STATISTICS
Descriptive Statistics

Victormanuel.casero@uclm.es
Despacho 2-A15 (Edificio Politécnico)
https://www.uclm.es/grupos/oed

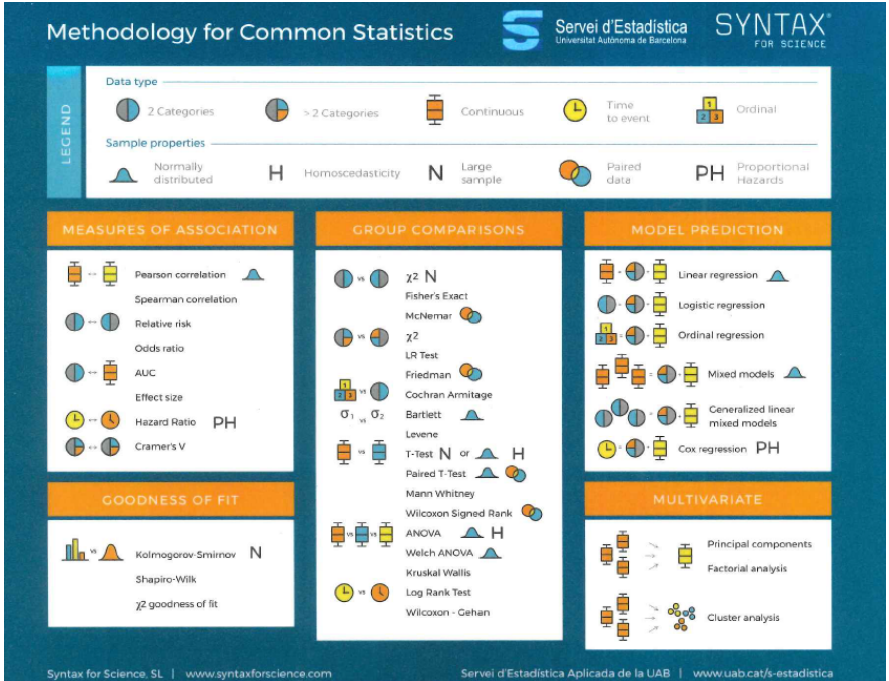
Statistics?

	Plant	Type	Treatment	conc	uptake
1	Qn1	Quebec	nonchilled	95	16.0
2	Qn1	Quebec	nonchilled	175	30.4
3	Qn1	C			
4	Qn1	C			
5	Qn1	C			
6	Qn1	C			
7	Qn1	C			
8	Qn2	C			
9	Qn2	C			
10	Qn2	C			
11	Qn2	C			
12	Qn2	C			
13	Qn2	C			
14	Qn2	C			

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.4					
Duster 360	14.3	8	360.0	245	3.21	3.5					
Merc 240D	24.4	4	146.7	62	3.69	3.1					
Merc 230	22.8	4	140.8	95	3.92	3.1					
Merc 280	19.2	6	167.6	123	3.92	3.4					
Merc 280C	17.8	6	167.6	123	3.92	3.4					
Merc 450SE	16.4	8	275.8	180	3.07	4.0					
Merc 450SL	17.3	8	275.8	180	3.07	3.7					
Merc 450SLC	15.2	8	275.8	180	3.07	3.7					

Statistics... Inference (decision making)

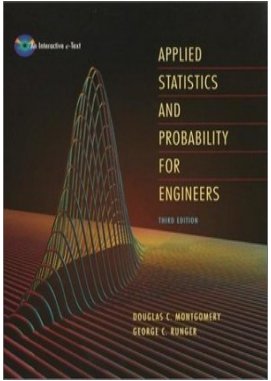
Infer conclusions for the '**population**' (**probability**)
from a '**sample**' (**descriptive statistics**).


Methodology for Common Statistics
Servei d'Estadística
Universitat Autònoma de Barcelona
SYNTAX
FOR SCIENCE
LEGEND
Data type
2 Categories
> 2 Categories
Continuous
Time to event
Ordinal
Sample properties
Normally distributed
H Homoscedasticity
N Large sample
Paired data
PH Proportional Hazards
MEASURES OF ASSOCIATION
Pearson correlation
Spearman correlation
Relative risk
Odds ratio
AUC
Effect size
Hazard Ratio PH
Cramer's V
GOODNESS OF FIT
Kolmogorov-Smirnov N
Shapiro-Wilk
χ2 goodness of fit
GROUP COMPARISONS
χ2 N
Fisher's Exact
McNemar
χ2
LR Test
Friedman
Cochran Armitage
Bartlett
Levene
T-Test N or H
Paired T-Test
Mann Whitney
Wilcoxon Signed Rank
ANOVA H
Welch ANOVA
Kruskal Wallis
Log Rank Test
Wilcoxon - Gehan
MODEL PREDICTION
Linear regression
Logistic regression
Ordinal regression
Mixed models
Generalized linear mixed models
Cox regression PH
MULTIVARIATE
Principal components
Factorial analysis
Cluster analysis
Syntax for Science, SL | www.syntaxforscience.com
Servei d'Estadística Aplicada de la UAB | www.uab.cat/s-estadistica

Outline, Acknowledgments & References

1. Descriptive Statistics
2. Probability
3. Inference

Jesús López–Fidalgo
Mercedes Fernández
Juan José Muñoz
Raúl Rivilla
Sergio Pozuelo



Descriptive Statistics Index

Data & variable (univariate)

Generally, unclassified data of a *variable*
(raw data) (phenomenon, characteristic... of interest)
from a *sample*.
(*individuals/Experimental units*)

Example: Repairing time after failure (units!)

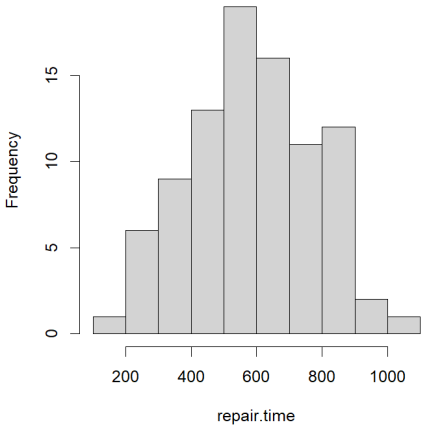
876	578	718	388	562	971	698	298	673
537	642	856	376	508	529	393	354	725
811	504	807	719	464	410	491	557	771
685	448	571	189	661	877	563	647	447
336	526	624	605	496	296	628	481	224
868	804	210	421	435	291	393	605	341
352	374	267	684	685	460	570	928	516
885	751	561	1020	592	814	843	466	498
562	739	562	817	690	720	758	731	480
559	505	703	809	706	626	631	585	639

Data = Information ?

