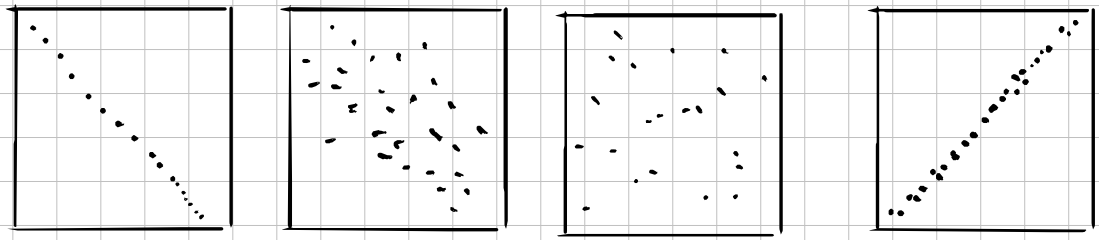
Correlation Coefficient

$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y}$$

Its help to calculate strenght of relationship and direction of the line

Correlation measures linear association

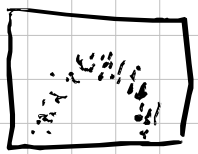
r is always between -1 and 1 and it gives the direction of the association and it absolute value gives the strenght



Es raro que sea -1 o 1

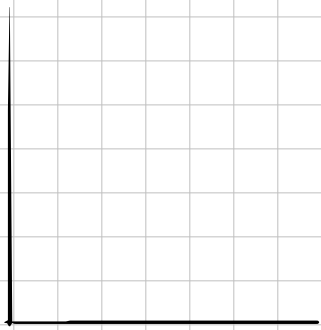
EL CORRELATION COEFFICIENT SOLO ES UTIL EN ASOCIACIONES LINEALES

Ej → Inutil



(Claramente hay una relación pero r = 0)

Regression Line



$\hat{y}_i = a + b x_i$

to calculate the line

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (a + b x_i))^2$$

The method called to calculate the line is called the method of least squares. It turns:

$$b = r \frac{s_y}{s_x} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

This line $\hat{y} = a + bx$ is called <u>Regression Line</u>

Regresion to the Mean, Regresion Fallacy

Main use of regresion. Predict y from x

$$\hat{y} = a + bx$$

$$b = r \frac{S_y}{S_x}$$

Basicamente podemos ver que por ejemplo los puntuajes mas bajos en un retest no son los mas bajos del test y lo mismo para los mas altos ya que hay un grado de variabilidad a suerte que influye

A veces esto se trata como un que es un dato erroneo pero no tiene porque ser asi a esto le llamamos "regresion fallacy"

Predicting y from x and x from y

Ej: $\overline{Midterm} = 49.5$

$\overline{final} = 69.1$

$s_{mid} = 10.2$

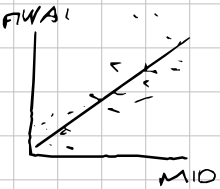
$s_{final} = 11.8$

$r = 0.67$

$(x-\mu)/\sigma = (41-49)/10.2 = -0.8$

Predict score of student 41 mid term

$\hat{y} = a + bx \qquad b = \frac{s_y}{s_x}$



41 is 8.5 below average

41 is $0.83 s_{mid}$ below average

$r \cdot 0.83 \cdot s_{Final} =$

$= 69.1 - 0.67 \cdot 0.83 \cdot 11.8 = 62$

$y = \overline{final} \pm r \cdot (x_{smd} \text{ away from mid}) \cdot s_{final}$

(above: +, below: −)

