Campus Irapuato-Salamanca | División de Ingenierías

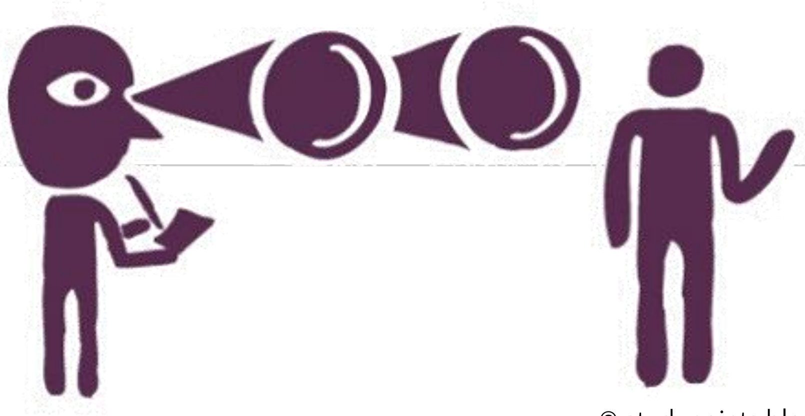
Data Mining

© copyrightuser.org

Topic 3: Analysis of structured data

August – December 2021

Juan Carlos Gómez
PhD in Computer Science
jc.gomez@ugto.mx
http://jcgcarranza.wix.com/juancarlosgomez
Office 314



© studypoints.blogspot.mx

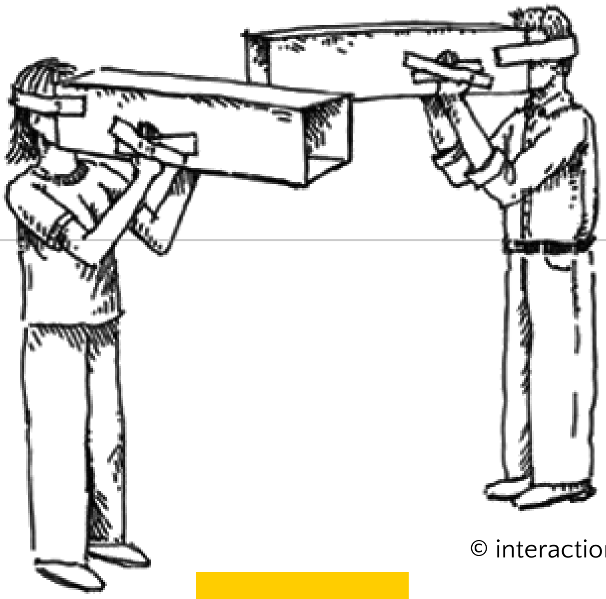
Observation

Information collected about an **object** of **interest**: a person, a business, a football game, an event, a period of time, etc.

© studypoints.blogspot.mx

Observer

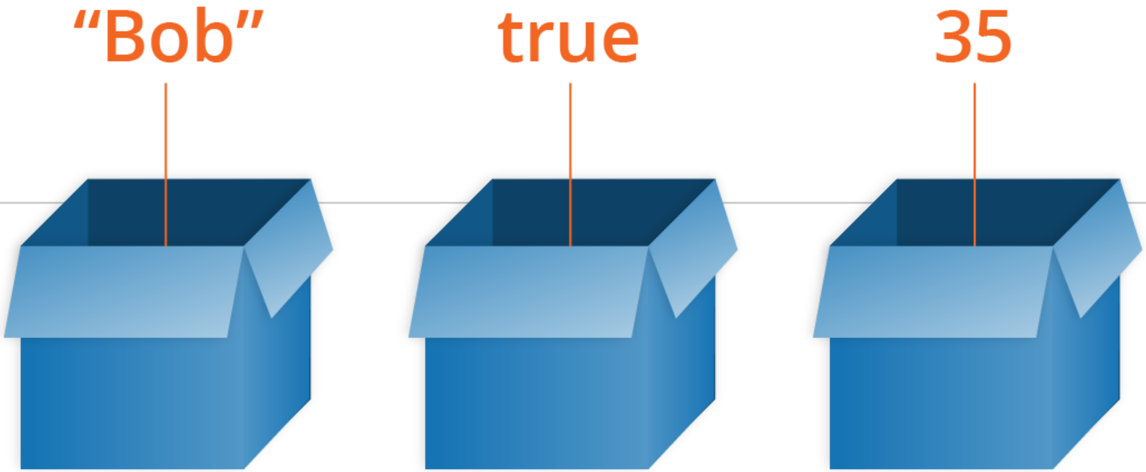
Someone who gathers information about an observed object but **does not intervene**.



© interaction-design.org

Bias

Inclination to present or hold a **partial perspective**. Due to many causes (social, cultural, economical, etc.). Implies a lack of a **neutral** viewpoint.



Variables

Record the measurements on which we are interested about **observations** (objects): age, sex, pet, chocolate preference, goals scored, etc.

Observations and variables

Variables

Observations

Name	Age	Sex	Chocolate Preference
John	18	M	Milk
Anna	45	F	Dark
Jenna	24	F	White

Single observation

Single variable

Types of variables

Nominal

Categorical or qualitative

e.g., Sex, color, chocolate preference

Ordinal

Order matters

e.g., Rank, satisfaction

Interval/ratio

Variables that can be measured rather than classified: scale, **quantitative**, parametric

e.g., weight, age, size,

Nominal variables

- Named wit labels/names but also with **codes/indexes**

(1) Red
(2) Blue
(3) Yellow
(4) White
(5) Black

Numbers **do not** have an order.

Ordinal variables

- Named wit labels/names but also with **codes/indexes**

(1) Very satisfied
(2) Satisfied
(3) Dissatisfied
(4) Very dissatisfied

Numbers **have** an order

Nominal and ordinal data

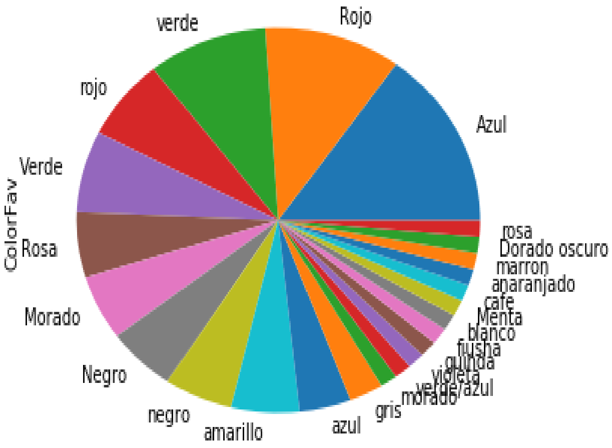
- We can estimate frequencies/percentages:

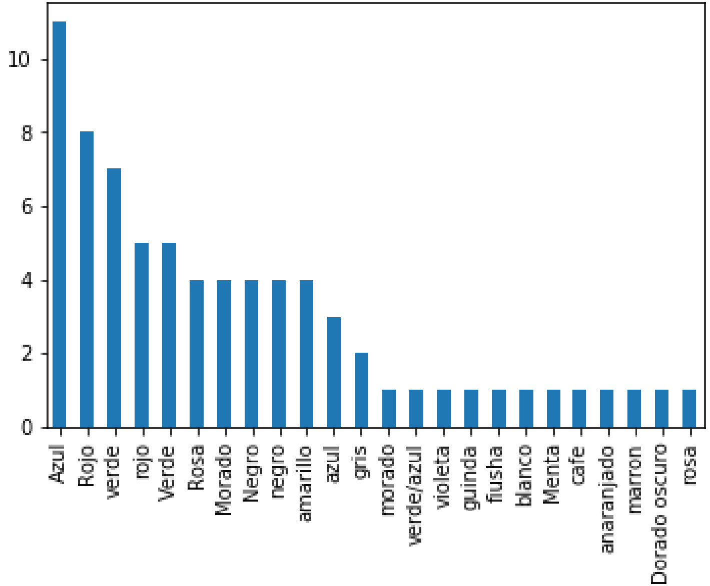
Red: 50 = 30%
Blue: 50 = 30%
Yellow: 15 = 10%
White: 20 = 20%
Black: 15 = 10%

Very satisfied: 20 = 20%
Satisfied: 45 = 45%
Dissatisfied: 20 = 20%
Very dissatisfied: 15 = 15%

Graphical representation of nominal data

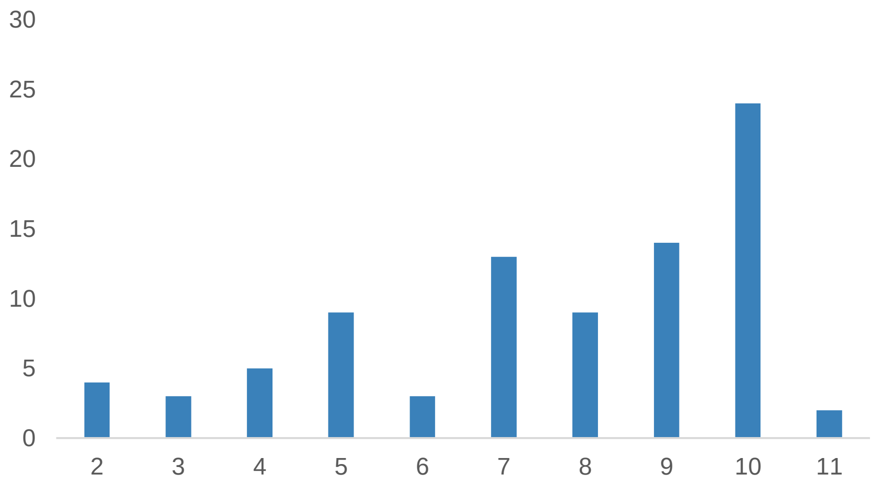
- **Bar/pie** charts





Graphical representation of ordinal data

- Bar chart



Ordinal data

- Sometimes the mean is useful. But be careful (**not recommended**) :

(1) Very satisfied: 20 = 20%
(2) Satisfied: 45 = 45%
(3) Dissatisfied: 20 = 20%
(4) Very dissatisfied: 15 = 15%

Mean = 2.3 (more satisfaction than dissatisfaction)

Interval/ratio variables

- Represent a **physical** attribute or a **quantity**. Something that can be measured.