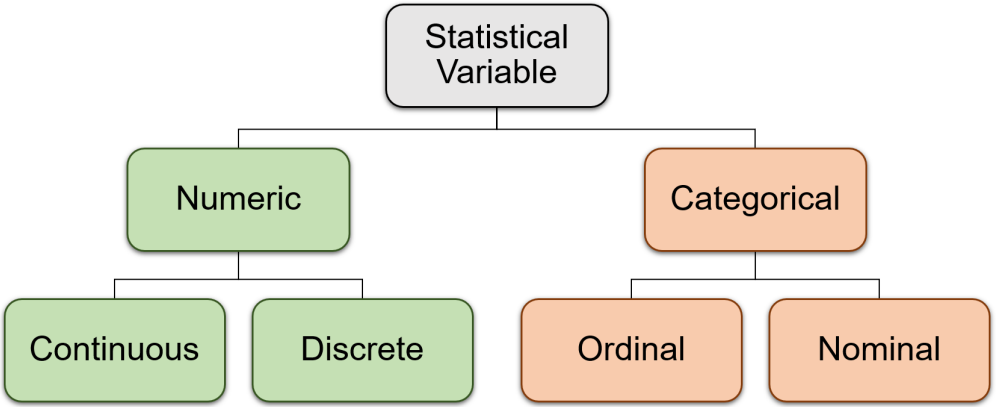
Data description example

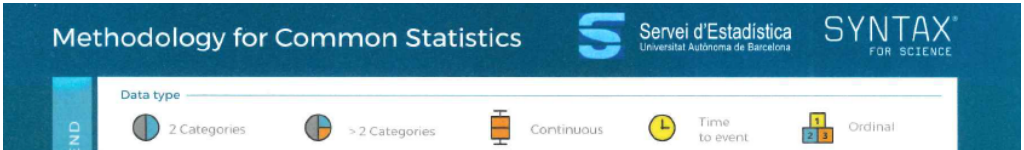
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
189.0   464.5   574.5   588.6   718.8  1020.0
```

... described?

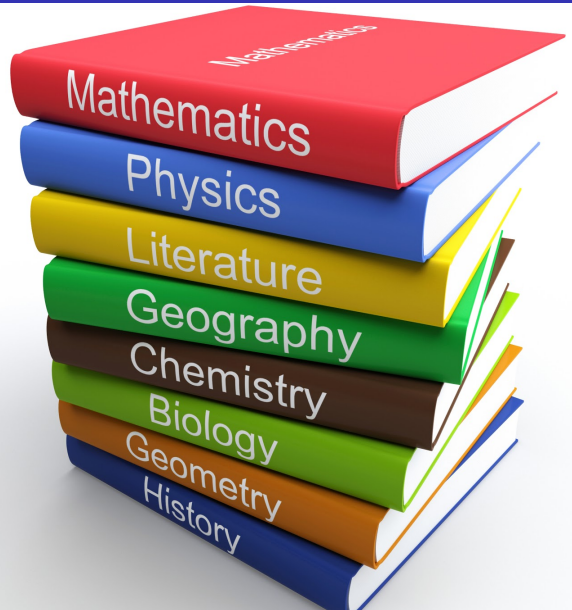


Types of variables





Examples


Mathematics
Physics
Literature
Geography
Chemistry
Biology
Geometry
History

...

Examples

...

Toma de datos

Grosor del papel

...

Frequencies ($\sim$ Classification)

for discrete, categorical...

Last Digit lottery prizes	Absolute frequency	Cumulative frequency	Percentage frequency	Cum.percent. frequency
x_i	f_i	F_i	p_i	P_i
0	19	19	9.5	9.5
1	8	27	4.0	13.5
2	13	40	6.5	20.0
3	20	...	...	...
4	26	86	13.0	43.0
5	31	117	15.5	58.5
6	26	143	13.0	71.5
7	20	163	10.0	81.5
8	20	183	10.0	91.5
9	17	200	8.5	100.0
Total	n=...		100.0	

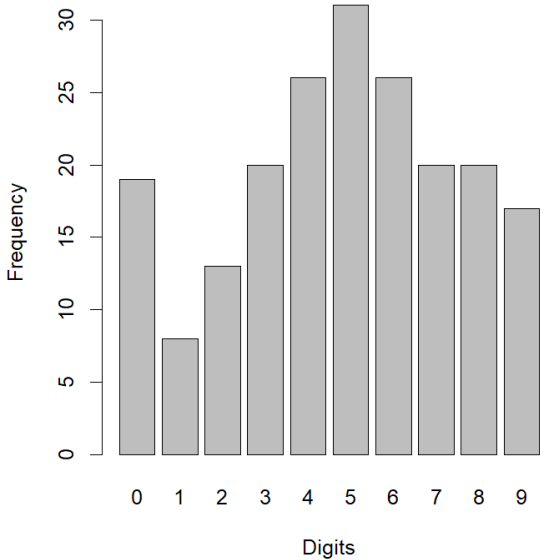
Cumulative frequencies have no sense for nominal variables!

Preguntas sobre tablas de frecuencias

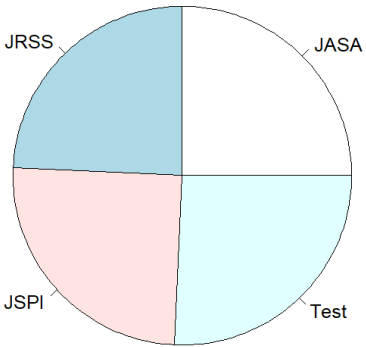
- ¿Cuántos premios han acabado en un número menor que 7?
- ¿Qué porcentaje de premios han acabado en 0, 1, 2 ó 3?
- ¿Qué dígito es tal que al menos el 50% de los premios acaba en él o en un dígito inferior a él?

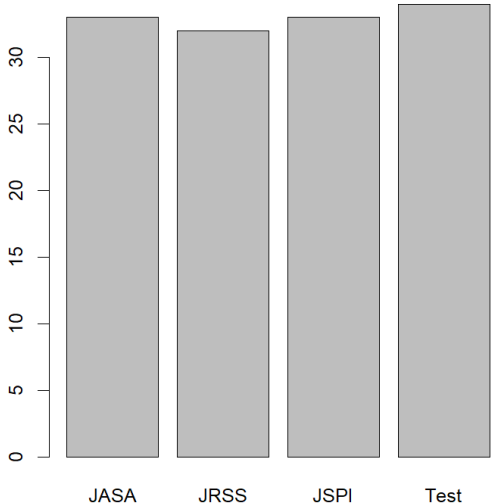
... have no sense for nominal variables!
... color azul o menor????

Bar chart (for . . .)



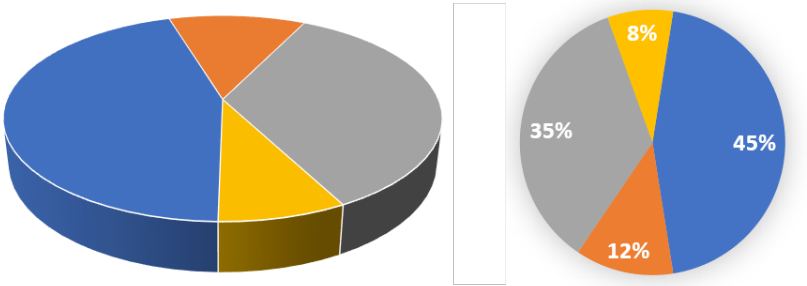
Pie chart (for . . . be cautious!)





Good Practice: Include % or/and n

Con perspectiva... peor!



Frequencies ($\sim$ Classification)

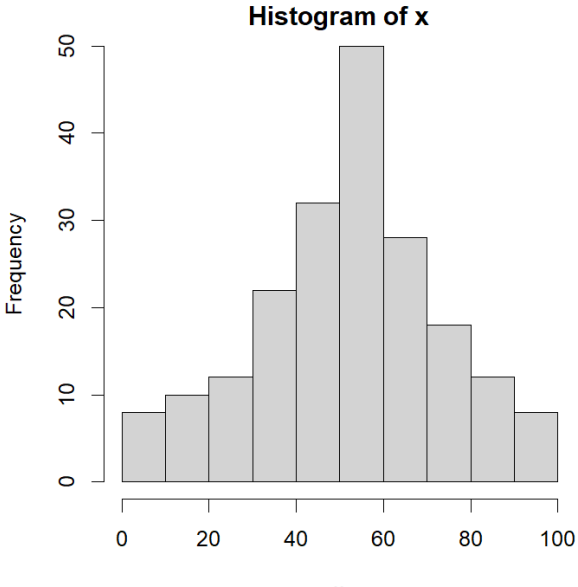
for grouped data (continuous, discrete ...

Example: Grades between 0 an 100

Class limits	Grade	Absolute frequency	Cumulative frequency	Percentage frequency	Cum.percent. frequency
	x_i	...			
0-10	5	8	8	4	4
10-20	15	10	18	5	...
20-30	25	12	30	...	15
30-40	35	22	...	11	26
40-50	45	...	84	16	42
50-60	55	50	134	25	67
60-70	65	28	162	14	81
70-80	75	18	180	9	90
80-90	85	12	192	6	96
90-100	95	8	200	4	100

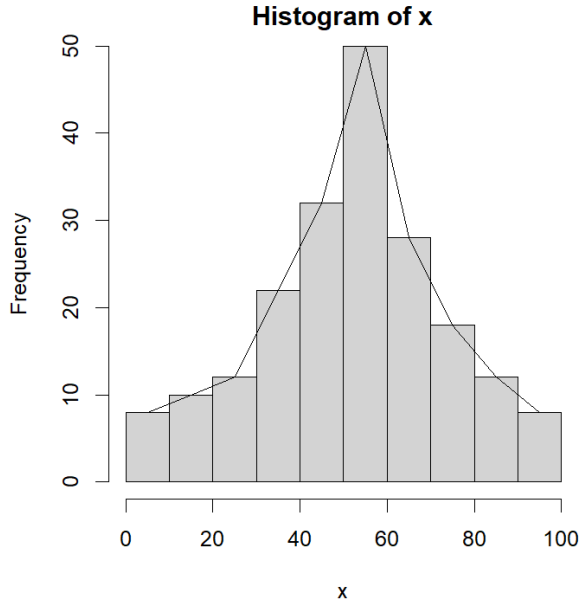
[0,10)?
or (0,10]?

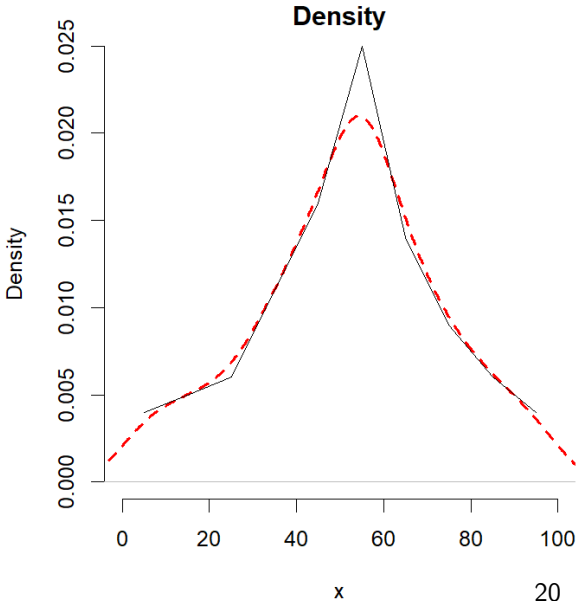
Histogram



$f_i(p_i, \ldots)$

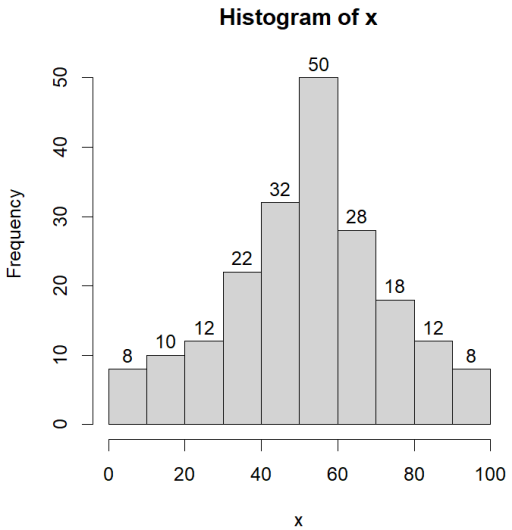
Histogram + Frequency Polygon ($\sim$ Density)





How many classes (bins)?

https://en.wikipedia.org/wiki/Histogram#Number_of_bins_and_width



$\sqrt{n}$,

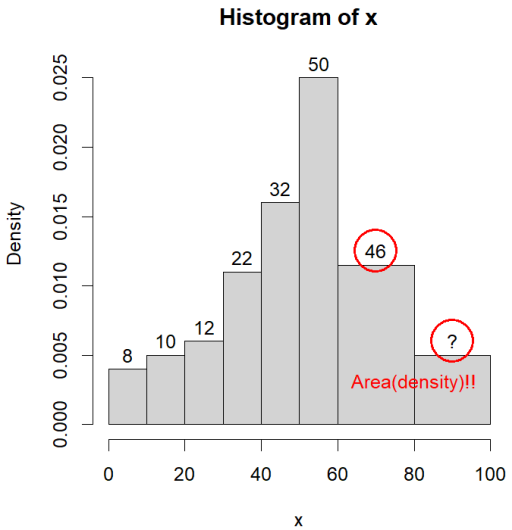
$$\begin{cases} n \leq 50 & \rightarrow 5-8 \text{ classes}, \\ n > 50 & \rightarrow 8-12 \text{ classes}, \end{cases}$$

Sturges's formula

...

Different widths?

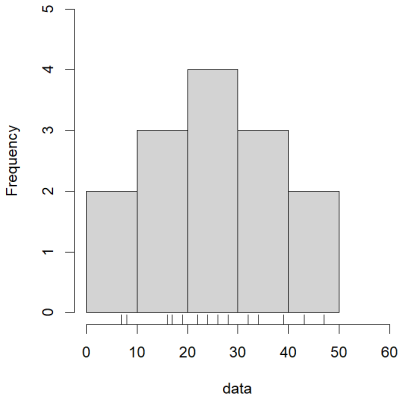
AREA is the KEY