Ej 2 → 5d

58 is 8.5 above → $+0.83 s_{mid}$

$= 69.1 + 0.67 \cdot 0.83 \cdot 11.8 = 75.7$

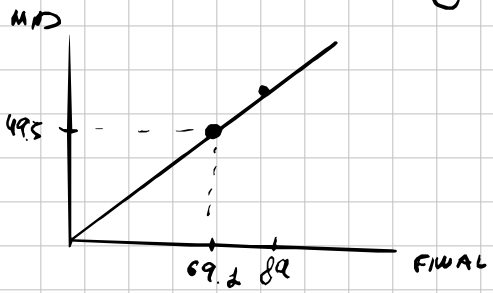
Predict y from x and x from y

DO NOT USE

The same regresion line

Next ⇒

Predict Mid from final 89



89 is 9.9 or 0.84 stdev above average

r · 0.84 sdev

$\overline{Midterm} = 49.5$

$\overline{final} = 69.1$

$s_{mid} = 10.2$

$s_{final} = 11.8$

$r = 0.67$

$= 49.5 + 0.67 \cdot 0.84 \cdot 10.2 = 55.2$

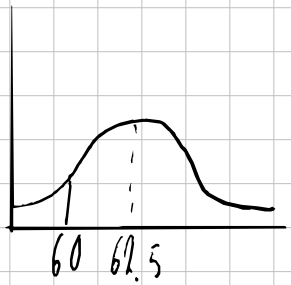
Normal aproximation given x

Remember: in order to have regression we need an american football ball size

If we have a plot with that characteristic we can do normal standarization $\frac{x - \hat{y}}{\sqrt{1-r^2} \cdot s_y}$

To do that: substract $\hat{y}$ and divide by $\sqrt{1-r^2} \cdot s_y$

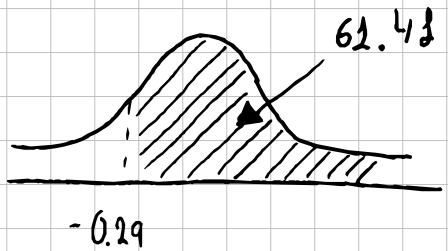
Ej. What % of students who scored around 41 on the midterm scored more than 60 on the final



We already predicted 62.5 as final score, so taking it as a normal curve!

First Standarice

$$\frac{60 - 62.5}{\sqrt{1-0.67^2} \cdot 11.8} = -0.29$$

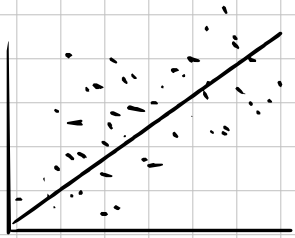


Residuals

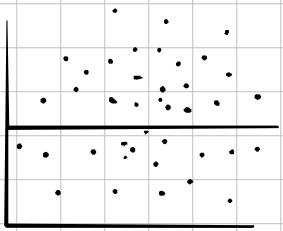
Residuals are the difference between observed and predicted

$$y_i - \hat{y}_i = e_i \qquad i = 1 \dots n$$

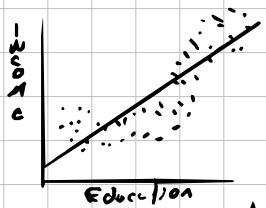
We use residuals to check if the use of regresion is apropiate. It should show an unstructured horizontal band

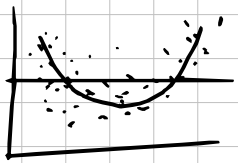


residual against x values



Example that data may not be applied and we may need to transform the data:





We observe a curve so may not linear

Ej of transformation:

regress $\sqrt{\text{income}}$ or $\log(\text{income})$

then we transform back the results

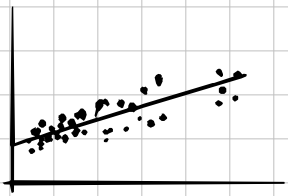
Other Example Residual Plot - HETEROSCEDASTIC

This fan shapes plots mean the scatters are heteroscedastic. (more variability on one side)

Normally we can transform it to work with more linear data. We may need another transform if the data lose linear