Age
Length
Weight
Area
Sales
Interest
...

Interval **variables**

- How to understand this **data**?

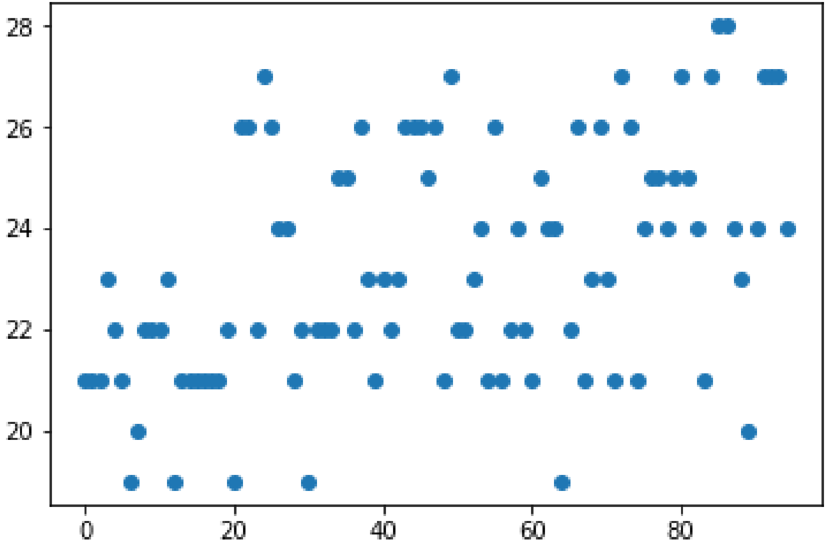
Age:

21	23	19	19	23	22	21	23
21	19	26	22	22	22	25	21
21	21	26	22	23	23	24	27
22	21	22	22	26	24	24	26
21	21	27	25	26	21	19	21
19	21	26	25	26	26	22	24
20	21	24	22	25	21	26	25
22	21	24	26	26	22	21	25
22	22	21	23	21	24	23	24
22	19	22	21	27	22	26	25

Interval variables

- First attempt: **scatter** chart

Age:



Interval variables

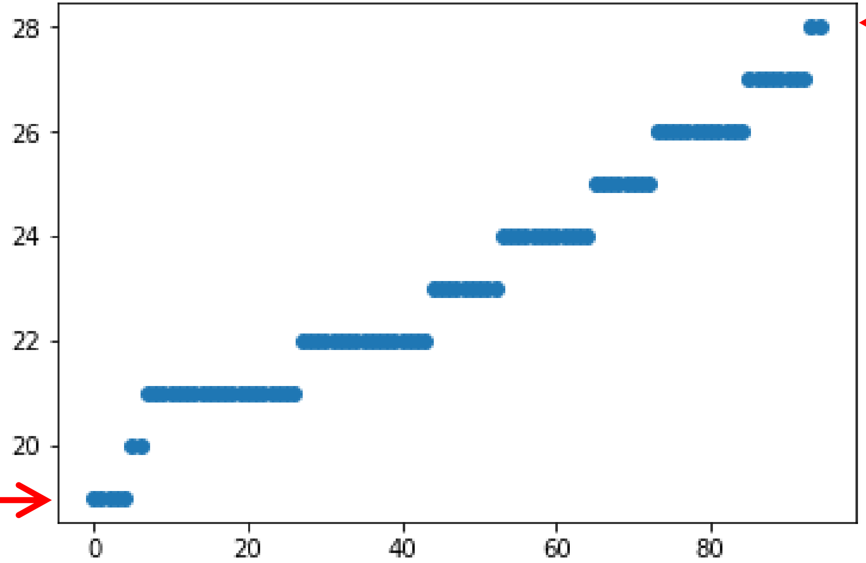
- Second attempt: ordered **scatter** chart

Age:

Rest of the ages

Max: 28

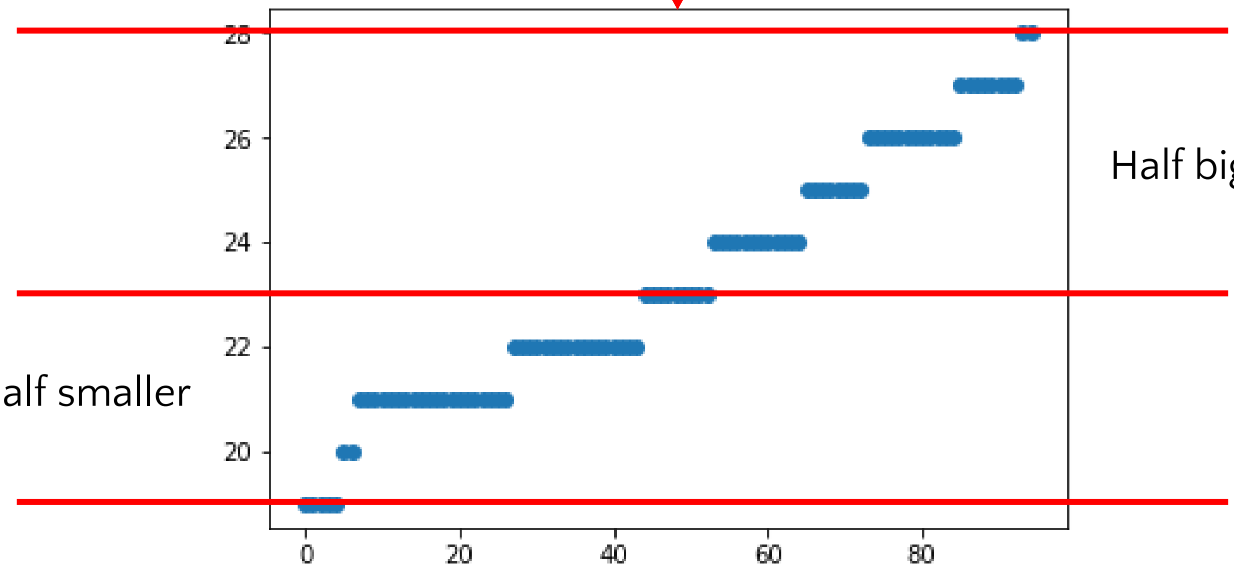
Min: 19



Interval variables

Half of the data: 23 = **Median**

Age:



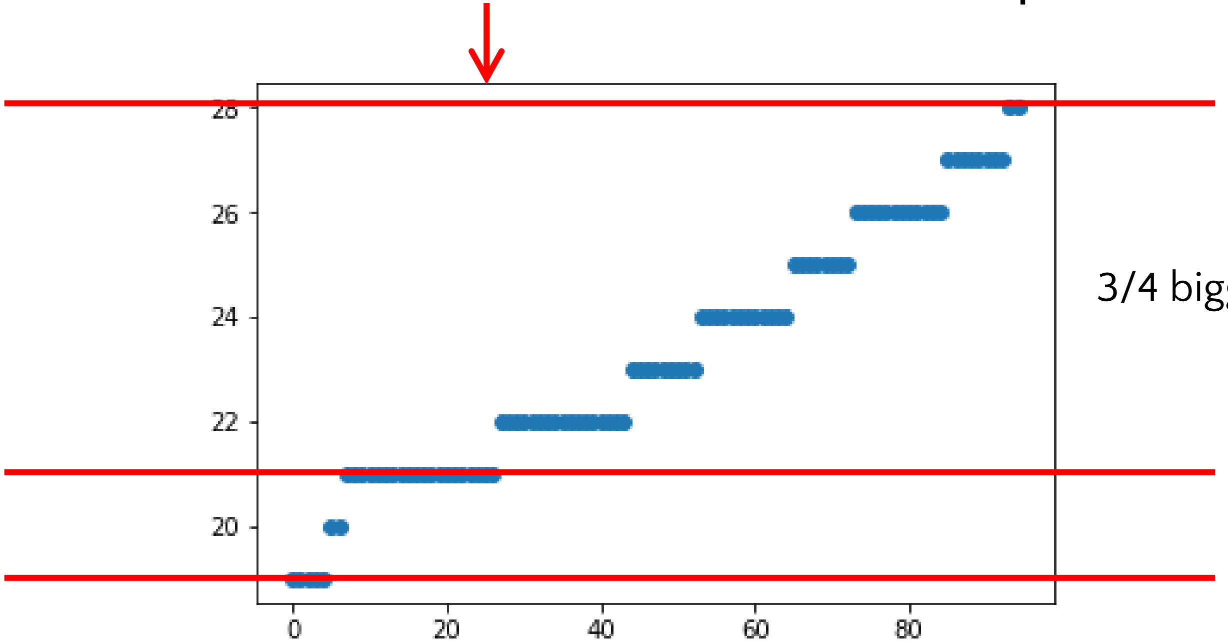
Half bigger

Half smaller

Interval variables

Quarter of the data = 21 = **1st quartile**

Age:



3/4 bigger

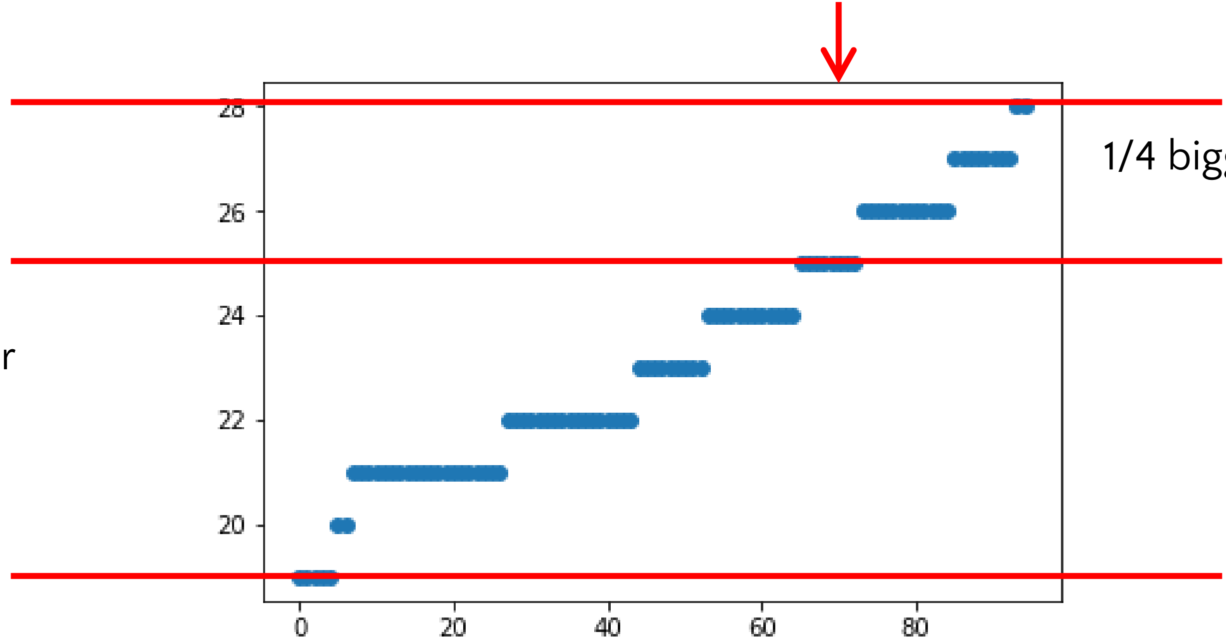
1/4 smaller

Interval variables

3/4 Quarter of the data = 25 = **3rd quartile**

Age:

3/4 smaller

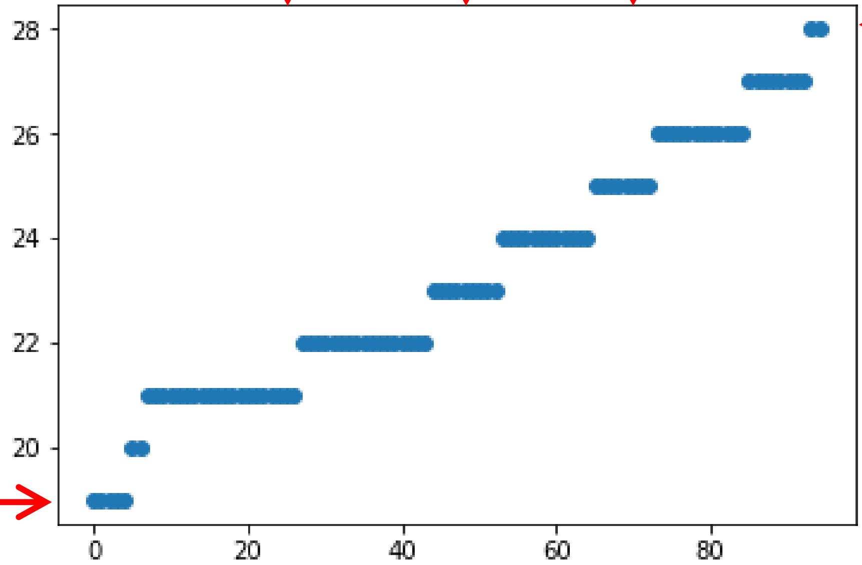


1/4 bigger

Interval variables

1st quartile = 21 Median = 23 3rd quartile = 25

Age:



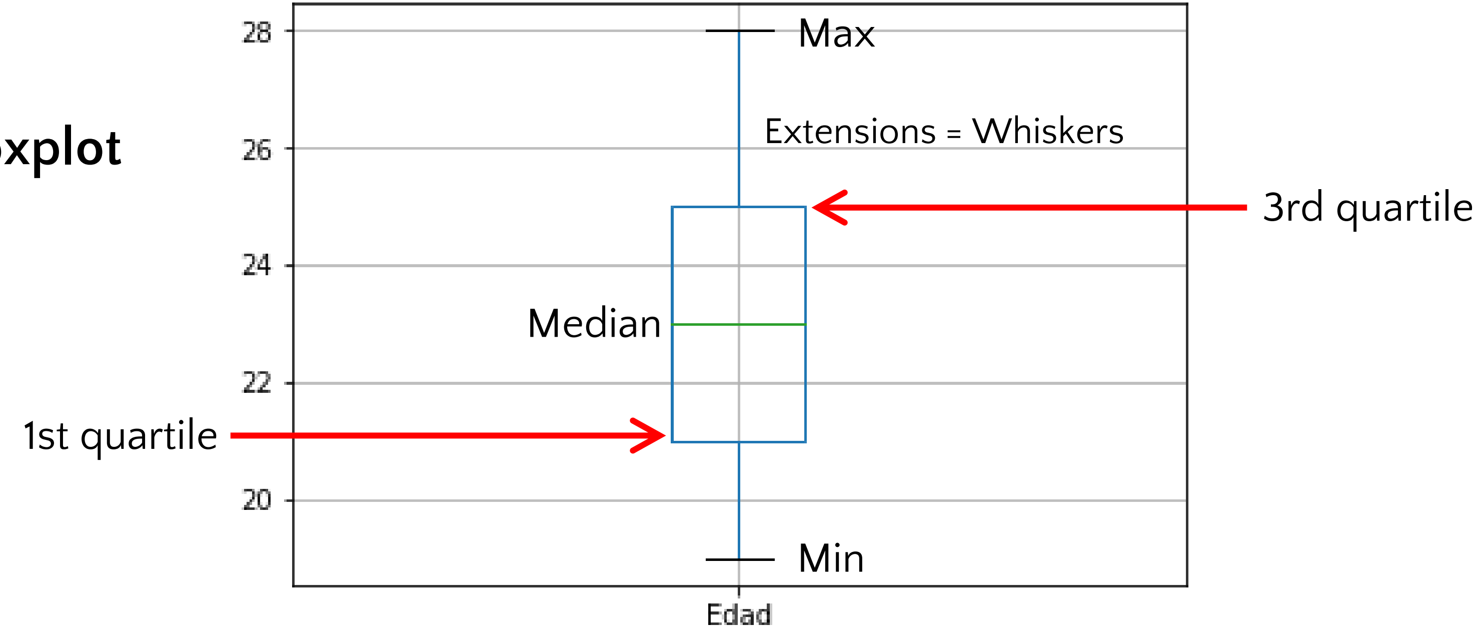
Max: 28

Min: 19

Five number summary

Interval variables

Boxplot



Interval variables

- Example. Grades. Extract the **five number** summary

79, 68, 88, 69, 90, 74, 87, 93, 76

Interval variables

- Example. 1st rearrange

Min

68, 69, 74, 76, 79, 87, 88, 90, 93

Max

Interval variables

- Example. Median. The one in the middle: there are 9 numbers, the one in the middle is the fourth (counting from 0).

Min

68, 69, 74, 76, 79, 87, 88, 90,93

Number of elements

Median

Max

Percentile

p = (9-1)*0.5 = 4

pl = floor(4) = 4

pu = ceil(4) = 4

Median position

Interval variables

- Example. 1st quartile. The one that is one quarter away from the first grade: the second (counting from 0).

1st quartile

Min

68, 69, 74, 76, 79, 87, 88, 90,93

Median

Max

Number of elements

Percentile

p = (9–1)*0.25 = 2
pl = floor(2) = 2
pu = ceil(2) = 2

1st quartile position

Interval variables

- Example. 3rd quartile. The one that is three quarters away from the first grade: the sixth (counting from 0).