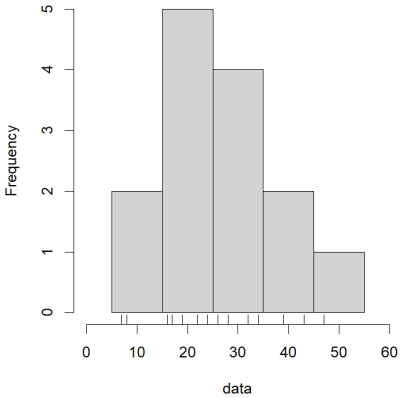


Where?

Be careful!

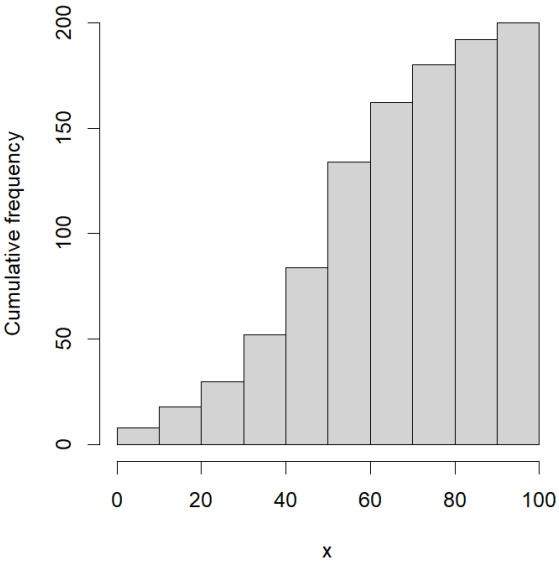
Data: 26,39,19,28,47,7,22,43,32,8,17,24,34,16



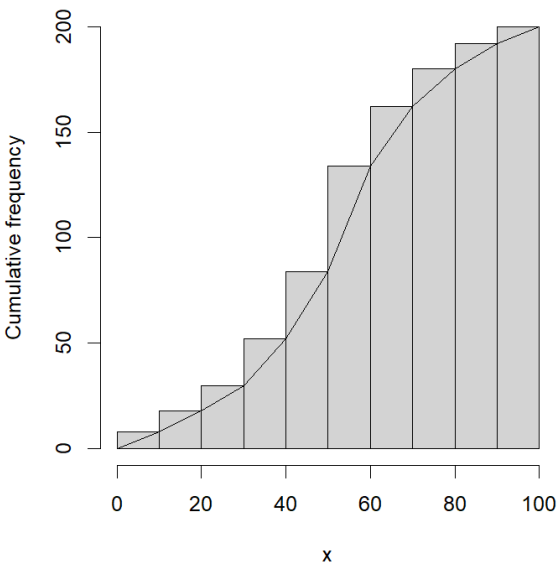


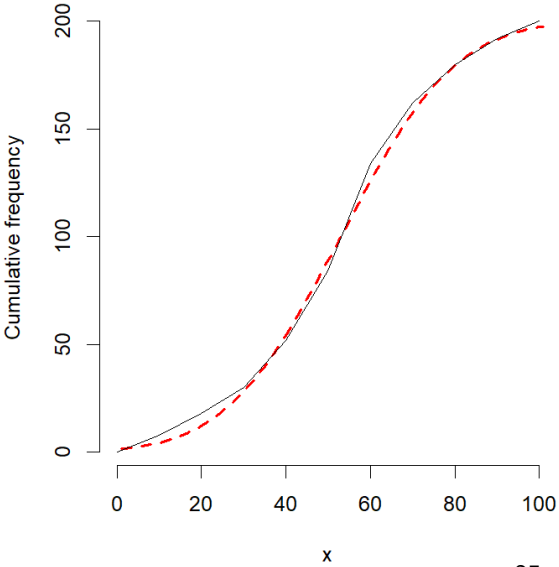
Symmetric or not??

Histogram (Cumulative Frequency)



Histogram + Frequency Polygon (Cumulative)



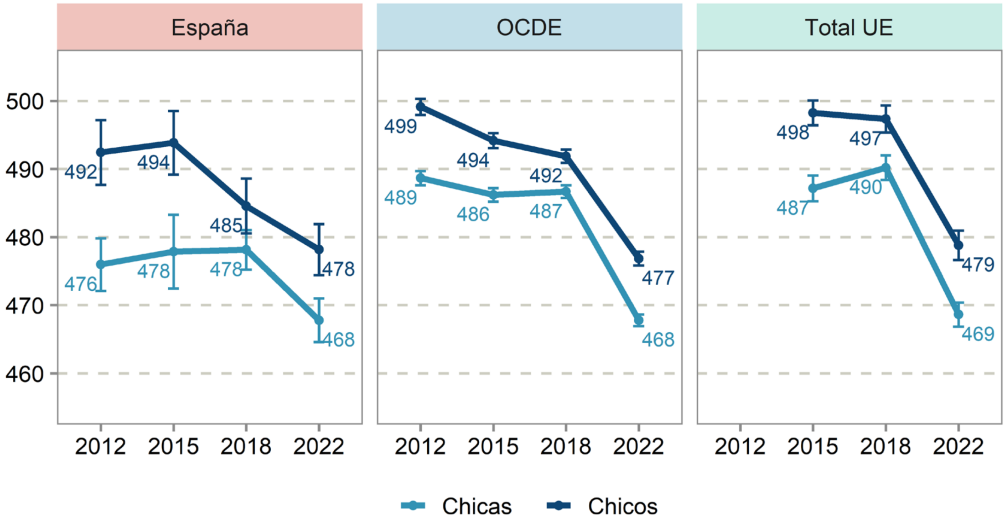


Other graphs (univariate)

- Stem-and-leaf diagram
  (https://en.wikipedia.org/wiki/Stem-and-leaf_display)
- Run chart
  (https://en.wikipedia.org/wiki/Run_chart)
- Pareto chart
  (https://en.wikipedia.org/wiki/Pareto_chart)
- ...

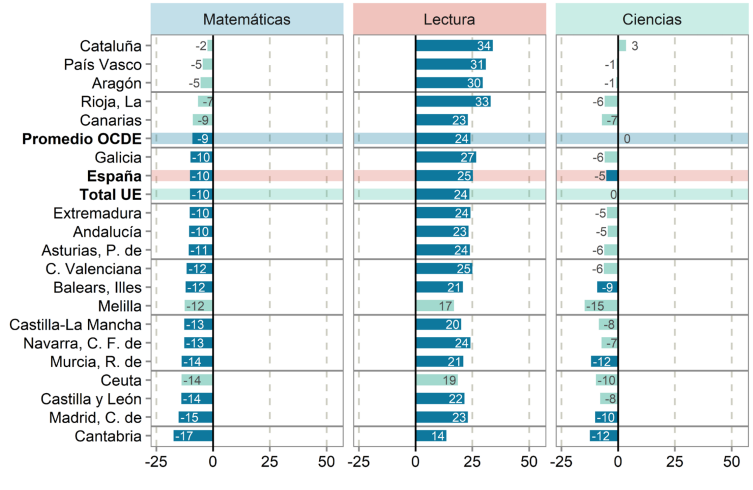
Examples

Figura 3.2. Tendencia en las puntuaciones medias en matemáticas según género



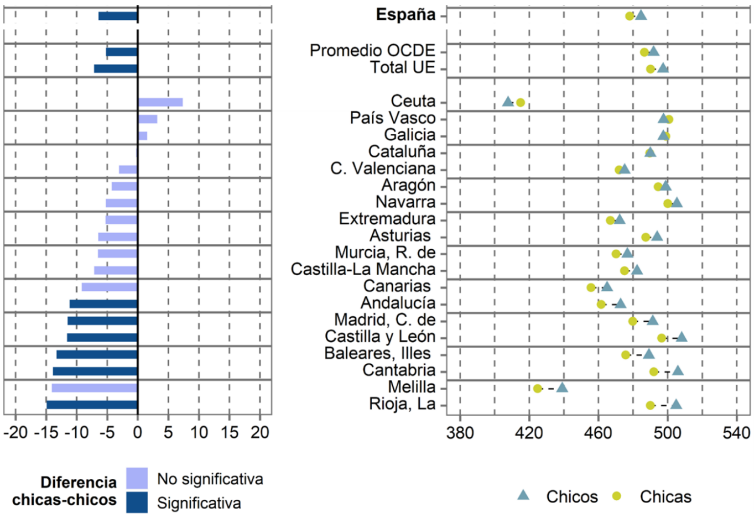
Examples

Figura 3.1.b. Diferencia en las puntuaciones medias de matemáticas, lectura y ciencias según género (chicas-chicos), significatividad del 95 % de las comunidades y ciudades autónomas participantes en PISA 2022



Examples

Figura 3.1. Diferencia en las puntuaciones medias de matemáticas según el género, significatividad del 5%



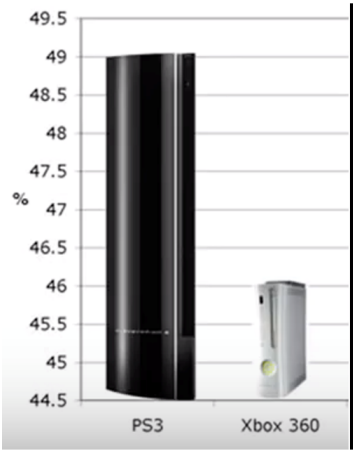
Test questions

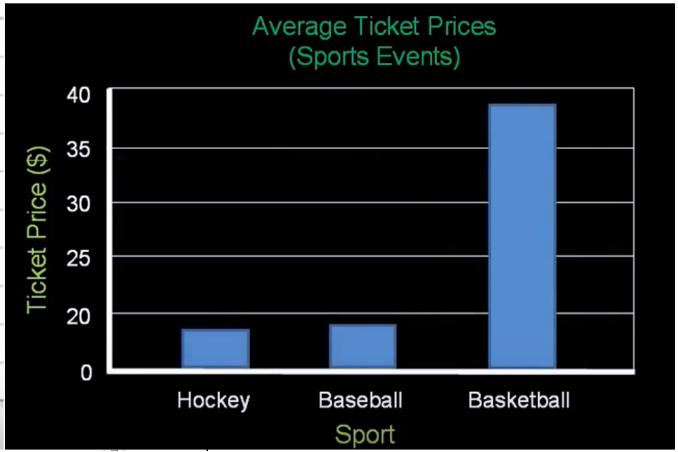
1. Los gráficos más adecuados para las variables *consumo* $\in [2, 12]$ (en l/100 km) y n^o *de cilindros* $= \{1, 2, 3..,10\}$ son
   a) diagrama de barras para consumo y para n^o de cilindros.
   b) boxplot para consumo e histograma para n^o de cilindros.
   c) boxplot para n^o de cilindros y diagrama de barras para consumo.
   d) diagrama de barras para n^o de cilindros e histograma para consumo.

2. Se estudian las temperaturas obtenidas en un proceso de fabricación a través de la clasificación: alta, media y baja, la variable considerada es
   a) Una variable cuantitativa discreta.
   b) Una variable cualitativa nominal.
   c) Una variable cualitativa ordinal.
   d) Una variable cualitativa dicotómica.

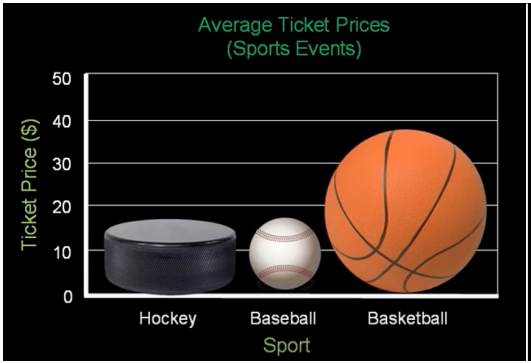
3. En una tabla de frecuencias
   a) la suma de frecuencias relativas acumuladas ha de ser el tamaño de la muestra.
   b) la última frecuencia relativa acumulada ha de ser 1.
   c) la suma de frecuencias absolutas acumuladas ha de ser el tamaño de la muestra.
   d) la última frecuencia absoluta acumulada ha de ser 1.

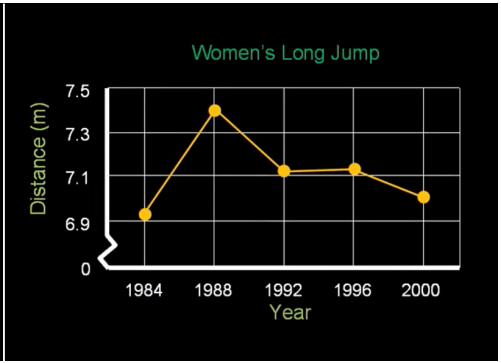
Misleading charts





Misleading charts, or not?





There are many examples...
For instance: `https://venngage.com/blog/misleading-graphs`
... or `https://en.wikipedia.org/wiki/Misleading_graph`

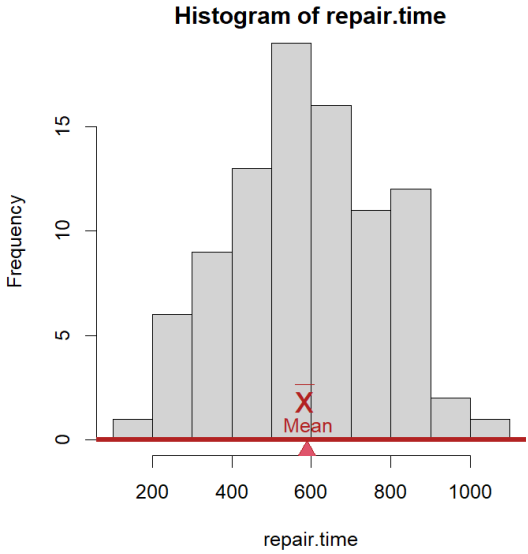
NUMERICAL MEASURES