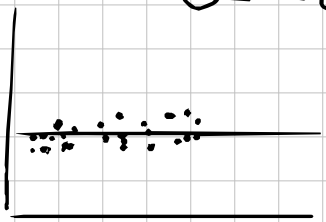
OUTLINERS & INFLUENTIAL POINTS

Eg. Outliers ⊙← Outlier



—

⊙← Outlier

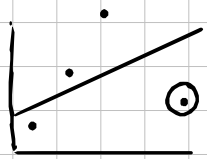


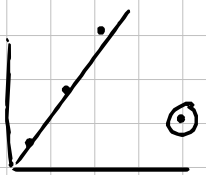
- We should always exam outliers, they always mean either an interesting phenomenom or a kind of typo. In case of typo we must remove it from the data

LEVERAGE AND INFLUENTIAL POINTS

A point whose x-value is far from the mean of x-values has high **Leverage** and will potentially make a big change in the regression line.

Eg.





In the example we see how without that point the regression has too much meaning in our prediction since that point deviate a lot for - the rest it has high leverage and its also an INFLUENTIAL POINT

OTHER ISSUES

- We must avoid predicting "y" by extrapolation since at x-values outside the range of x-values used for the regression the linear relationship often breaks
- Beware if the data comes in summaries from other data since it will tend to overstate the strength of the relationship
- Regression analysis often report 'R-squared': $R^2 = r^2$
  It gives the fraction of the variation in the y values that is explained by the regression line
  Higher R^2 mean the regression line does a good job explaining a lot of the y-values variation

CONFIDENCE INTERVALS

$\mu = 60\%$ $\quad SE = \frac{\sigma}{\sqrt{1000}} = 1.6\%$

Sample size = 1000

90 ~~3~~ SE

ranges: 95% confidence (Empirical rule 2 SE)

99 3 SE

95% $\Rightarrow$ $60\% \pm (1.6\%) \cdot 2$

CENTRAL LIMIT THEOREM FOR CONFIDENCE INTERVALS

Ej μ = approval among 140 million voters

est-date = aproval % among votes in sample

μ = speed of light

estimate = average of 30 measures

CONFIDENCE INTERVAL

estimate $\pm$ z SE

$SE = \frac{\sigma}{\sqrt{n}}$ $\rightarrow$ Std Dev pop

z = value in z table for X% of confidence

Ej 95% $\Rightarrow$ z = 1.96

90% $\rightarrow$ 1.65

99 $\rightarrow$ 2.58

BOOTSTRAP PRINCIPLE

We can estimate σ by its sample version 's' and still get an approximately correct confidence interval

Ex

1000 voters 58% approves president

$SE = \frac{\sigma}{\sqrt{n}} \cdot 100\%$ where $\sigma = \sqrt{p(1-p)}$

p = proportion who approve (58%)

$$SE = \frac{\sqrt{0.58(1-0.58)}}{\sqrt{1000}} = \frac{0.49}{\sqrt{1000}}.$$

$$58\% \pm 2\frac{0.49}{\sqrt{1000}} \rightarrow [54.9\%, 61.1\%] \text{ 95\%}$$

For 20%

$$20\% \pm 2\frac{\sqrt{0.2(1-0.1)}}{\sqrt{1000}} = 20\% \pm 2.52\% \text{ at 95\%}$$

Ej:

estimate = average of 30 measures of speed of light

measurement = speed of light + measure error

EXTRA - CONFIDENCE INTERVAL

2 SE = margin of error, larger n smaller the error

since $SE = \frac{\sigma}{\sqrt{n}}$

$\text{estimate} \pm \frac{1}{\sqrt{n}}$ ← Fast calculation

because $\sigma = \underbrace{\sqrt{p(1-p)} \leq \frac{1}{2}}$

↑ Always

If not given the confidence level standard is 95%

Shrink Margin of Error

from 500 with 5400 ME to 2000 ME

how many N

$$\left(\frac{5400}{2000}\right)^2 \cdot 500 = n \qquad \left(\frac{ME_{old}}{ME_{new}}\right)^2 \cdot n_{actual}$$

$$= n_{new}$$