Min 1st quartile 3rd quartile

68, 69, 74, 76, 79, 87, 88, 90,93

Median Max

Percentile

Number of elements

p = (9–1)*0.75 = 6

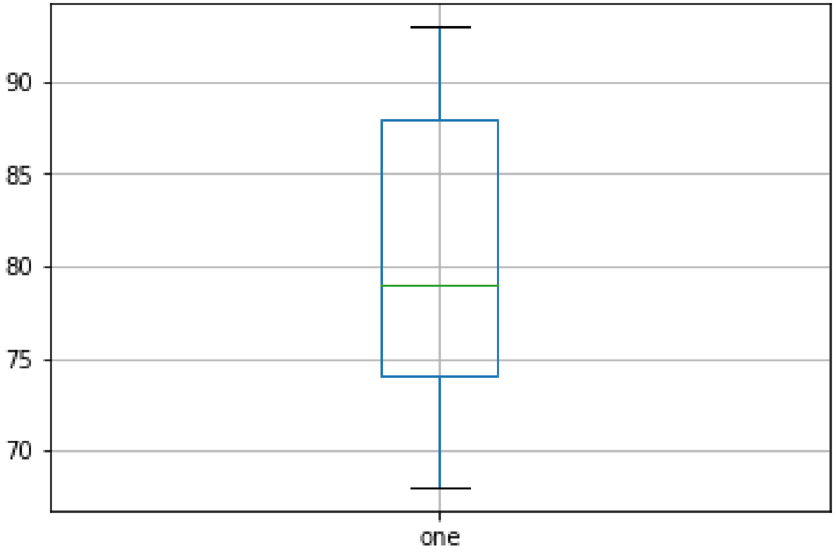
pl = floor(6) = 6

pu = ceil(6) = 6

3rd quartile position

Interval variables

- Example. Boxplot



Interval variables

- 2nd Example. Grades. Extract the **five number** summary

79, 93, 68, 84, 90, 74

Rearrange

Min

68, 74, 78, 84, 90, 93

Max

Median?

Interval variables

- 2nd Example

Median

68, 74, 78, 84, 90, 93

Number of elements

0 2.5 5

Percentile

p = (6–1)*0.5 = 2.5

Median position

Interval variables

- 2nd Example. Median: number that is half way between the second number and the third number (counting from 0).

Percentile

68, 74, 78, 84, 90, 93

pl = floor(2.5) = 2
pu = ceil(2.5) = 3

Median = 78+(84-78)*0.5

Quartiles?

Interval variables

- 2nd Example

1st quartile

68, 74, 78, 84, 90, 93

Number of elements

Percentile

0 1.25 5

(6–1)*0.25 = 5*0.25 = 1.25

1st quartile position

Interval variables

- 2nd Example. 1st quartile: number that is 0.25 of the way between the first and the second numbers (counting from 0)

Percentile

68, 74, 78, 84, 90, 93

pl = floor(1.25) = 1
pu = ceil(1.25) = 2

1st quartile = 74+(78−74)*0.25 = 74 +1 = 75

Interval variables

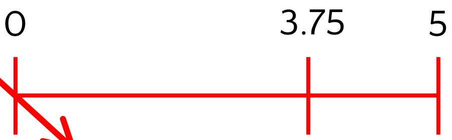
- 2nd Example

3rd quartile

68, 74, 78, 84, 90, 93

Number of elements

Percentile



(6–1)*0.75 = 5*0.75 = 3.75

3rd quartile position

Interval variables

- 2nd Example. 3rd quartile: number that is 0.75 of the way between the third and the fourth number (counting from 0).

Percentile

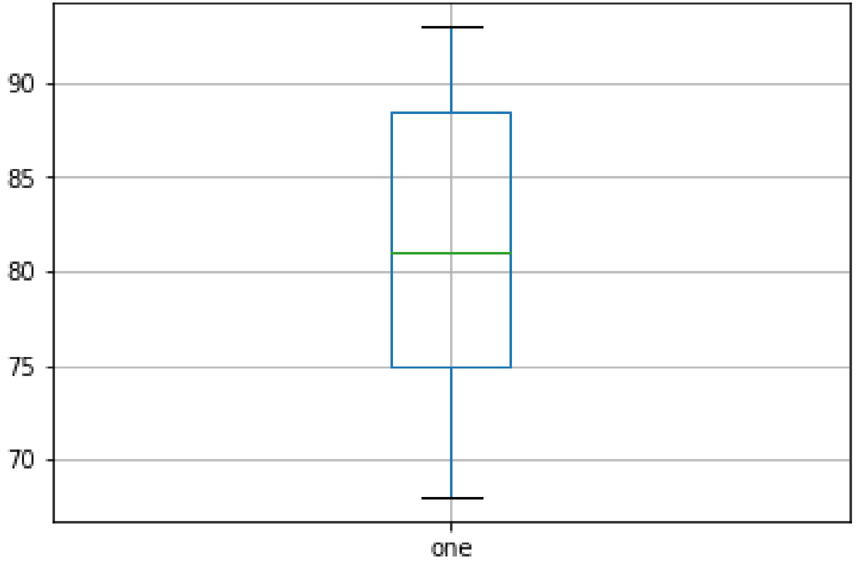
68, 74, 78, 84, 90, 93

pu = ceil(3.75) = 4
pl = floor(3.75) = 3

3rd quartile = 84+(90-84)*0.75 = 84 + 4.5 = 88.5

Interval variables

- 2nd Example. Boxplot



Interval variables: Modified boxplot

- Five number summary of the following data. Draw boxplot

21, 22, 23, 19, 20, 21, 22, 23, 25, 21, 26, 45, 14

First, we sort the data

14, 19, 20, 21, 21, 21, 22, 22, 23, 23, 25, 26, 45

Min and max values are a bit “out” the other numbers. → **Outliers**

Interval variables: Modified boxplot

- IQR: Inter Quartile Range

$$\text{IQR} = \text{3rd Quartile} - \text{1st Quartile}$$

- Inner fences (upper and lower)

Upper inner fence = 3rd Quartile + 1.5(IQR)
Lower inner fence = 1st Quartile – 1.5(IQR)

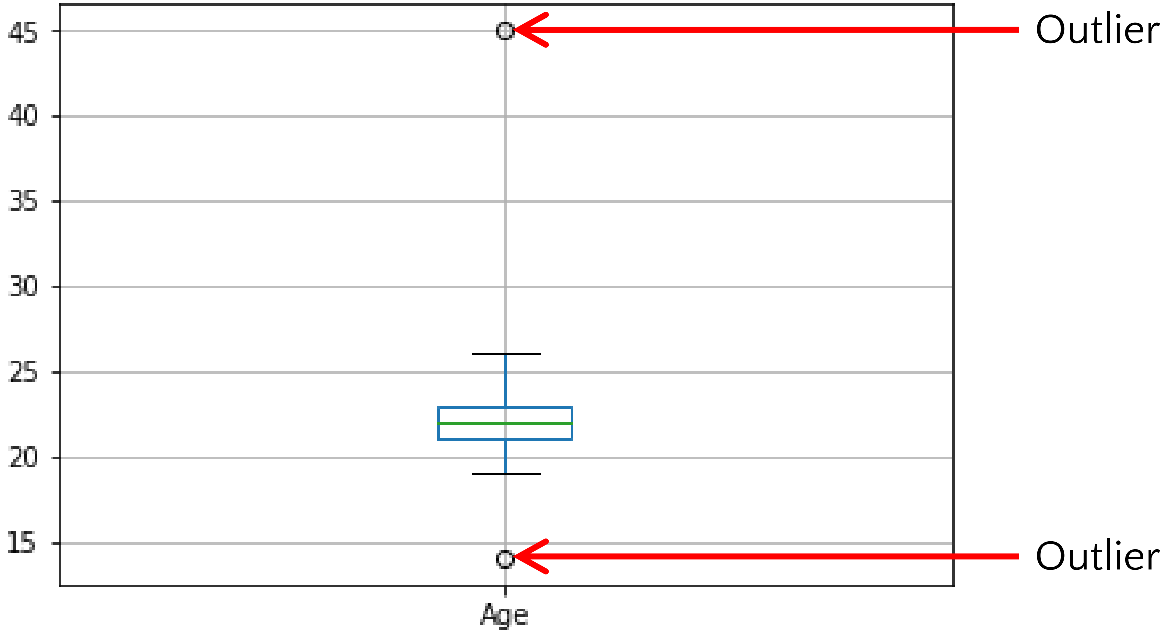
Interval variables: Modified boxplot

- Upper whisker = data point closest (less than or equal) to the upper inner fence
- Lower whisker = data point closest (greater than or equal) to the lower inner fence

- The data points that are outside the upper and lower whiskers are the **outliers**.