```
Descriptive Statistics
repair.time
N: 90

                  repair.time
---------------- -------------
            Mean      588.62
         Std.Dev      184.22
             Min      189.00
              Q1      464.00
          Median      574.50
              Q3      719.00
             Max     1020.00
             MAD      189.77
             IQR      254.25
              CV        0.31
        Skewness        0.01
     SE.Skewness        0.25
        Kurtosis       -0.58
         N.Valid       90.00
       Pct.Valid      100.00
```

?

Central Tendency/Centrality

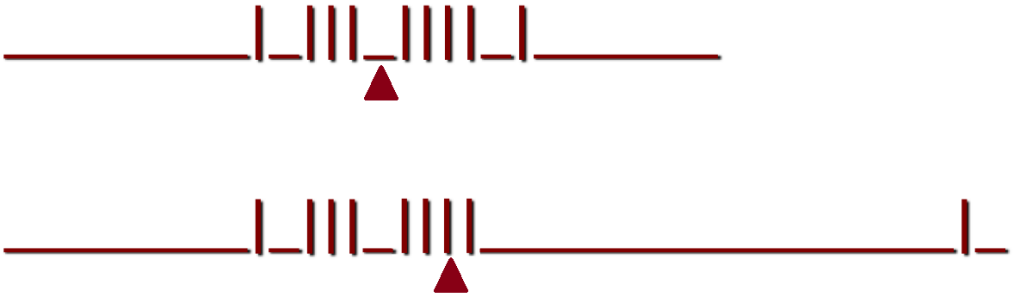


Formula?
La conoces!!

- Pero cuidado con su cálculo...
  - Datos: 0, 5, 5 y 10
    Media?
    Media agrupada? [0,5) y [5,10]
- y con su interpretación
  - Si yo tengo 100€ y tu 0€...
    en media tenemos 50€¿?

Falta formación ...

Influence of outliers in the mean



Data: 0, 5, 5, 10
Data: 0, 5, 5, __

Median

Central Tendency statistics

Example: 56, 63, 65, 62, 68, 65, 65, 72, 65, 70
Sorted! : 56, 62, 63, 65, 65, 65, 65, 68, 70, 72

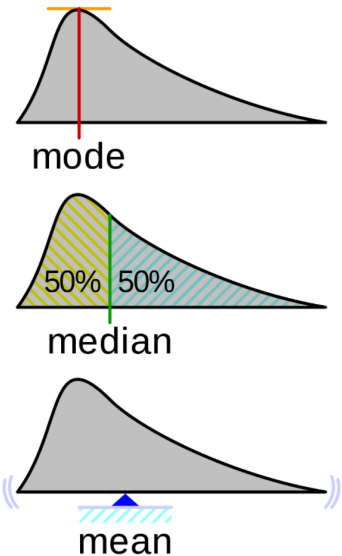
- Mean (arithmetic):
  $\bar{x} = \frac{\sum_i f_i x_i}{n} = \frac{56 + 62 + 63 + 4 \cdot 65 + 68 + 70 + 72}{10} = 65.1$ units!

- Median: 'x' that exceeds half of the measurements (sorted!).
  Remark: 'n' odd or even!
  $Me = x_{\left(\frac{n}{2}\right)} = 65.$

- Mode: Most frequent 'x'
  $Mo = 65.$

Example 2: 56, 62, 63, 65, 65, 65, 65, 68, 70, 172

Graphically



With grouped data

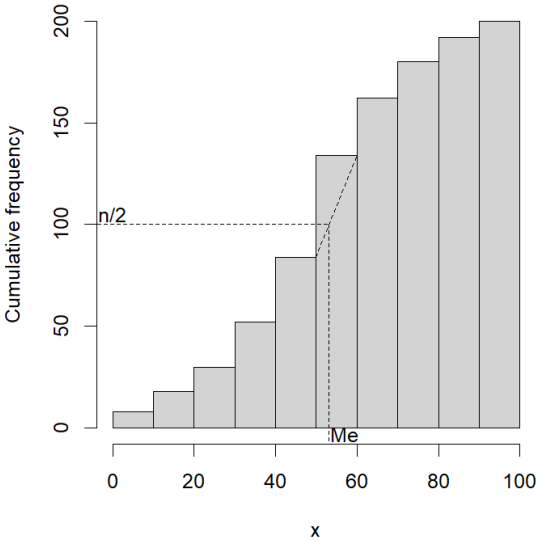
Example (cont.): Marks between 0 an 100

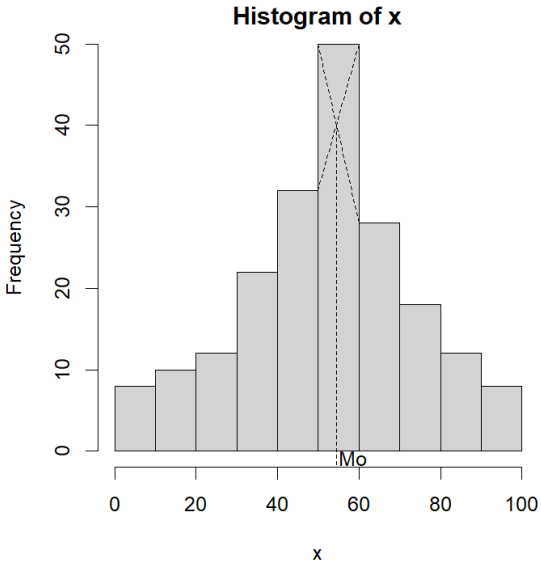
Class limits	Mark	Absolute frequency	Cumulative frequency	Percentage frequency	Cum.percent. frequency
	x_i	...			
0-10	5	8	8	4	4
10-20	15	10	18	5	...
20-30	25	12	30	...	15
...					

- Mean (arithmetic): $\bar{x} = \frac{\sum_i f_i x_i}{n} = \frac{8 \cdot 5 + 10 \cdot 15 + 12 \cdot 25 + \ldots}{200} = 52.$

- Median and Mode...

Me and Mo... Graphically!

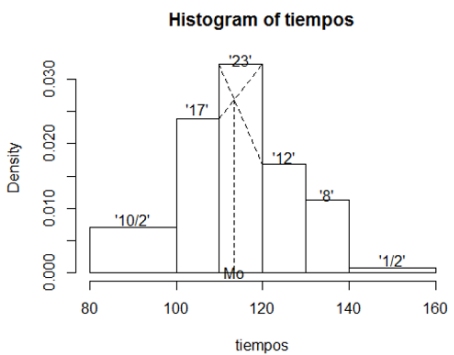


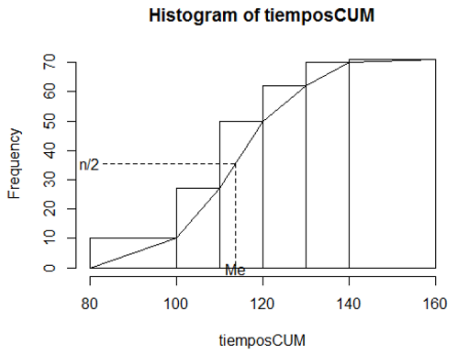


Example

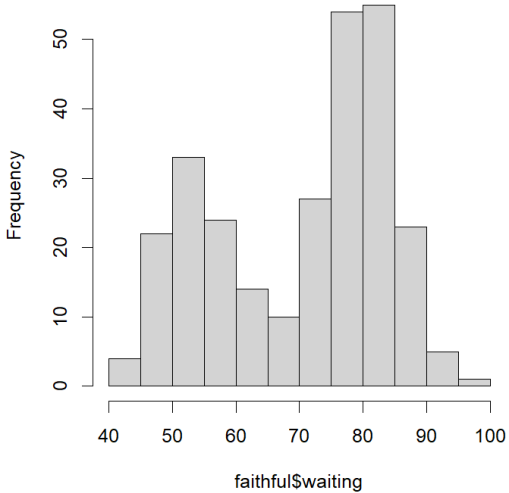
El departamento de fabricación de una empresa de rodamientos ha realizado cambios en la composición de los mismos para aumentar su resistencia. Ha realizado una serie de experimentos en condiciones extremas para observar el tiempo (en segundos) hasta la rotura de los mismos. Los resultados obtenidos son:

$[80, 100)$	$[100, 110)$	$[110, 120)$	$[120, 130)$	$[130, 140)$	$[140, 160)$
10	17	23	12	8	1



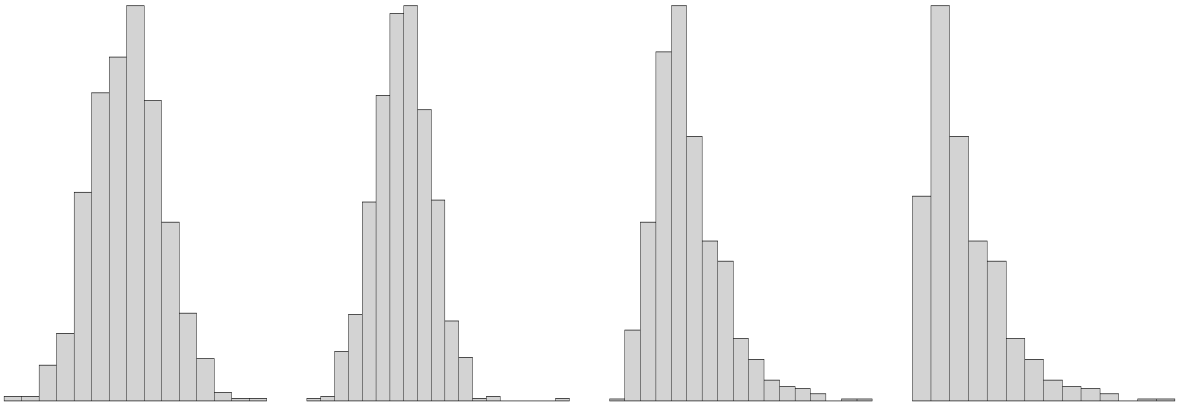


Is it representative (as central tendency)?