Interval variables: Modified boxplot

Outliers may be worthy of attention



Interval variables: Center of the data (Mean)

- Median
- **Mean** or average (expected value or arithmetic center)

$$mean = \frac{\sum data\ values}{\#\ of\ data\ values}$$

$$data\ values = x_1, x_2, x_3, \dots, x_n$$

$$x_i = i^{th} data\ value$$

$$n = \#\ of\ data\ values$$

$$mean = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1, x_2, x_3, \dots, x_n}{n}$$

Interval variables: Trimmed data

33750	95000	205000
33750	103500	292500
33750	112495	301999
33750	138188	4600000
44000	141666	5600000
44000	181500	
44000	185000	
44000	190000	
45566	194375	
65000	195000	

Interval variables: Center of the data

- **Mean** is NOT a **robust** statistic: it is not resistant to extreme values of observations

- Median is a robust statistic

Interval variables: Trimmed mean

1. a% trimmed mean, delete the largest k and the smallest k of the data. k=a/100*n, where n is the number of data points. If k is not an integer, take the integer less than k.

2. Compute the mean again with the remaining data.

Trimmed mean is more robust than **mean**

Interval variables: Spread of the data

- How far is the data from its central (expected) value?
- Range = Maximum – Minimum
→ All the data fits in this interval
- IQR = 3rd Quartile – 1st Quartile
→ Middle half of the data fits in this interval

These measures do not consider all the data values, just some summary values

Interval variables: Spread of the data

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

$$standard\ deviation = \sigma = \sqrt{variance}$$

Interval variables: Spread of the data

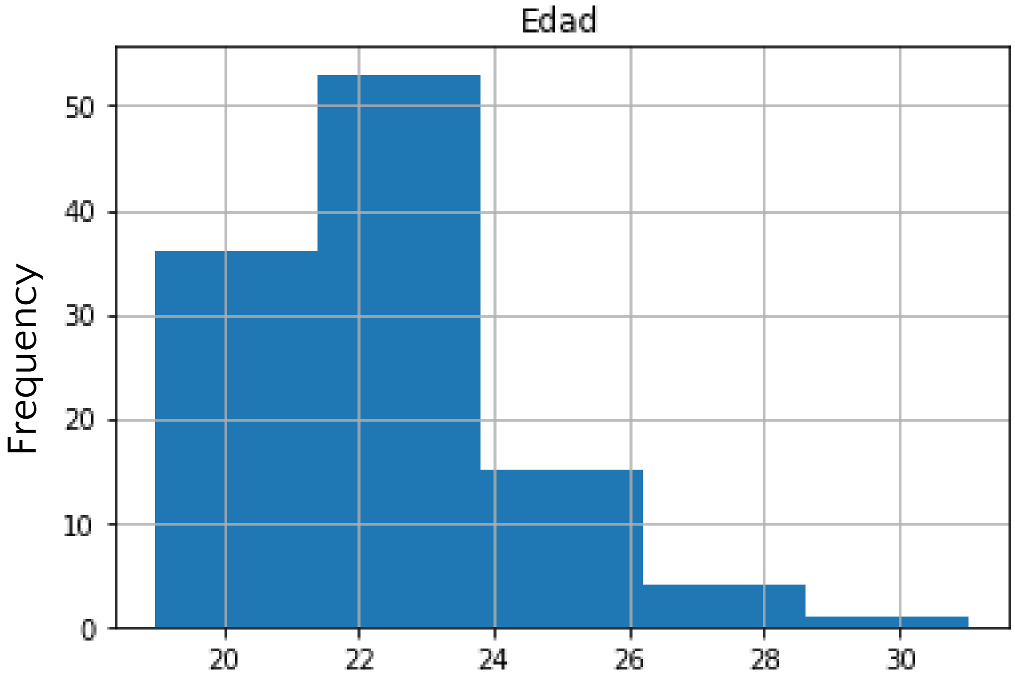
	Original	Trimmed	Robust
Median	$112,495	$112,495	
Mean	$518,311	$128,109	
Range	$5,566,250	$268,249	
IQR	$150,375	$146,000	
S.D.	$1,360,762	$81,967	

Interval variables: Shape of the data

- **Distribution**: The pattern of values in the data, showing their frequency of occurrence relative to each other.

- **Histogram**: Plot to visualize the distribution.

Interval variables: Shape of the data



Interval variables: Shape of the data

- **Histogram**

Divide the data in intervals or “bins” that are mutually exclusive (do not overlap) and exhaustive (include all the data).

For a bin

Data >= lower limit interval

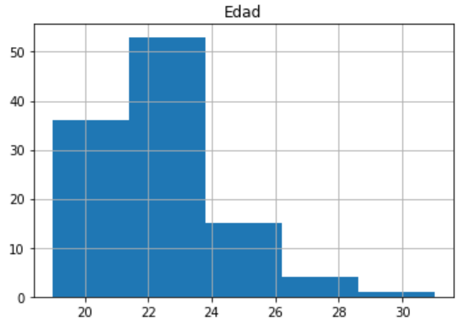
Data < upper limit interval

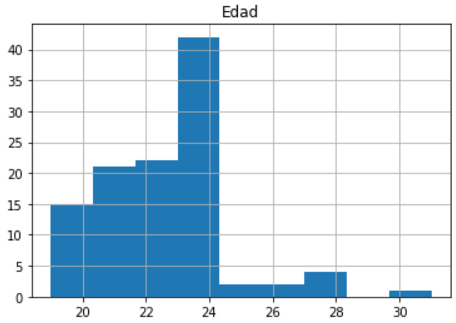
Interval variables: Shape of the data

Histogram

Different bin sizes produce different distributions and reveal different properties of the data. There is no a “best” number of bins.

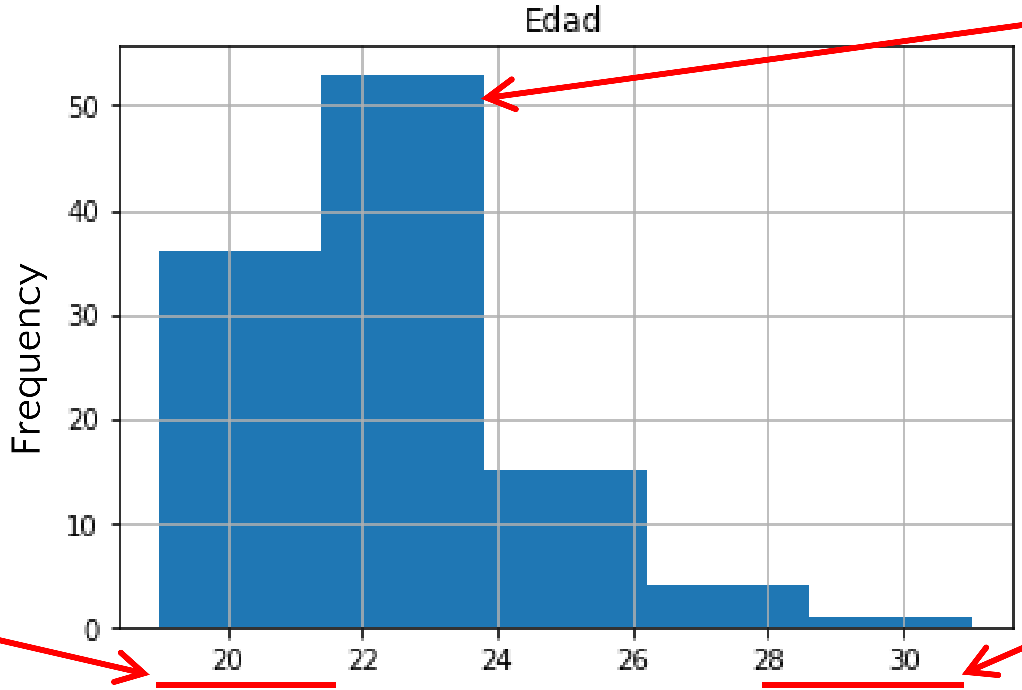
5 bins





9 bins

Interval variables: Shape of the data



Peak of the data

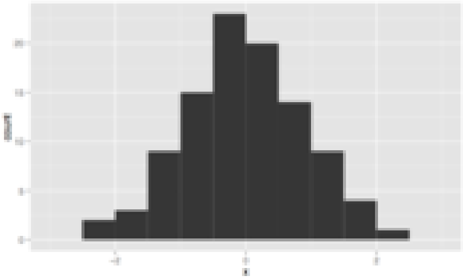
If there is a data value here, that is the **mode**: most frequent value in the distribution

Left tail

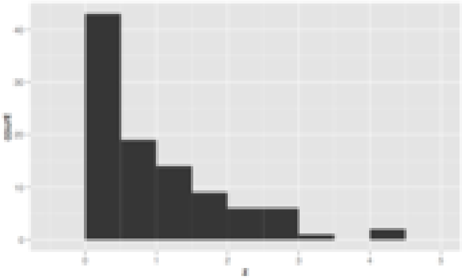
Right tail

Interval variables: Shape of the data

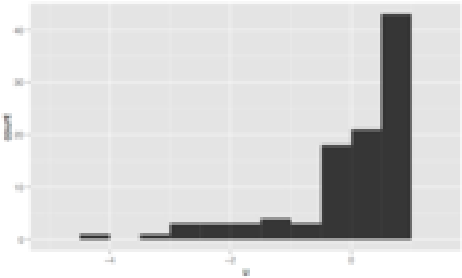
Unimodal, symmetric



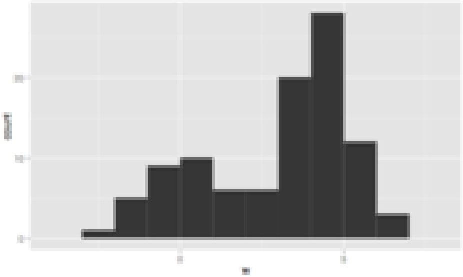
Skewed right



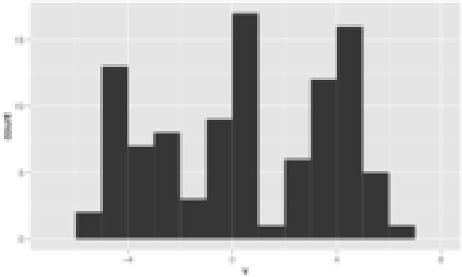
Skewed left



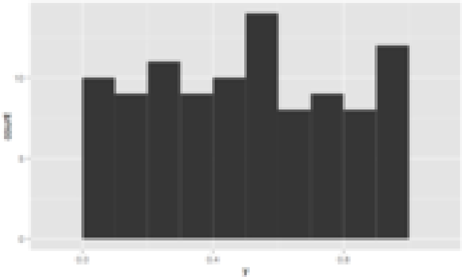
Bimodal (2 peaks)



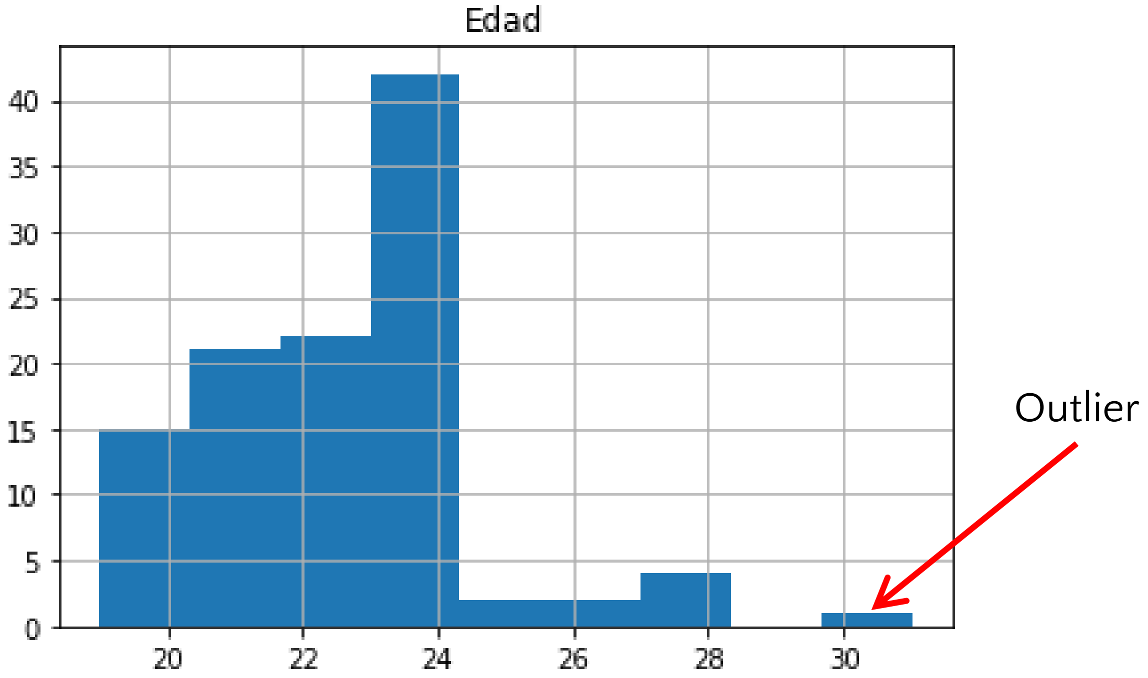
Multimodal (several peaks)



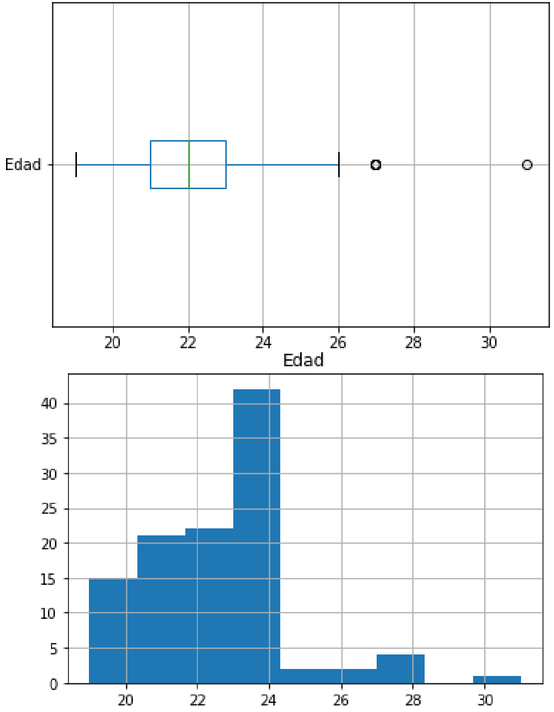
Uniform, symmetric



Interval variables: Shape of the data



Interval variables: Shape of the data



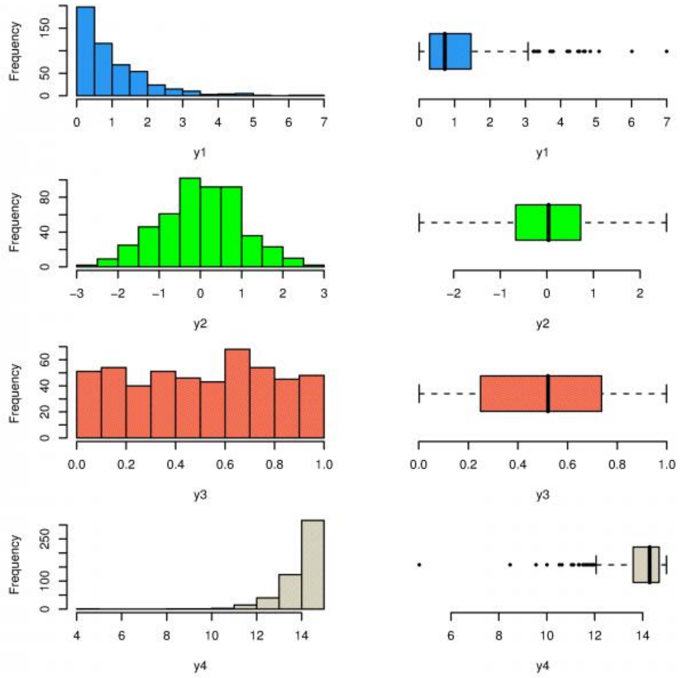
Unimodal
Right skewed
Some outliers

Interval variables: Shape of the data

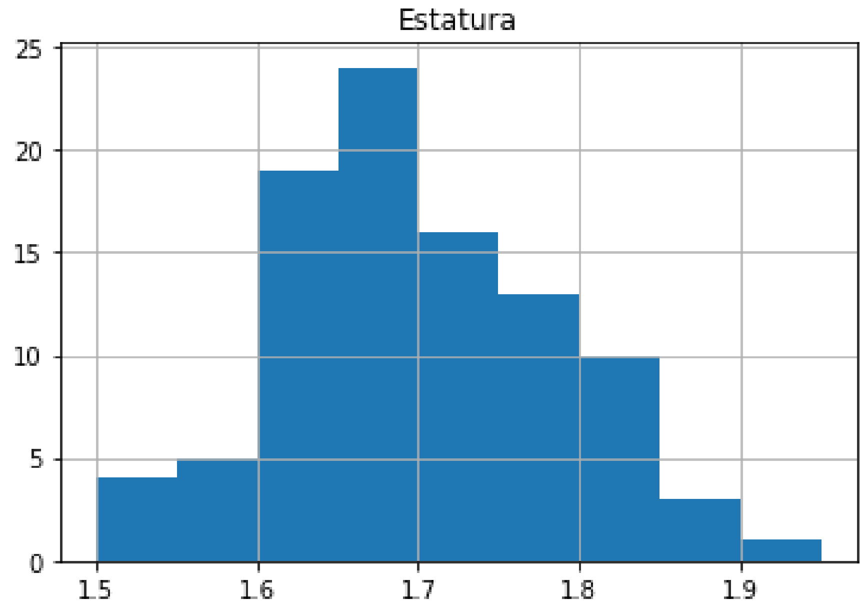
Min	1st Quartile	Median	Mean	3rd Quartile	Max
19	21	22	22.4	23	31