```
> mean(faithful$waiting)
[1] 70.89706
> median(faithful$waiting)
[1] 76
> Mo?...  Bimodal!
```

And here?



Unimodal!

Para pensar un poco

- ¿Debe coincidir la media (mediana/moda) con alguno de los valores con los que se calcula? ¿Puede ser mayor/menor que dichos valores?
- ¿Qué ocurre con la media (mediana/moda) si trasladamos todos los datos (les sumamos a todos el mismo número)? ¿Y ante un cambio de escala?
- ¿Qué ocurre con la media (mediana/moda) si a los datos le añadimos un dato más con el valor de la media?
- ¿Puede haber dos conjuntos diferentes de datos que tengan la misma media y la misma mediana?
- La media usa las magnitudes/valores de todos los datos, ¿la mediana? ¿la moda?
-

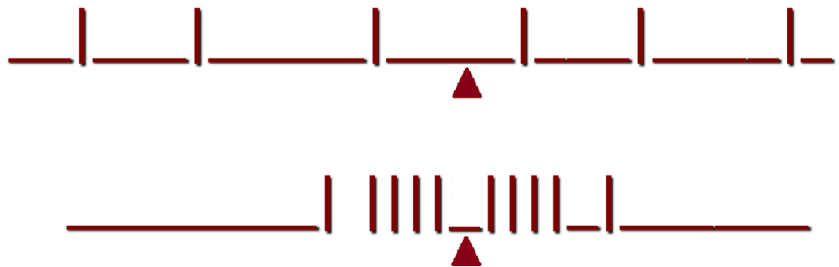
Other means

- Geometric mean... formula?...
- Harmonic mean...
- Weighted arithmetic mean:

8. CRITERIOS DE EVALUACIÓN Y VALORACIONES

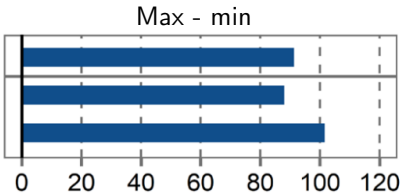
Sistema de evaluación	Evaluacion continua
Trabajo	10.00%
Prueba final	65.00%
Realización de actividades en aulas de ordenadores	25.00%
Total:	**100.00%**

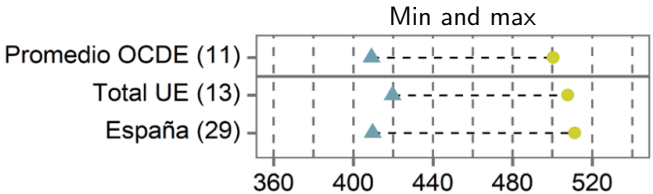
Dispersion



Range

easiest, but!





Advantages?

Disadvantages?

Variance and SD

($x_i - \bar{x}$: deviation from the mean)

- $S^2 = \dfrac{\sum_i f_i (x_i - \bar{x})^2}{n}$: Variance. (In practice) $S^2 = \overline{x^2} - \bar{x}^2$

  $S_c^2 = \dfrac{\sum_i f_i (x_i - \bar{x})^2}{n-1}$: Quasi-Variance.
  (Sample Variance -ISO-; Important for *Inference*)

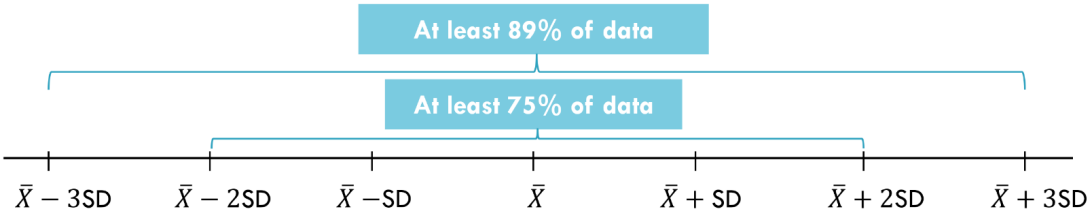
  - $S = +\sqrt{S^2}$: Standard Deviation.
    $S_c = +\sqrt{S_c^2}$: Quasi-Standard Deviation.
    (Sample Standard Deviation)

units!

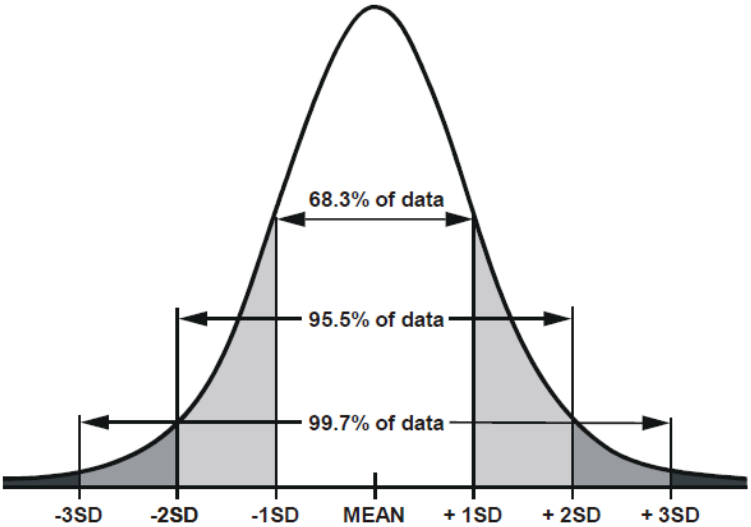
Practical meaning of the SD

Based on **Chebyshev's inequality**!

For **any data**! (*'distribution'*) ... $P(|x_i - \bar{X}| \leq k \cdot SD) \overset{(at\ least)}{\geq} 1 - \frac{1}{k^2}$



Assuming a bell shape *'distribution'*



Coefficient of variation

Relative SD

$$CV = \frac{S}{|\bar{x}|} \qquad \textit{dimensionless!}$$

- Useful to compare data sets with different means/units (PISA examples)

$$CV = \frac{10}{|100|} \quad \textit{vs} \quad CV = \frac{10}{|\ldots|}$$

- Interpretation... 'homogeneity'

Exercise

56, 62, 63, 65, 65, 65, 65, 68, 70, 72

Range?
... ?

What percentage of data is between $\bar{X}$ and 2 SD on each side?

Para pensar un poco

- ¿Debe ser el rango (desviación típica/coeficiente de variación...) siempre positivo?
- ¿Qué ocurre con el rango (desviación típica/coeficiente de variación...) si trasladamos todos los datos? ¿Y ante un cambio de escala?
- Sean los números 1, 2 y 3, cuya cuasi-desviación típica $S_{c1} = 1$. Agregando dos veces el número 2 se tiene 1, 2, 2, 2 y 3 que tendrá cuasi-desviación típica S_{c2}. ¿Será S_{c2} mayor, menor o igual que S_{c1}?
-

Sample z-score

Dimensionless variable

$$z = \frac{x - \bar{x}}{S}$$

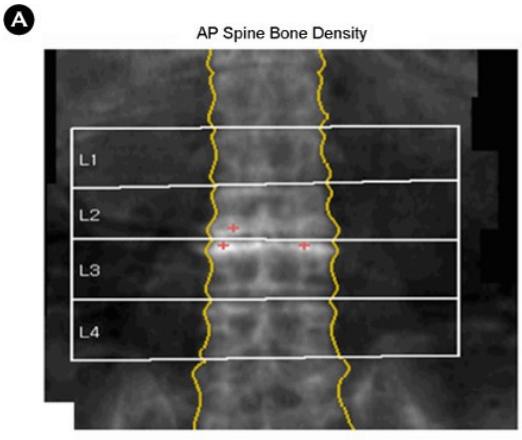
- Useful to compare data sets with different mean and/or SD...

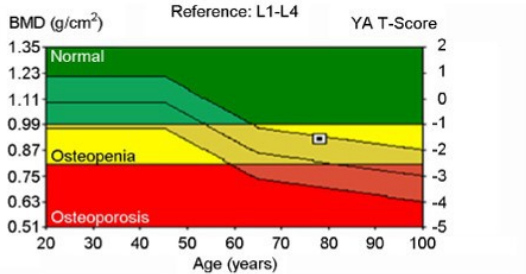
Example:

Marks Group A	vs	Marks Group B
$\bar{x}_A = 5$		$\bar{x}_B = 6$
$S_A = 1.5$		$S_B = \ldots$

Is a student with a 9.5 mark better in group A than in group B?

Medical context

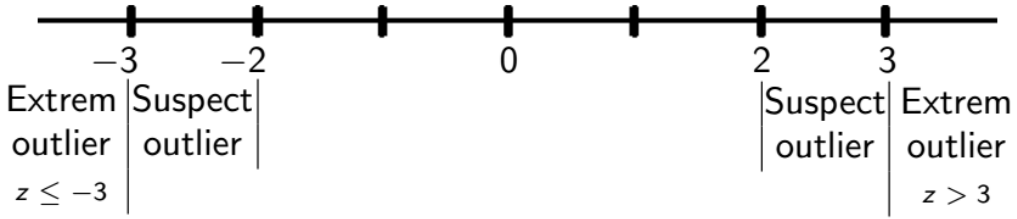




Region	[1] BMD (g/cm²)	[2] Young – Adult (%)	[2] Young – Adult T-B core	[3] Age – Matched (%)	[3] Age – Matched Z-B core
L1	0.681	64	-3.2	89	-0.7
L2	1.030	92	-0.8	125	1.7
L3	1.097	98	-0.2	133	2.3
L4	0.846	76	-2.3	102	0.2
L1-L4	0.920	83	-1.6	113	0.9

Sample z-score

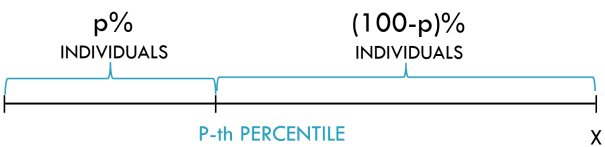
Useful for detecting outliers

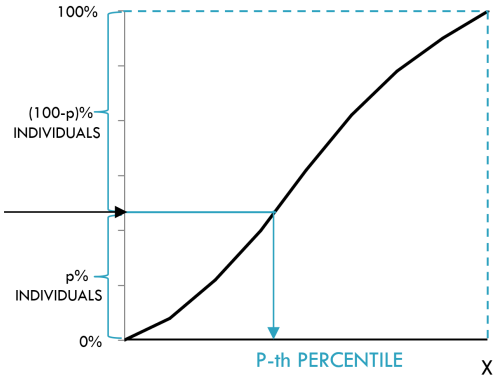


Position statistics

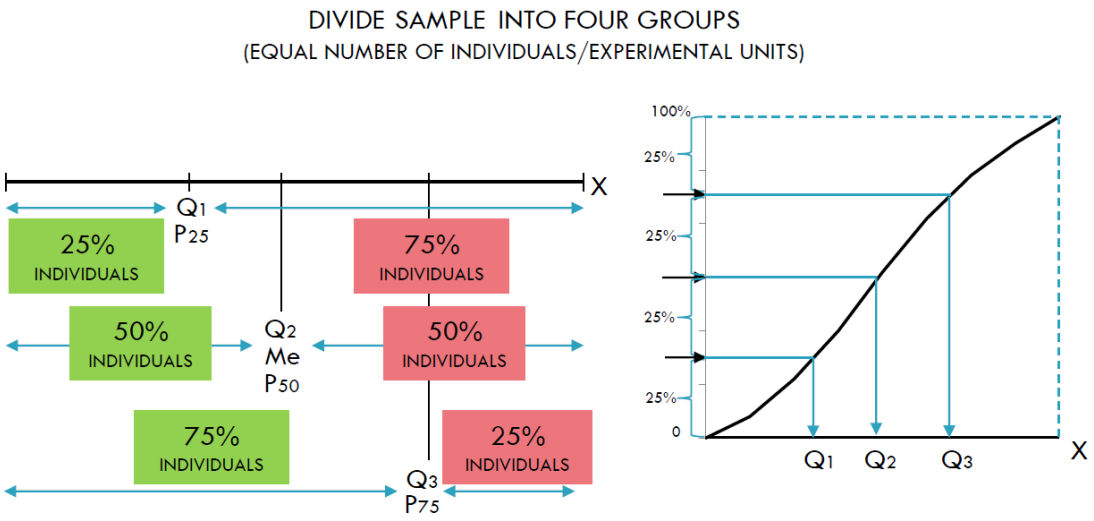
Quantile $0 < q < 1$; Percentile $0\% < p < 100\%$

x that exceeds $p\%$ of the measurements (sorted!)
(x that divides the sample into 2 groups)





Quartiles


DIVIDE SAMPLE INTO FOUR GROUPS
(EQUAL NUMBER OF INDIVIDUALS/EXPERIMENTAL UNITS)
X
Q1
P25
25%
INDIVIDUALS
75%
INDIVIDUALS
50%
INDIVIDUALS
Q2
Me
P50
50%
INDIVIDUALS
75%
INDIVIDUALS
25%
INDIVIDUALS
Q3
P75
100%
25%
25%
25%
25%
0
Q1
Q2
Q3
X

Example

1. Sort data:
   280 283 287 288 288 289 289 290 290 290 292 293 293 293
2. Find position:
   for Q_1 : $\frac{n+1}{4} = \frac{14+1}{4} = 3.75 \Rightarrow$ 280 283 $\boxed{287}$ $\boxed{288}$ 288 289 ...

   for $Q_2 \ldots$

   for $P_{30} \ldots$
3. Obtain the 'x' value (using interpolation):
   $Q_1 = x_{(\frac{n+1}{4})} = 287 + 0.75(288 - 287) = 287.75$

   $Q_2 \ldots$

   $P_{30} \ldots$

Obs: `quantile(...,type=6,...)`

From the other point of view!

$$\text{Percentile of } x = \frac{\text{number of values before } x}{n}$$

Percentile of 287?...

Inter-Quartile Range

$$IQR = Q_3 - Q_1$$

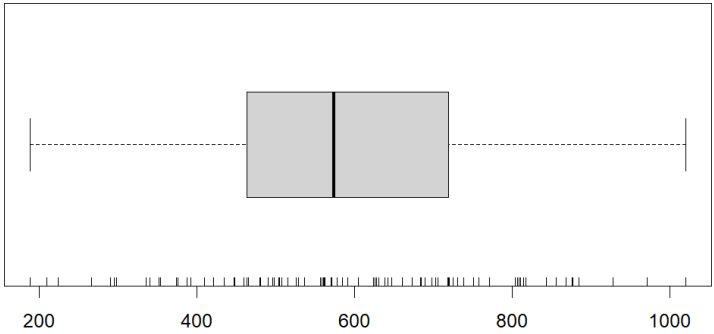
0 2 4 6 6 7 8 9

2.5 7.75

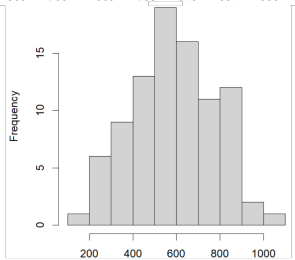
50%

5.25

Box-Plot example: repairing time after failure



876	578	718	388	562	971	698	298	673
537	642	856	376	508	529	393	354	725
811	504	807	719	464	410	491	557	771
685	448	571	189	661	877	563	647	447
336	526	624	605	496	296	628	481	224
868	804	210	421	435	291	393	605	341
352	374	267	684	685	460	570	928	516
885	751	561	1020	592	814	843	466	498
562	739	562	817	690	720	758	731	480
559	505	703	809	706	626	631	585	639

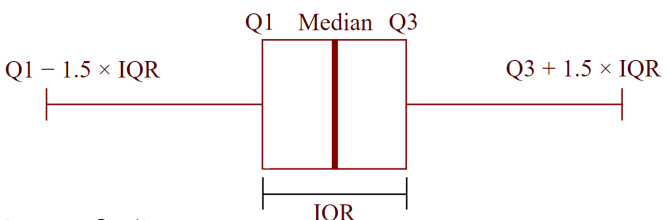



```
> summary(repair.time)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   189.0   464.5   574.5   588.6   718.8  1020.0
> IQR(repair.time)
[1] 254.25
```

Box-Plot

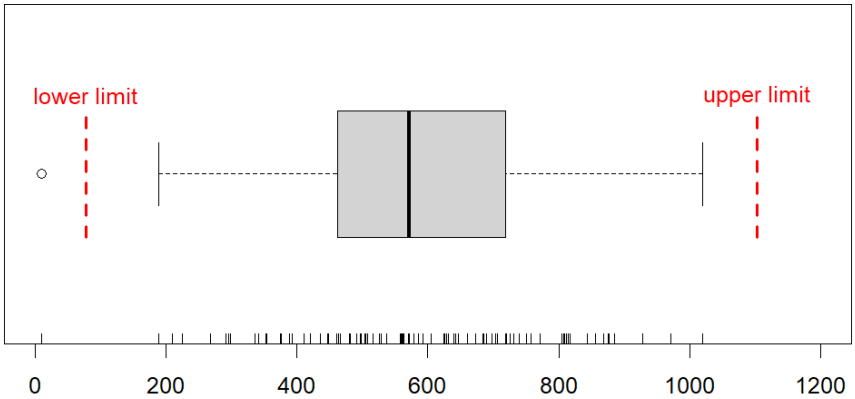
How to build it

1. Sort data.
2. Calculate quartiles: $\hookleftarrow 25\% \rightarrowtail \boxed{Q_1} \hookleftarrow 25\% \rightarrowtail \boxed{Q_2} \hookleftarrow 25\% \rightarrowtail \boxed{Q_3} \hookleftarrow 25\% \rightarrowtail$
3. Draw:
   - 3.1 A box limited by Q_1 and Q_3; straight line at the median (Q_2).
   - 3.2 Line from the box... to the smallest data value within the lower limit:
     $$min(x)|x \geq LL = Q_1 - 1.5 \times IQR$$
     " largest ... upper limit:
     $$max(x)|x \leq UL = Q_3 + 1.5 \times IQR$$

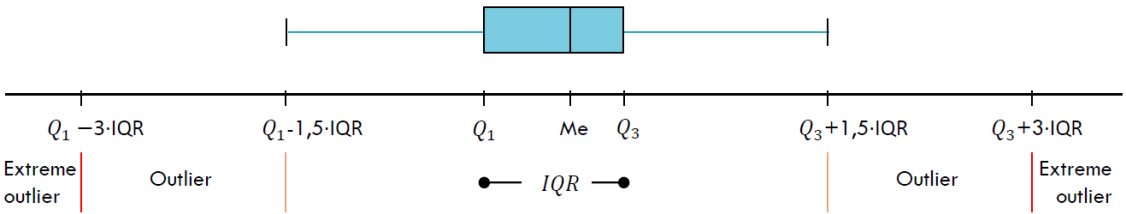


4. Out of the limits $\Longrightarrow$ Outliers.

Box-Plot example with outlier

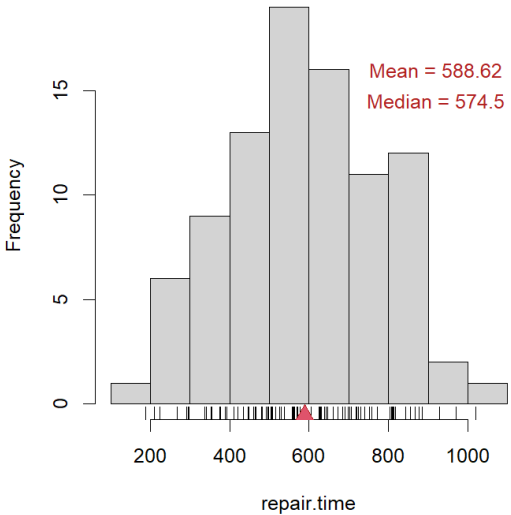


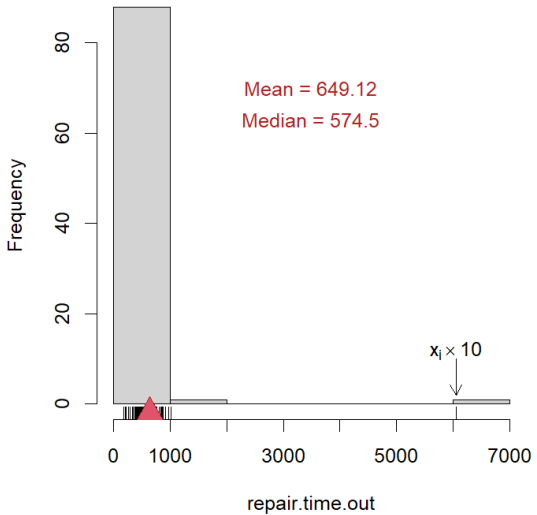
Outliers detection with Box-Plot



Outliers effects

Among others



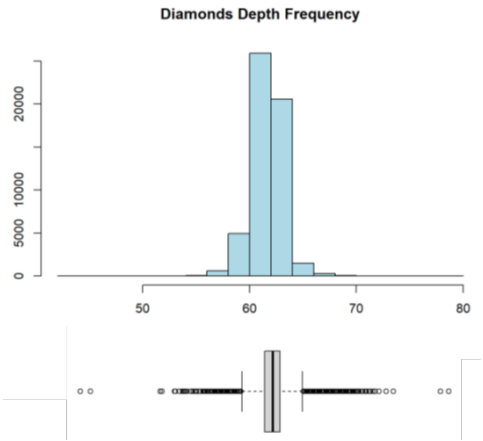


Outliers causes

Take into account for the analysis

- Error recording the observation.
  . . . Should be reviewed . . . removed

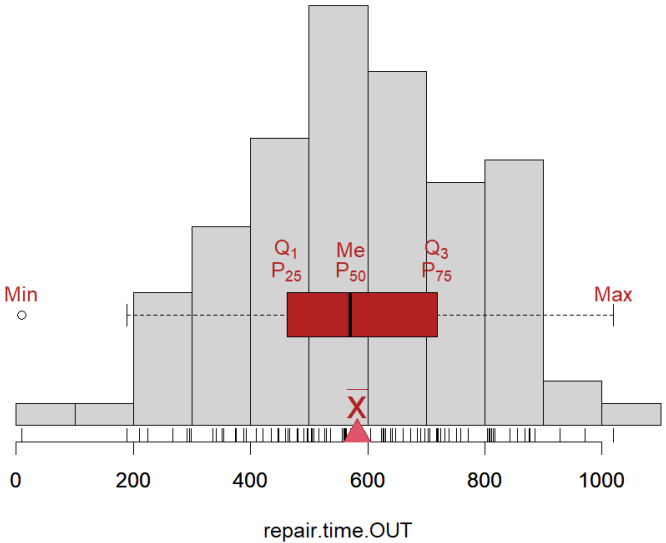
- Correct value.
  . . . Contain valuable information
  (**Examples:** Age first motorcycle, ftp).



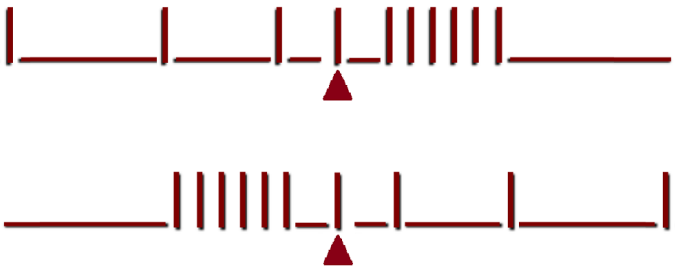
Para pensar un poco

- ¿Qué tendrías que calcular para obtener el intervalo donde se encuentra el 80% central de los datos?
- ¿Qué tendrías que calcular para obtener el intervalo, con centro la media, donde se encuentra al menos el 80% de los datos?
- ¿Qué tendrías que calcular para obtener los valores que dividen los datos en tres partes iguales?
-

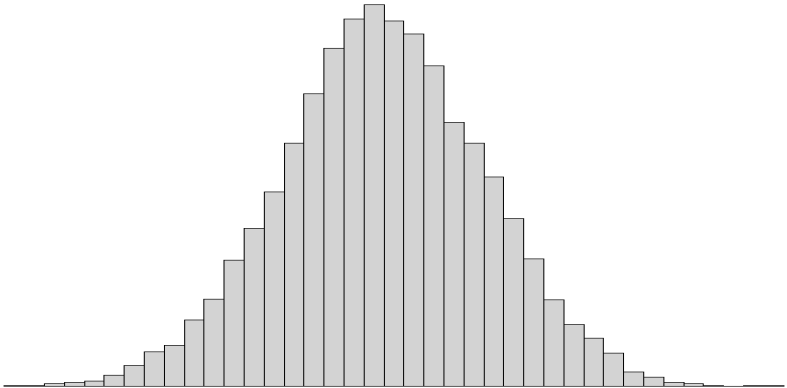
Almost all together!



Shape



Shape measures: 1. Asymmetry/Skewness

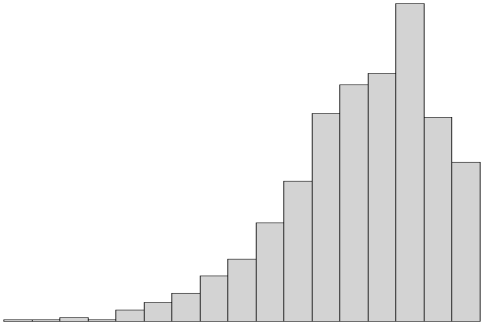


Symmetric $\Rightarrow \bar{x} = Me = Mo$

$\not\Leftarrow$

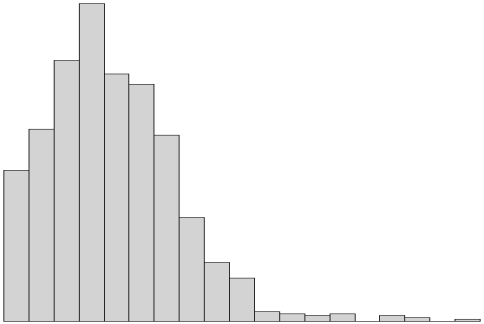
Skewed/Tail to the ...

... left



$\bar{x} < Me < Mo$

... right



$Mo < Me < \bar{x}$

Skewness coefficients

- Fisher: $g_1 = \dfrac{m_3}{S^3} = \dfrac{\sum f_i (x_i - \bar{x})^3 / n}{S^3}$

< 0	≈ 0	> 0
Tail to the left	'Symmetric' (graph!) *N(0,1)* *unimodal!*	Tail to the right

Remark: *'distribution'*

$< 0 \Leftarrow$ skewed to the left.

$= 0 \Leftarrow$ symmetric.

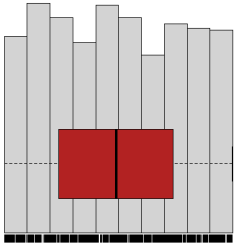
$> 0 \Leftarrow$ skewed to the right.

$\not\Rightarrow$!!

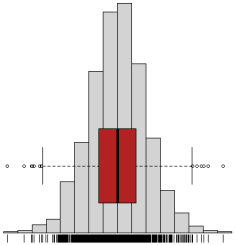
- (Others: ... Pearson 1: $\dfrac{\bar{x} - Mo}{S}$

... Pearson 2: $\dfrac{3(\bar{x} - Me)}{S}$ ...)

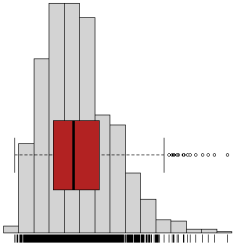
Skewness & Box-Plot



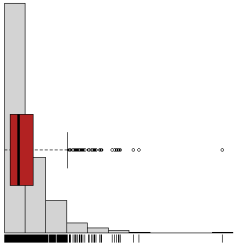
g1= 0.033



g1= -0.029



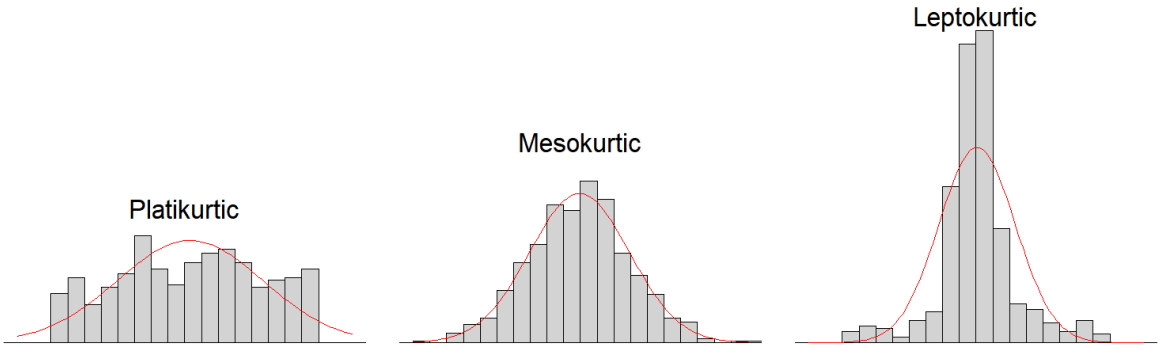
g1= 0.784



g1= 2.366

Shape measures: 2. Kurtosis

tailedness $\sim$ heaviness of the tails



Fisher's kurtosis coefficient

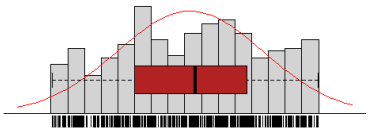
$$g_2 = \frac{m_4}{S^4} - 3 = \frac{\sum_i f_i (x_i - \bar{x})^4 / n}{S^4} - 3$$

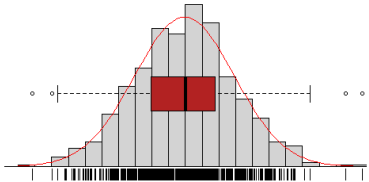
< 0	≈ 0	> 0
Platikurtic	Mesokurtic	Leptokurtic
	N(0,1)	
	unimodal!	

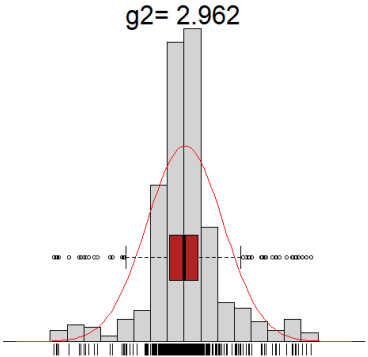
$\Leftarrow$

$\not\Rightarrow$

Kurtosis & Box-Plot







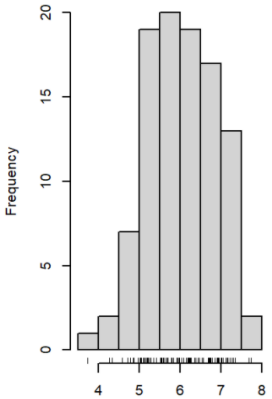
High kurtosis

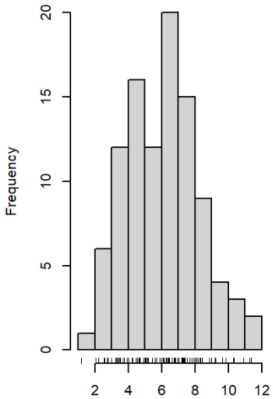
$\Downarrow$

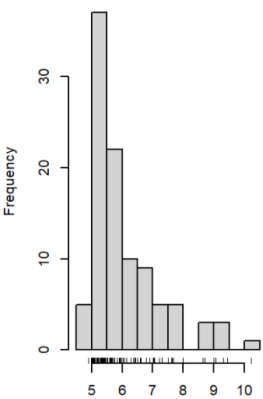
Outliers

Quiz

Measures	a)	b)	c)
Mean	6.06	6.01	6.04
SD	1.14	0.86	2.2
Skewness	1.59	-0.13	0.24







Para pensar un poco

- ¿Qué unidades tienen los coeficientes de asimetría y curtosis?
- ¿Qué indica la diferencia entre media y mediana acerca de la asimetría de los datos?
- ¿Un conjunto de datos con varianza pequeña (grande) siempre tendrá curtosis grande (pequeña)?
-

Summary - Exercise

	Continuous	Discrete	Ordinal	Nominal
Bar chart				
Histogram				
Boxplot				
Mean				
Median				
Mode				
Variance				
Std Deviation				
Percentiles				
Skewness				
Kurtosis				

$\surd$	$\times$	?

Test questions

4. El espesor medio de una muestra de planchas de material aislante es 17 mm y su desviación típica es 6 mm. Si añadimos a la muestra una nueva plancha con espesor 17.5 mm,
   a) la media disminuirá y la desviación típica aumentará.
   b) la media aumentará y la desviación típica disminuirá.
   c) la media y la desviación típica disminuirán.
   d) la media y la desviación típica aumentarán.

5. Se tiene la siguiente información: `> summary(datos)`

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	2.0	3.0	3.1	4.0	12.0

   a) Al menos hay un atípico.
   b) Viendo el rango y el valor de la media, la varianza será inferior a 3.1.
   c) Los datos son asimétricos por lo que la moda estará por debajo de la mediana.
   d) No se puede decir nada de la simetría o asimetría de los datos sin un gráfico.

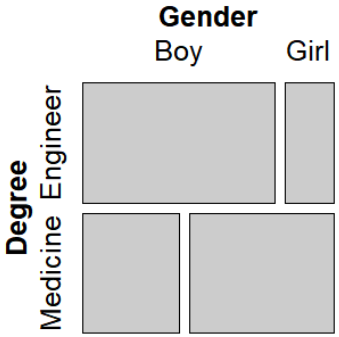
6. Dados 3 datos distintos
   a) la media será necesariamente uno de los tres datos.
   b) la media no puede ser mayor que dos de ellos.
   c) la mediana será uno de los tres datos.
   d) la moda será el dato mayor.

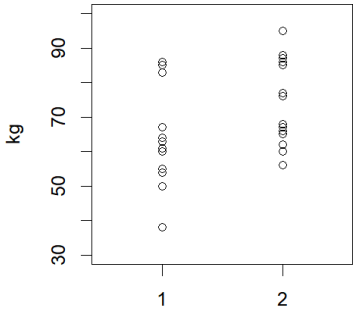
7. Si el valor del coeficiente de asimetría de una distribución es negativo, entonces:
   a) La distribución es simétrica.

BIVARIATE

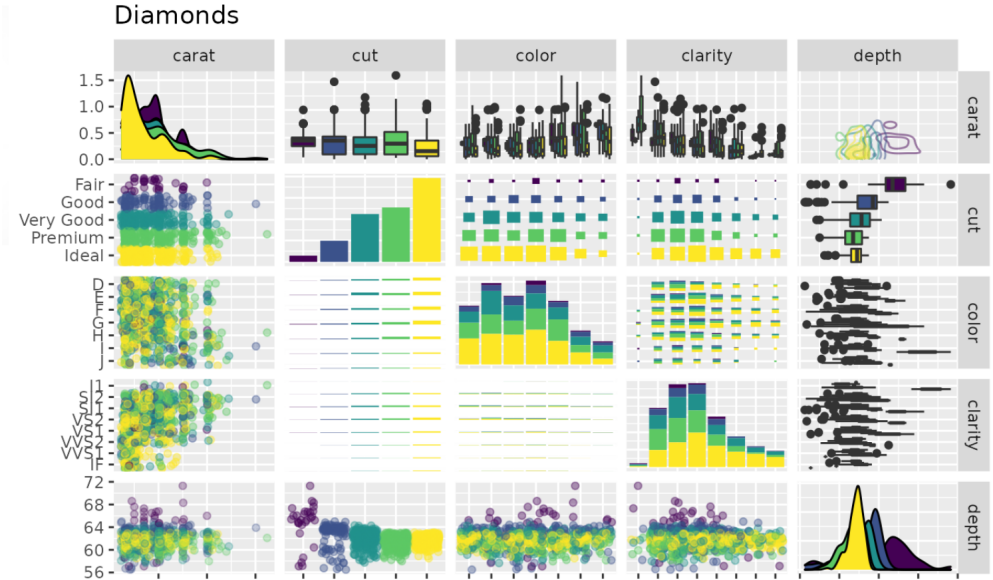
1. Categorical vs. Categorical ex: Gender vs. Degree
2. Categorical vs. Numeric ex: Gender vs. Weight
3. Numeric vs. Numeric ex: Weight vs. Height

	Boy	Girl
Engineer		
Medicine		

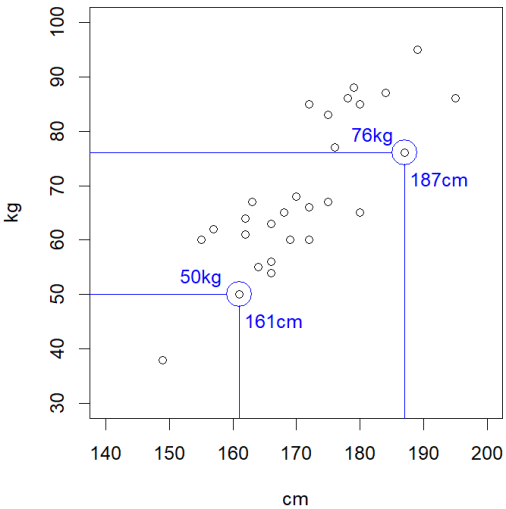




Other examples



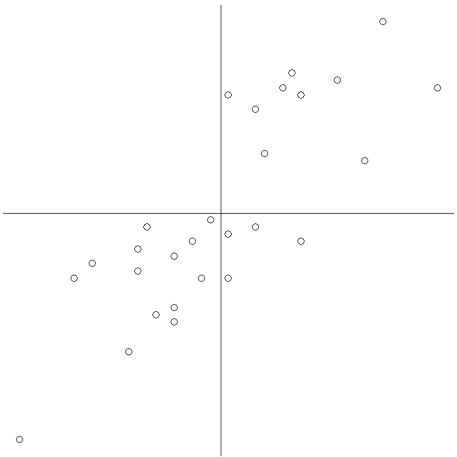
Scatterplot



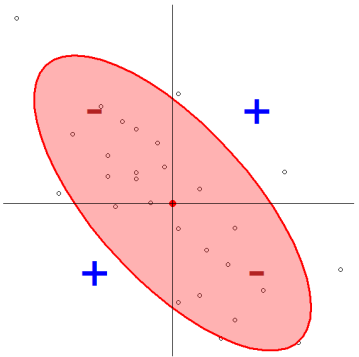
Both numeric!
Importance of the scale! $\Rightarrow \square$
Don't confuse with units!

Covariance

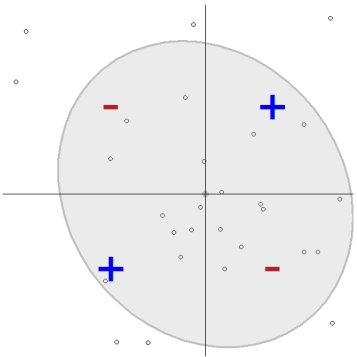
$$S_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n} \overset{\text{(in practice)}}{=} \overline{xy} - \bar{x}\bar{y}$$



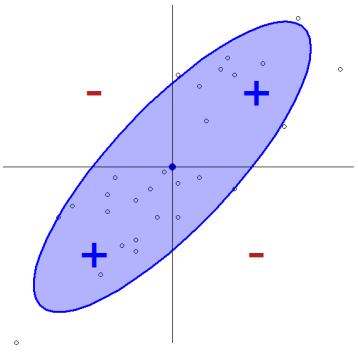
Covariance



$S_{xy} < 0$



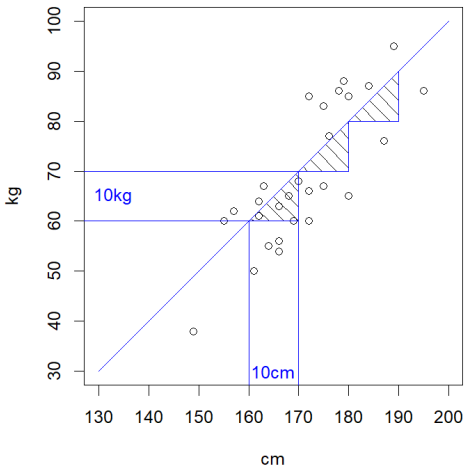
$S_{xy} \approx 0$



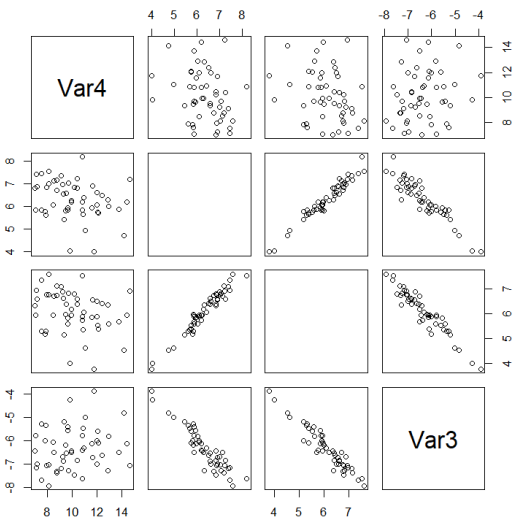
$S_{xy} > 0$

Linear relationship

Goals: Interpretation and prediction



Examples



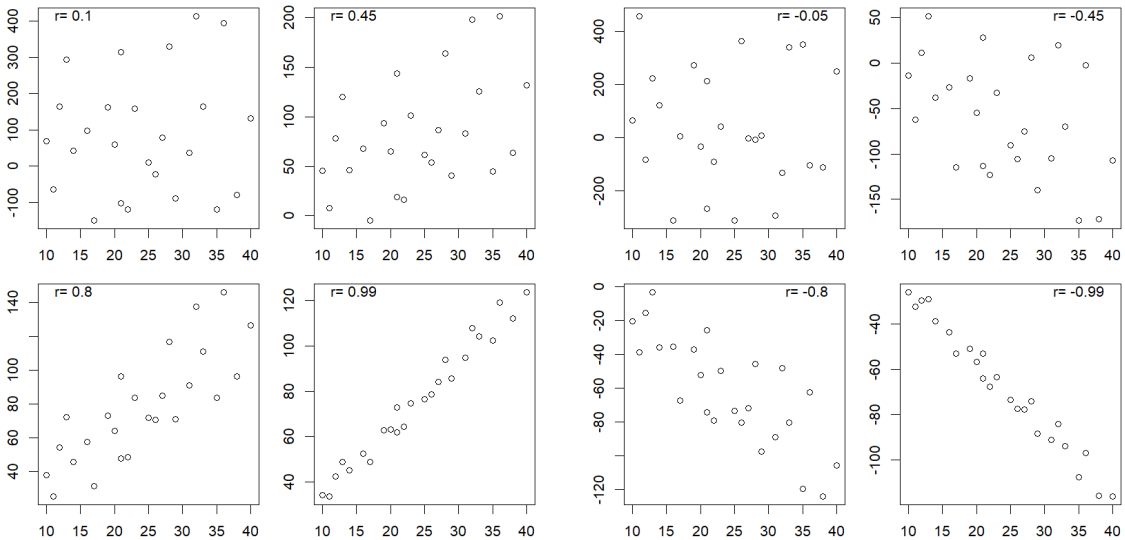
Pearson's linear correlation coefficient

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \in [-1, 1] \text{ (No units!)} \begin{cases} > 0 & \Leftarrow \text{positive linear correlation} \\ < 0 & \Leftarrow \text{negative linear correlation} \\ = 0 & \Leftarrow \text{linear uncorrelation (may be others)} \\ = \begin{cases} -1 \\ 1 \end{cases} & \Leftarrow \text{perfect linear correlation} \begin{cases} \textit{reverse} \\ \textit{direct} \end{cases} \\ & \not\Rightarrow \end{cases}$$

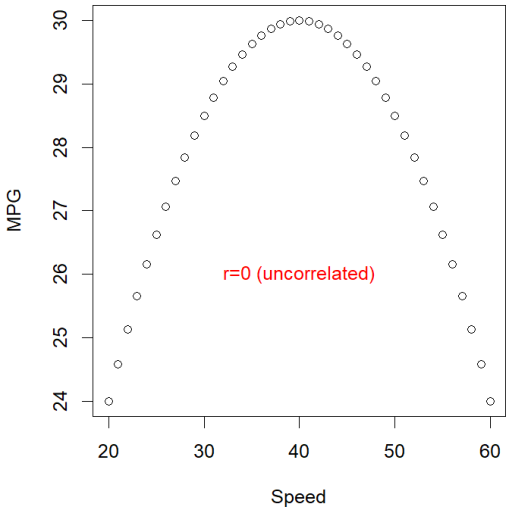
A	>	`cor(Var1,Var3)`		`0.9734758`
B	>	`cor(Var1,Var2)`		`-0.2718557`
C	>	`cor(Var1,Var4)`		`-0.9519246`

Spearman's corr. coef: ...

Exercising (Thank you Mercedes)



r does not measure nonlinear correlation



Speed and MPG linearly independent!
Speed and MPG NOT independent!

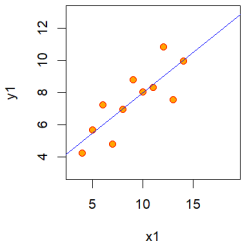
Beware!

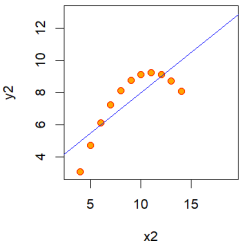
Anscombe's Quartet

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Four data sets where $r_{xy} = 0.816$!!

Expected



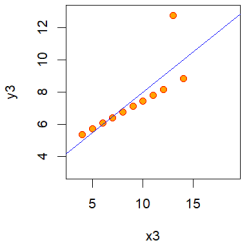


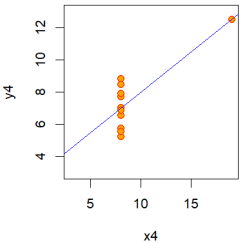
Non linear!

Outlier!

⇓

↓ r_{xy} from 1



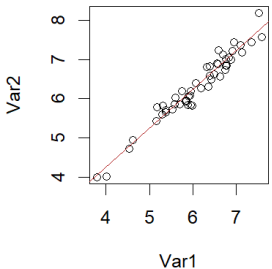


Outlier!

⇓

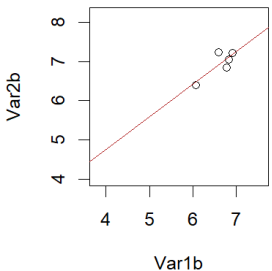
↑ r_{xy} from 0

Correlation and ‘p-value’: examples




```
Pearson's product-moment correlation

data:  Var1 and Var2
t = 29.479, df = 48, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9534903 0.9849396
sample estimates:
      cor
0.9734758
```




```
Pearson's product-moment correlation

data:  Var1b and Var2b
t = 2.4033, df = 3, p-value = 0.09559
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2497737  0.9870495
sample estimates:
      cor
0.8112698
```

Correlation and ‘p-value’: comments

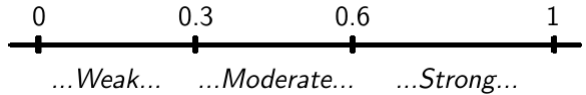
- First **significant** (important) correlation

  The lower the **p-value**, the significant the relationship will be.
  (hypothesis test
  -inference-)

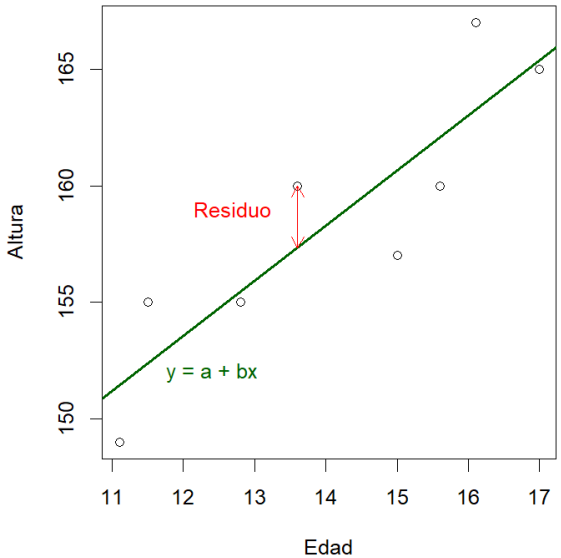
- Then **magnitude**

  The higher the correlation, the strong the relationship will be.

Orientative values:



Regression: Least Squares method



Simple Linear Regression Model

$$y = a + bx + e$$

Remarks:

- x and y are not exchangeable.
  both numeric!
- The model forces a type of line.
- Prediction beyond the range of observation is possible (carefully).

Least Squares Estimators (LSE)

LS problem statement: $\min_{a,b} \sum_i [y_i - (a + bx_i)]^2$

Normal equations:
$$\begin{cases} \bar{y} = a + b\bar{x} \\ \overline{xy} = a\bar{x} + b\overline{x^2} \end{cases}$$

LSE:
$$\begin{cases} \boxed{\hat{b} = \dfrac{S_{xy}}{S_x^2}} \\ \boxed{\hat{a} = \bar{y} - \hat{b}\bar{x}} \end{cases}$$

Regression line: $y = \hat{a} + \hat{b}x \iff y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$

Example/Exercise

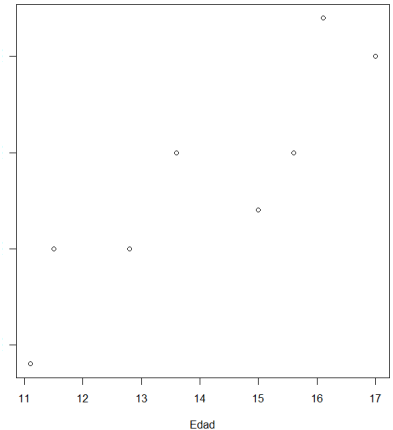
y: points	10	4	15	20	10
x: player's height	190	194	200	208	195

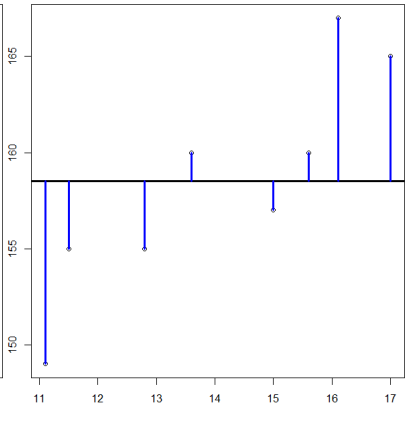
```
                Estimate Std. Error   t value   Pr(>|t|)
(Intercept) -132.1202929 54.1786367 -2.438605 0.09261545
height         0.7290795  0.2743266  2.657706 0.07648835
```

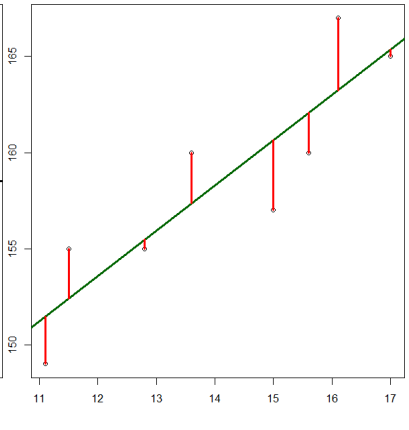
Interpretation?
Goodness-of-fit?

Goodness-of-fit, first view (Thank you Licesio!)

Example: Edad vs Altura







Descomposition of variability

$$\sum_i (y_i - \bar{y})^2 = \sum_i [(\hat{a} + \hat{b}x_i) - \bar{y}]^2 + \sum_i [y_i - (\hat{a} + \hat{b}x_i)]^2$$

$$VT = VE + VNE$$

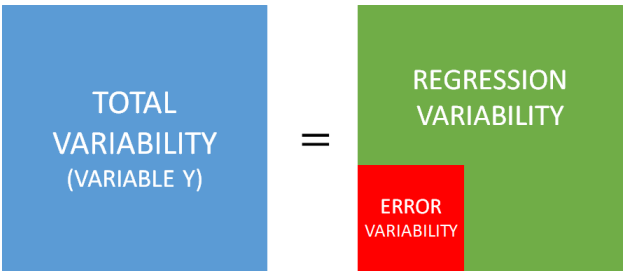
Total Variab. = Variab. Explained + Variab. Not Explained

Reg. Variab.

Error Variab.: $\sum_i e_i^2$

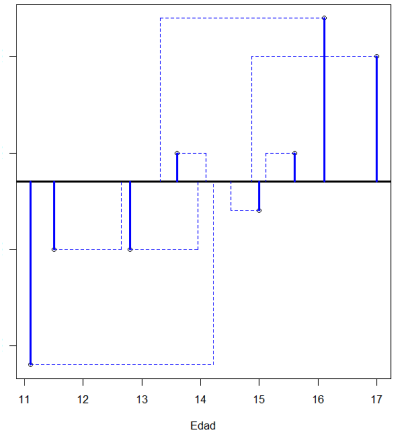
↑

Residuals

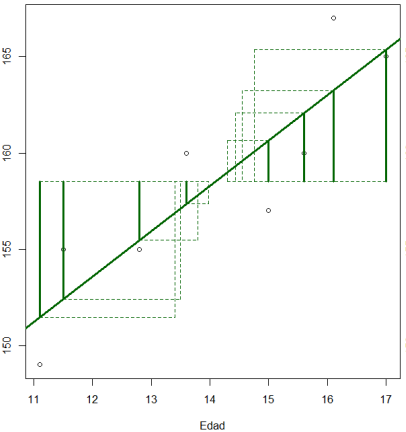


The lower Error Variab (VNE) ⇒ the higher Reg Variab (VE) ⇒ the better goodness-of-fit (R^2)

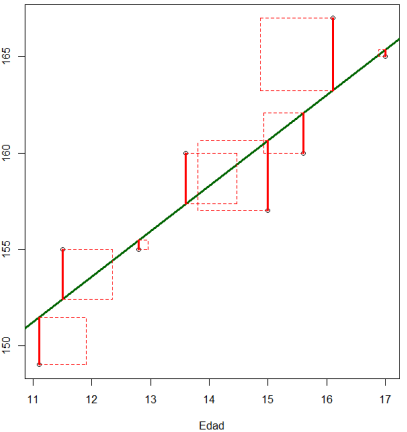
Grafically



V...
236 (units^2)



V...
184.132...



V...
51.868...

R^2: Determination coefficient

Goodness-of-fit measure

$$R^2 = \frac{VT - VNE}{VT} = \frac{VE}{VT}$$

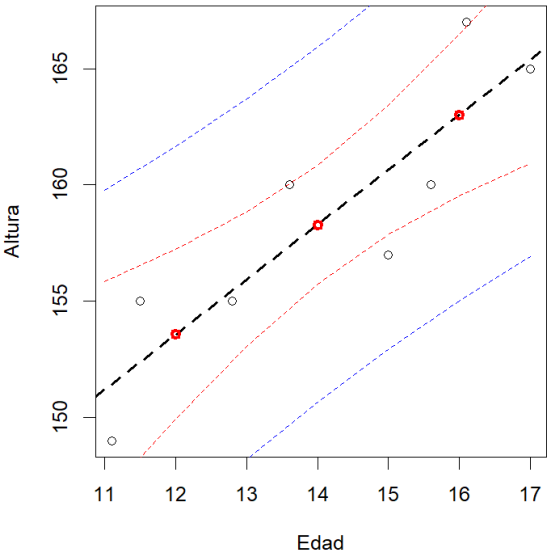
Variability proportion explained by the regression model

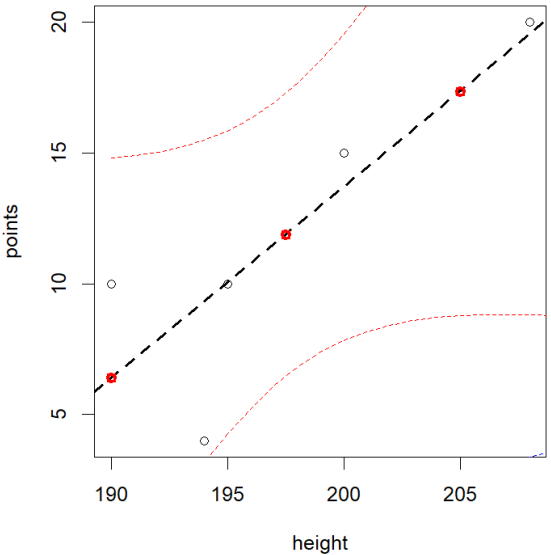
In practice (for simple linear regression!):

$$R^2 = r_{xy}^2$$

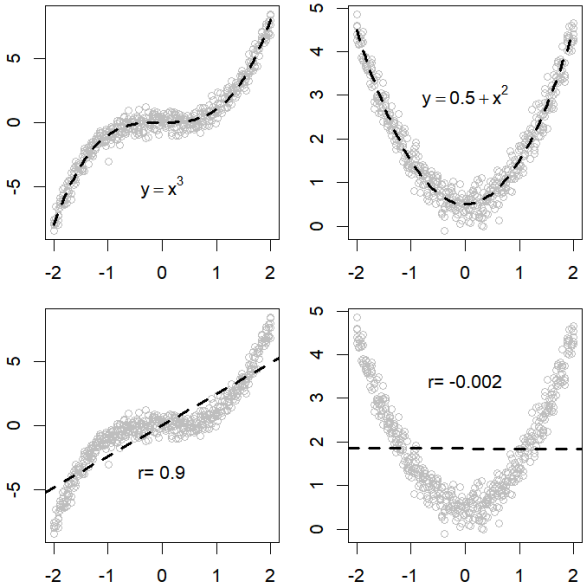
Altura vs. Edad example: $R^2 = 78.02\%$
Basket exercise:....

Prediction (examples)

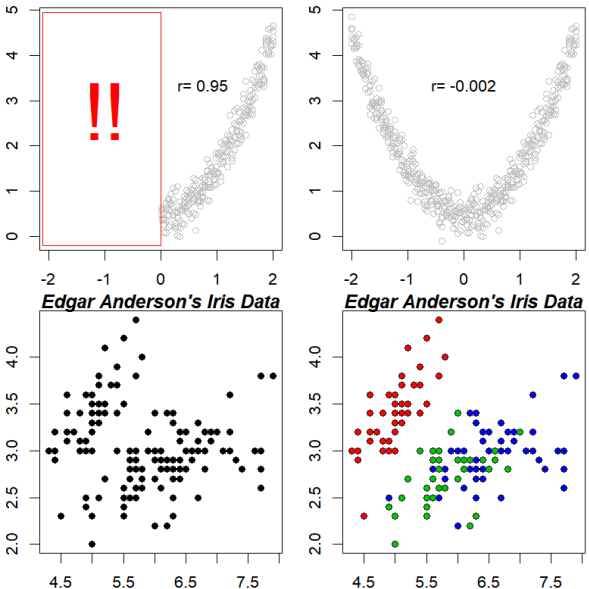




Nonlinear vs Linear regression



Be careful with scatterplots!

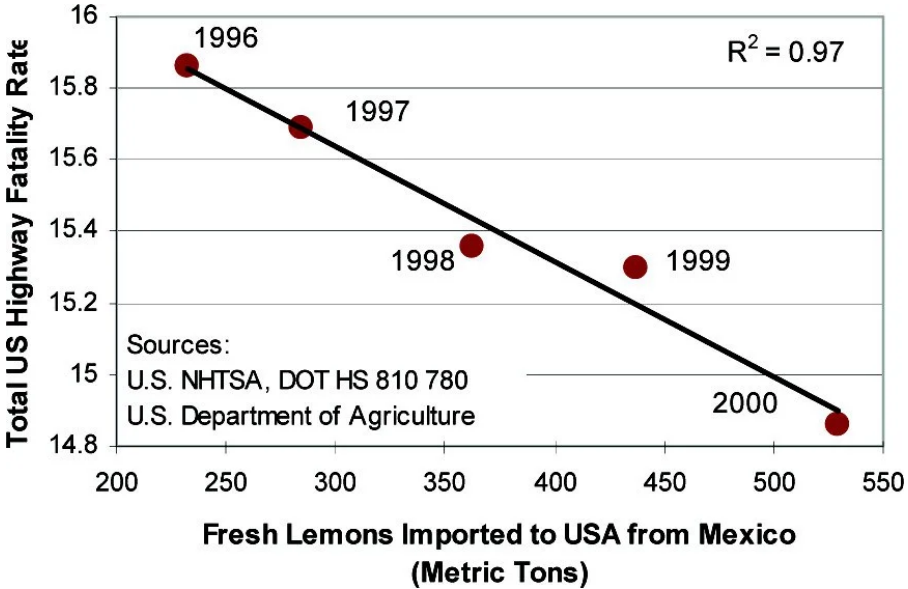


Test questions

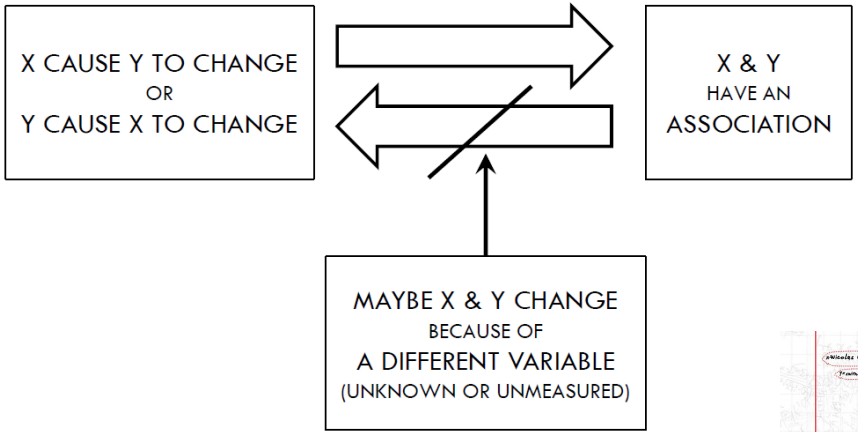
8. Dada la recta de regresión $y = 2{,}5 + 0{,}8x$, y sabiendo que $S_x = 1{,}1S_y$ y $x \in [10.3,\ 18.2]$,
   a) el coeficiente de correlación es 0.88.
   b) el coeficiente de determinación es 0.8.
   c) no se puede saber el coeficiente de correlación.
   d) los valores de $y \in [10.74,\ 17.06]$.

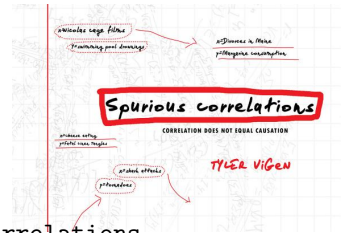
9. Si el valor del coeficiente de determinación (R^2) de una recta de regresión es -0.95, entonces
   a) La pendiente de la recta será positiva.
   b) El coeficiente de correlación tendrá signo positivo.
   c) El coeficiente de determinación sólo puede tomar valores entre 0 y 1.
   d) La pendiente de la recta será negativa.

Causality!


Total US Highway Fatality Rate
16
15.8
15.6
15.4
15.2
15
14.8
1996
1997
1998
1999
2000
$R^2 = 0.97$
Sources:
U.S. NHTSA, DOT HS 810 780
U.S. Department of Agriculture
200
250
300
350
400
450
500
550
Fresh Lemons Imported to USA from Mexico
(Metric Tons)

Association/Relationship





Sergio's examples: `https://www.tylervigen.com/spurious-correlations`