Median-Min = 3
Max-Median = 9

Interval variables: Shape of the data



Interval variables: Shape of the data



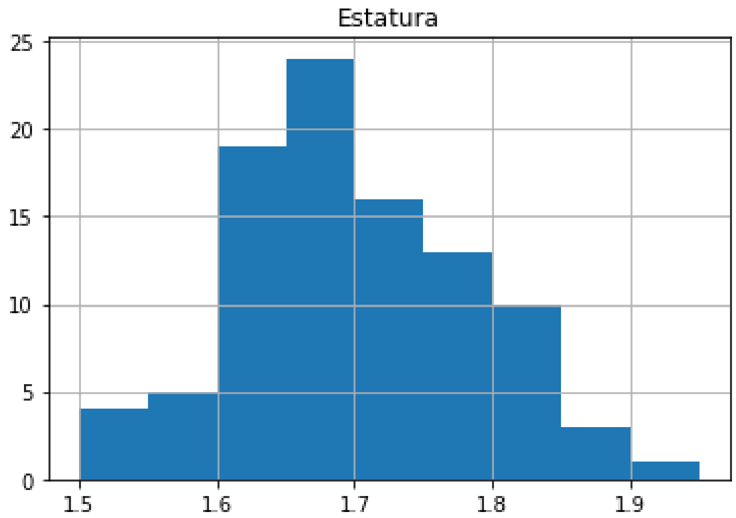
Unimodal
Symmetric
No outliers
Bell shape

→ **Normal distribution**

Interval variables: Shape of the data

For this kind of distribution, we can apply **The Empirical Rule**:

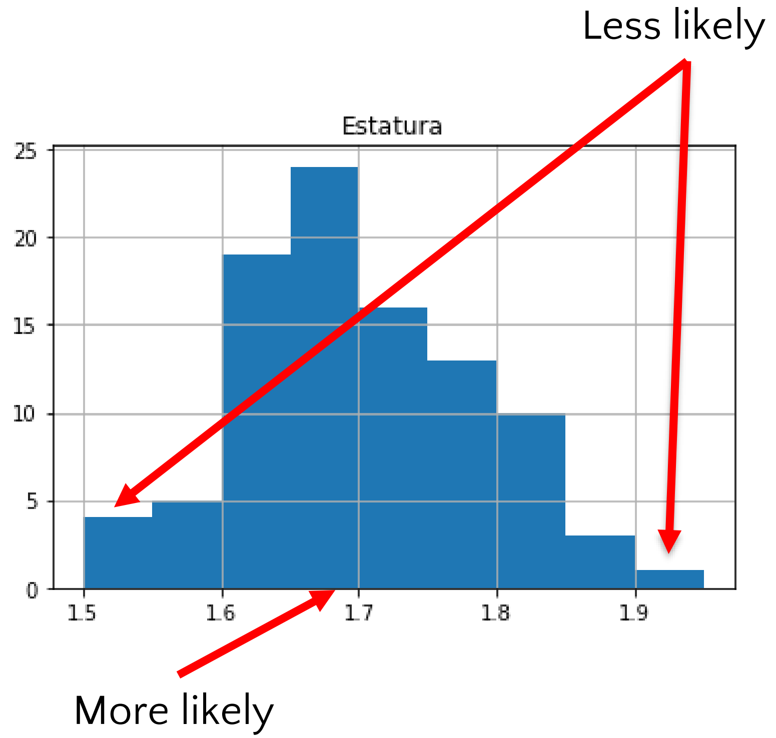
- ~68% of data is between mean–(1 sigma) and mean+(1 sigma)
- ~95% of data is between mean–(2 sigma) and mean+(2 sigma)
- ~99.7% of data is between mean–(3 sigma) and mean+(3 sigma)



Interval variables: Shape of the data

Data distribution is related with probability:

- How likely is to find an observation in some specific range of values?
- Where is more likely to find observations?



Interval variables: Correlations

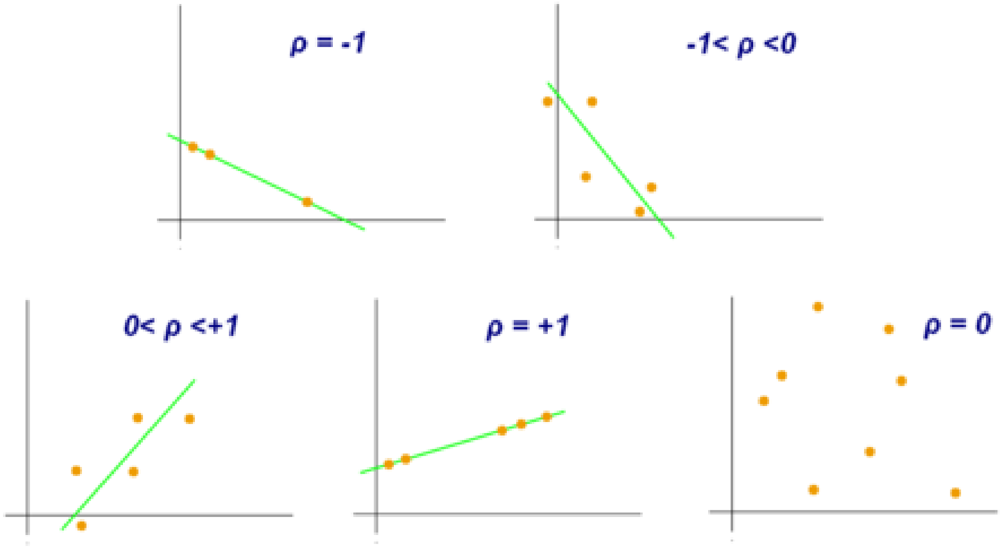
- **Pearson correlation coefficient**

Measure of the linear dependency between two variables X and Y. Range between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Are X and Y changing together?

Interval variables: Correlations

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



Interval variables: Correlations

Estimate Pearson correlation coefficient

Temperature	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

Interval variables: Correlations

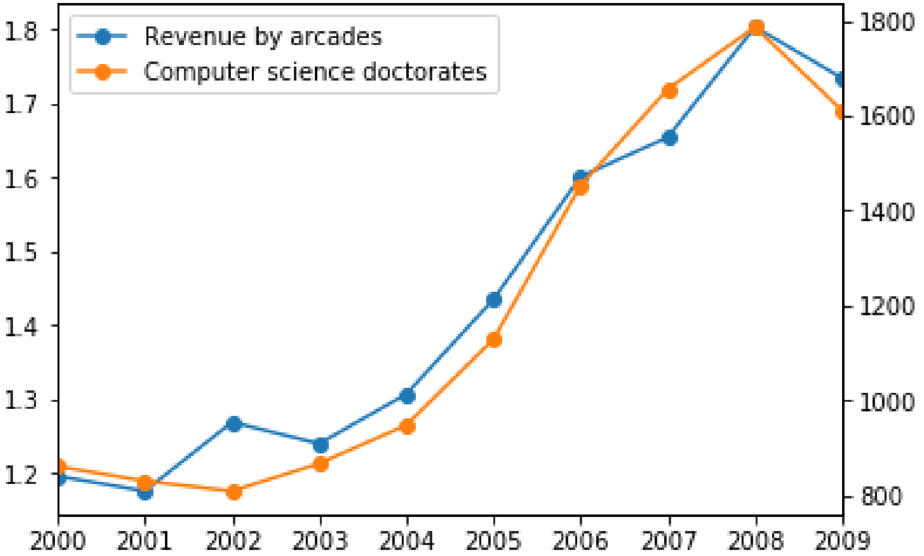
- **Correlation does not imply causation**

Correlation between two variables does not imply that one causes the other.

There could be “spurious correlations”

Interval variables: Correlations

- Correlation does not imply causation



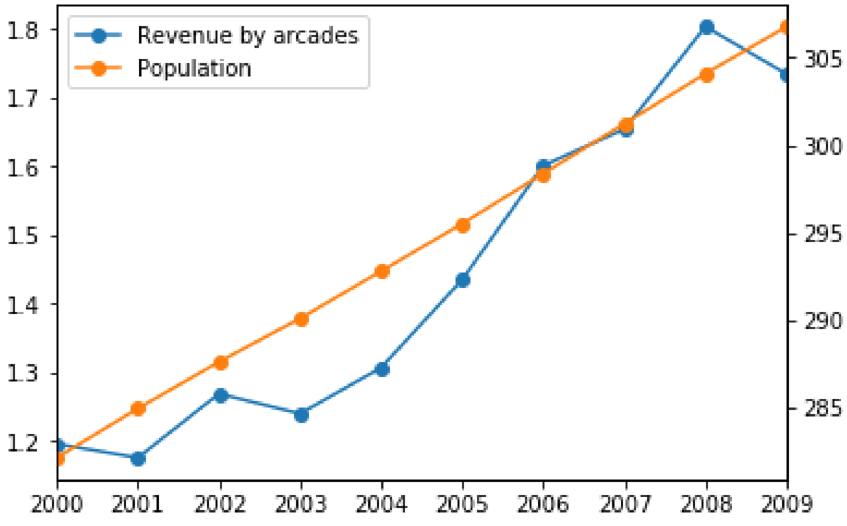
$r = 0.98$

Campus Irapuato-Salamanca | División de Ingenierías

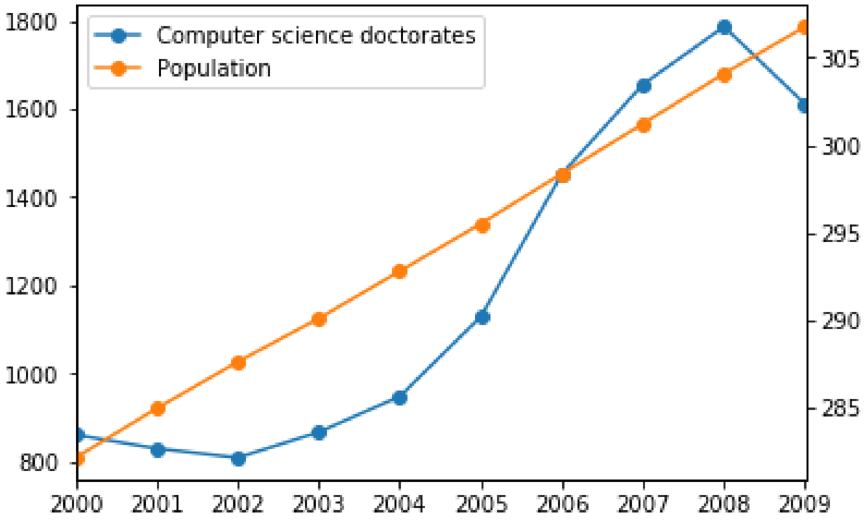
Interval variables: Correlations

- Correlation does not imply causation

$r = 0.96$



$r = 0.92$



All variables: Contingency table

- Displays the (multivariate) frequency distribution of variables

	Sex		
Console	**Men**	**Women**	**Total**
Y	60	15	75
N	20	25	45
Total	80	40	**120**

Nominal variables

Contingency table: Contingency table

- When there are continuous variables, an alternative is to discretize it in groups

lista = [1.75, 1.63, 1.89, 1.88, 1.66, 1.72, 1.65, 1.80, 1.77, 1.71]

groups = {1:[1.61, 1.70], 2:[1.71, 1.80], 3:[1.81, 1.90]}

gruplist = [2, 1, 3, 3, 1, 2, 1, 3, 2, 2]

Contingency table: Odds ratio

- For two **binary** variables (X and Y), it measures the ratio of the odds of X in the presence of Y and the odds of X in the absence of Y

	X		
Y	**1**	**0**	**Total**
1	n_{11}	n_{10}	n_{1*}
0	n_{01}	n_{00}	n_{0*}
Total	n_{*1}	n_{*0}	n

In case one (or more) cell(s) contains a zero, add 0.5 to all cells (**Haldane-Anscombe correction**)

$$OR = \frac{n_{11}n_{00}}{n_{10}n_{01}}$$

Contingency table: Odds ratio

- For two **binary** variables (X and Y), it measures the ratio of the odds of X in the presence of Y and the odds of X in the absence of Y

	Sex		
Console	**Men**	**Women**	**Total**
Y	60	15	75
N	20	25	45
Total	80	40	**120**

The variables are **independent if and only if the ratio is 1**. For a ratio > 1, the variables are **positively associated**. For a ratio < 1, the variables are **negatively associated**.

$$OR = \frac{60 * 25}{15 * 20} = \frac{1500}{300} = 5$$

Contingency table: Pearson's phi coefficient

- Measure of association for two **binary** variables, interpreted similarly to Pearson correlation coefficient (–1 to 1).

	X		
Y	**1**	**0**	**Total**
1	n_{11}	n_{10}	n_{1*}
0	n_{01}	n_{00}	n_{0*}
Total	n_{*1}	n_{*0}	n

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1*}n_{0*}n_{*1}n_{*0}}}$$

Contingency table: Pearson's phi coefficient

- Measure of association for two **binary** variables, interpreted similarly to Pearson correlation coefficient (–1 to 1).

	Sex		
Console	**Men**	**Women**	**Total**
Y	60	15	75
N	20	25	45
Total	80	40	**120**

$$\phi = \frac{(60)(25)-(15)(20)}{\sqrt{(75)(45)(80)(40)}} = \frac{1500-300}{\sqrt{10800000}} = \frac{1200}{3286.3} = 0.365$$

Contingency table: Pearson's chi-squared test

- Determines whether there is a **statistically significant difference** between the **expected** frequencies and the **observed** frequencies in one or categories of a contingency table

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i}$$

k : Number of categories
x_i : Observed frequency for category i
m_i : Expected frequency for category i

Contingency table: Pearson's chi-squared test

Null hypothesis: The type of work is independent od the neighborhood of residence

	Neighborhood				
Work Type	**A**	**B**	**C**	**D**	**Total**
White collar	90	60	104	95	**349**
Blue collar	30	50	51	20	**151**
No collar	30	40	45	35	**150**
Total	**150**	**150**	**200**	**150**	**650**

White collar

$$\frac{349.150}{650} = 80.54$$

Total sample

Neighborhood A

Expected value of white collars in neighborhood A

Observed

Expected

$$\frac{(90 - 80.54)^2}{80.54} = 1.11$$

Expected

Contingency table: Pearson's chi-squared test

If the test is improbably large according to that chisquared distribution, the **null hypothesis is rejected**. (Table of probabilities)

	Neighborhood				
Work Type	**A**	**B**	**C**	**D**	**Total**
White collar	90	60	104	95	**349**
Blue collar	30	50	51	20	**151**
No collar	30	40	45	35	**150**
Total	**150**	**150**	**200**	**150**	**650**

Value in tables for 6 degrees of freedom and probability (p) of 0.05 of exceeding the critical value = **12.59**

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} = 24.6$$

Null hypothesis is rejected

There is dependency

Degrees of freedom = (number of rows–1)(number of columns –1) = (3–1)(4–1) = 6

Contingency table: Pearson's chi-squared test

Value in tables for 1 degree of freedom and probabilty (p) of 0.05 of exceeding the critical value = **3.84**

Null hypothesis is rejected

There is dependency

	Sex		
Console	**Men**	**Women**	**Total**
Y	60	15	75
N	20	25	45
Total	80	40	**120**

$$E_{ym} = \frac{(75)(80)}{120} = 50$$

$$E_{nm} = \frac{(45)(80)}{120} = 30$$

$$E_{yw} = \frac{(75)(40)}{120} = 25$$

$$E_{nw} = \frac{(45)(40)}{120} = 15$$

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} = \frac{(60-50)^2}{50} + \frac{(20-30)^2}{30} + \frac{(15-25)^2}{25} + \frac{(25-15)^2}{15} = \mathbf{16}$$

Degrees of freedom = (number of rows-1)(number of columns -1) = (2-1)(2-1) = 1

Contingency table: Point Biseral Correlation

Same as Pearson correlation coefficient, but for a **binary** variable and a **continuous** variable, e.g. console and height

Mean of data from group 0

Mean of data from group 1

Number of data points in group 0

$$r_{pb} = \frac{M_0 - M_1}{S_y}\sqrt{\frac{n_0}{n}\frac{n_1}{n}}$$

Number of data points in group 1

Standard deviation of the continuous variable

Total number of data points

Contingency table: Conditional probability

- Measure the probability of an event occurring, given that another event has already occurred. For discrete values:

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

$P(A|B)$: Probability of A occurring given that B has occurred
$P(A \cap B)$: Probability of A and B occurring together
$P(B)$: Probability of B occurring

Probability: Conditional probability

- Compute the conditional probability of dying given that a person is a man or a woman, based on the following data.

Sequence	Status	Genre
1	Die	Man
2	Die	Man
3	Die	Man
4	Live	Man
5	Die	Women
6	Die	Women
7	Live	Women

	Sex		
Status	**Men**	**Women**	**Total**
Die	3	2	5
Live	1	1	2
Total	4	3	7

Probability: Conditional probability

- Compute the conditional probability of dying given that a person is a man or a woman, based on the following data.

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

$$P(die|man) = \frac{P(die \cap man)}{P(man)}$$

$$P(die|man) = \frac{3}{7}$$

$$P(man) = \frac{4}{7}$$

$$P(die|man) = \frac{3/7}{4/7} = \frac{21}{28} = 0.75$$

$$P(die|wom) = \frac{P(die \cap wom)}{P(wom)}$$

$$P(die|wom) = \frac{2}{7}$$

$$P(wom) = \frac{3}{7}$$

$$P(die|wom) = \frac{2/7}{3/7} = \frac{14}{21} = 0.66$$

Probability: Conditional probability

- Compute the conditional probability of having a console given that a person is a man or a woman.

	Sex		
Console	**Men**	**Women**	**Total**
Y	60	15	75
N	20	25	45
Total	80	40	**120**

$$P(c|m) = \frac{P(c \cap m)}{P(m)}$$

$$P(c|m) = \frac{60}{120}$$

$$P(m) = \frac{80}{120}$$

$$P(c|m) = \frac{60/120}{80/120} = \frac{7200}{9600} = 0.75$$

$$P(c|w) = \frac{P(c \cap w)}{P(w)}$$

$$P(d|m) = \frac{15}{120}$$

$$P(w) = \frac{40}{120}$$

$$P(c|m) = \frac{15/120}{40/120} = \frac{1800}{4800} = 0.38$$

Basics on statistics

- All the previous was part of **statistical analysis** for **structured data**.

Statistic was used to describe the structured data and to find relations (**patterns**) among variables.

End topic 3

Next topics

- Analysis of unstructured data
- Unsupervised learning

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival