Statistics Pre-course

Part 1

Descriptive Statistics

Alfonso Russo

Department of Economics and Finance

Tor Vergata University of Rome

September 2024

General Information

- **Instructor:** Alfonso Russo
- **Email:** alfonso.russo@uniroma2.it
- **Office Hours:** After class or by appointment.
- **Objectives:** The aim of the preparatory course, held in the first half of September, is to review the fundamental concepts of both descriptive and inferential statistics and to provide students with all the necessary tools to successfully attend the MSc in Economics and in Finance and Banking. Attendance is highly recommended for those students that do not have a strong background in statistics, but it can be a good opportunity to review and deepen the understanding of several key issues for students with solid statistical foundations.
- **Pre-requisites:** Students are assumed to be familiar with undergraduate-level calculus and linear algebra.

Part I Syllabus

- **Description:** with descriptive statistics we summarise and describe data. The description often results in graphs (boxplot, histograms, etc) and numbers (averages, percentages, etc)
- **Inference:** inferential statistics refers to the process of making decisions or predictions about a population, after analysing a sample of observed unit from that population
- The bridge connecting Description and Inference is **Probability**

Statistical Variables

- A statistical **variable** y is any characteristic observed in a study
- The different values taken by a variable, that we are able to measure and record, are called **observations**
- **Unit** usually indexed by i is a member of the population
- The collection of all units (families, individuals, groups, etc) is the **population** we are interested in finding out about
- A **Sample** e.g. $(y_1, \ldots, y_N)$ is a subset of population that we observe
- The way a variable is presented for a specific statistical unit e.g $y_i = x_i$ is the **modality**

Example

- **Example**: Grades in Tor Vergata Statistics' Course
- **Observations**: $18, 24, 22, 28, 30, 30L, 25, 25, 23, 24, \ldots$
- **Unit**: single students
- **Population:** all students of Tor Vergata
- **Sample:** one class (e.g. you!!)
- **Modality:** $18 : 30$

Types of Variables

- **Qualitative (Categorical)**
  - Nominal
  - Ordinal
- **Quantitative (Numerical)**
  - Discrete
  - Continuous

Categorical Variables: variables that are **not** numbers

A variable is categorical if its observations belong to one of a set of distinct categories

- **Nominal**: categories are disconnected
  - Eyes colour, McDonald's Sandwiches, Citizenships, etc
- **Ordinal**: categories are ranked
  - Level of Education (PhD, Master, Bachelor, ..), Energy classes (A++, A+, B, ..), Rating of products (5 starts, 4 stars, 3 stars, ...), etc

Your guess!

- Netflix membership plans
  - Modalities: Base, Standard, Premium
- Musical genres
  - Modalities: Jazz, Pop, Soul, Rock, ..
- Basketball roles:
  - Modalities: 1, 2, 3, 4, ..

Quantitative variables:

variables that are numbers

A variable is quantitative if its observations take numerical values that reports different magnitudes of a certain measurement

- **Discrete**: the variable assumes values in a countable set. Ask "how many?"
  - TV Show episodes, Grades
- **Continuous**: the variable takes values in a continuous set. Ask "how much?"
  - Time, length, weight, acceleration, most physical measures

Wrap Up: Exercise

For the following variables, determine type, modalities and the statistical unit of measurements.

1. Flight time to reach Rome from the others European capitals.
2. Number of times each main character in Harry Potter says "Expelliarmus/Avada Kedavra" (in all the official movies)
3. Family Income (yearly data)
4. Classification of objects in our solar system (Pluto, Jupiter, Earth, Sun, The Moon, ..)
5. Evaluation Survey of Statistics Pre-Course lectures' quality (Clarity 1 to 5, Organisation 1 to 5, etc)

How do we treat them?

The objective is to shed light on meaningful properties of a certain phenomenon that are not immediately inferable from raw data

- For **Categorical** data, a key feature is the relative number of observations in the different categories
  - How many days were "Sunny" in a certain year?
- For **Quantitative** data, we are mainly interested in the *center* and *variability* of the measurements
  - What is the average annual precipitation rate? Do we observe much variability from year to year?

Frequency Distributions:

for categorical and numerical data

Consider a sample $(y_1, \ldots, y_n)$ taking values in a set $(x_1, \ldots, x_k)$

- **Absolute (Raw) frequency** n_i: how many times the i-th modality appears in the sample

$$n_i = \sum_{j=1}^{N} \mathbb{1}(y_j = x_i) \quad \text{for } i = 1, \ldots, k \tag{1}$$

- **Relative frequency** f_i: proportion of observations taking a certain modality

$$f_i = \frac{n_i}{N} \tag{2}$$

Frequency Table

Frequency Distribution: can be written as a table reporting frequencies of observations across the different modalities that the variable of interest can assume

Modalities	**Absolute frequency** n_i	**Relative Frequency** f_i
x_1	n_1	f_1
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k
Σ	N	?

Video Streaming subscribers data as of November 2019

- **Variable type**: Categorical Nominal
- **Sample**: 614.5 million viewers across the World
- **Unit**: Single streaming viewer
- **Population**: Global video streaming subscriber
- **Modalities**: Netflix, iQiyi, Hulu, Youku, Viu, Alt Balaji, Prime Video, Iflix, E Now, Tencent Video

x	n_i	f_i
iQiyi	100.0	0.163
E Now	18.8	0.031
Netflix	151.6	0.247
Youku	82.1	0.133
Alt Balaji	20.0	0.032
Viu	30.0	0.049
Prime Video	75.0	0.122
Iflix	15.0	0.024
Tencent Video	94.0	0.153
Hulu	28.0	0.046
Σ	614.5	1.000

Cumulative relative frequency:

adding a new column to the table

- If the variable is categorical ordinal or numerical discrete we can define the **Cumulative Relative Frequency** as the proportion of observations that take a certain value.
- Take the modalities in increasing order $x_1, \ldots, x_k$ and define

$$F_i = \sum_{j \leq 1} f_j \tag{3}$$

Example: last year Statistics exam

Consider the following sample, consisting of the grades achieved in last year's Statistics exam.

$$19, 24, 26, 30, 24, 30, 29, 24, 28, 18, 29, 29, 21, 30,$$
$$25, 19, 20, 28, 23, 26, 23, 22, 30, 30, 18, 23, 28, 30, 22$$

- **Sample**: 29 students from last year
- **Population**: Statistics' students at Tor Vergata
- **Variable**: ???

Example: last year Statistics exam

building a frequency table

x	n_i	f_i	F_i
18	2	0.069	0.069
19	2	0.069	0.138
20	1	0.035	0.173
21	1	0.035	0.208
22	2	0.069	0.277
23	3	0.104	0.381
24	3	0.104	0.485
25	1	0.035	0.520
26	2	0.069	0.589
28	3	0.104	0.693
29	3	0.104	0.797
30	6	0.203	1
Σ	29	1	

Exercise

- The following table reports the number of M&Ms bags Alfonso bought from January to September 2020.

Month	x
Jan	2
Feb	1
Mar	3
Apr	0
May	?
Jun	4
Jul	1
Aug	3
Sep	2
Σ	19

- Describe the variable: type, unit, etc
- Fill in the missing value ?
- Complete the table adding columns for relative and cumulative frequencies

- When variables are continuous, frequency tables are built based on intervals rather than on single modalities
- Divide all the modalities in a finite number of classes $[\ell_1, u_1), \ldots, [\ell_k, u_k]$
- **Absolute (Raw) frequency** n_i: how many times the i-th modality appears in the sample (in a given class)

$$n_i = \sum_{j=1}^{N} \mathbb{1}\big(y_j \in [\ell_i, u_i)\big) \quad \text{for } i = 1, \ldots, k \tag{4}$$

Frequency tables

for continuous data

Classes $[\ell, u)$	**Absolute frequency** n_k	**Relative Frequency** f_k
$[\ell_1, u_1)$	n_1	f_1
$\vdots$	$\vdots$	$\vdots$
$[\ell_K, u_K]$	n_K	f_K
Σ	N	1

CAVEAT: classes can have different sizes!

Kill Bill Vol.1

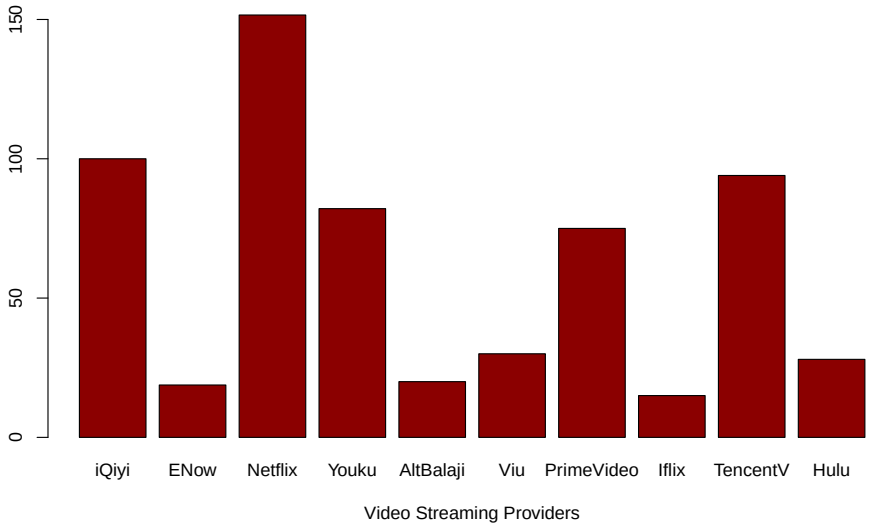
- Observations are the times (minutes) at which a swearword is said in the movie
- y_i is the i-th swearword/blasphemy said in the movie
- Classes are equally spaced (10 minutes intervals)

$[\ell, u)$	n_k	f_k
[0, 10)	3	0.02
[10, 20)	22	0.18
[20, 30)	10	0.08
[30, 40)	11	0.09
[40, 50)	10	0.08
[50, 60)	0	0
[60, 70)	9	0.08
[70, 80)	7	0.06
[80, 90)	46	0.40
[90, 100]	1	0.01

- Frequency tables allow us to describe and summarise the data
- Graphical representations make even easier to analyse it
- **Barplot**: for categorical and discrete data
  - each modality is associated to a bar whose height corresponds to the absolute frequency
- **Histogram**: for continuous data
  - each class is associated to a bar whose area corresponds to its density

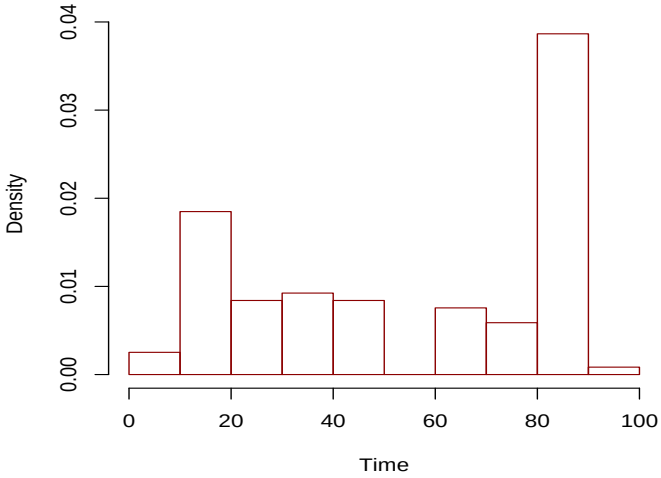
Video Streaming providers

Barplot



Kill Bill Vol.1

Histogram



Exercise

- The following table reports the number of lighters Bob bought last year

	n	x
	0	2
	1	2
	2	5
	3	2
	4	1
Σ		12

- Represent this table graphically and justify your choices

Exercise

The number of people treated in the emergency service of a hospital every day of November was:

$15, 23, 12, 10, 28, 7, 12, 17, 20, 21, 18, 13, 11, 12, 26$

$30, 6, 16, 19, 22, 14, 17, 21, 22, 9, 23, 13, 11, 23, 24$

- Construct the frequency distribution table of the sample.
- Draw a suitable chart for the frequency distribution

- If one is interested in conveying conciser information, **numerical summaries** are suitable since they describe variables with numbers
- The features to focus on are:
  - **centrality**: describing what is the "typical" value for the variables
  - **variability**: describing whether the observations take similar values or they differ from each other

What is the "typical" value for the observations?

- **Mode**
  - the value that appears more often (it can be more than one!)
- **Median**
  - the value that splits in half the distribution
- **Mean**
  - the *balance* point of the distribution

These can all be computed from a frequency table!

The Mode

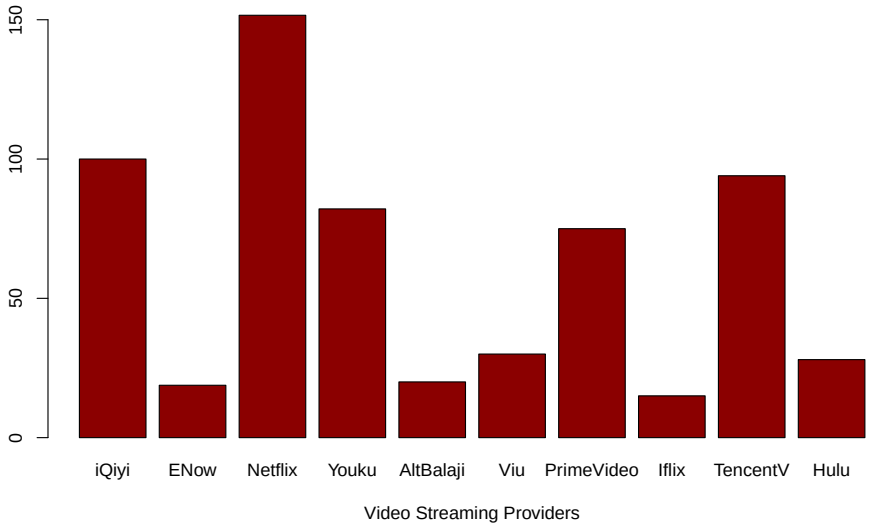
- The mode is the modality with the highest observed frequency

$$y_{Mode} = \{x_j : f_j \geq f_i \, \forall \, j \neq i\}$$

- This is the most general notion of *centrality* and applies to all type of data (numerical and categorical)
- If data are grouped (e.g. divided into classes) then the notion of mode becomes the *modal class*

Note: there can be more than one mode!

Back to the video streaming example



- The median is the value that splits in half the distribution when observations are ordered.
- Let $y_{(1)}, \ldots, y_{(N)}$ be the ordered sample:
  - If N is odd, the median value is:

$$y_{MED} = y_{\left(\frac{N+1}{2}\right)} \tag{5}$$

  - If N is even, the median value is:

$$y_{MED} = \frac{1}{2}\left(y_{\left(\frac{N}{2}\right)} + y_{\left(\frac{N}{2}+1\right)}\right) \tag{6}$$

- The median can be computed for all types of data

Examples:

- Grades from Statistics' exam
  - $18, 18, 19, 20, 21, 22, 22, 22, 26, 26, 26, 27, 27, 28, 28 \implies y_{MED} = ?$
  - $18, 18, 19, 20, 21, 22, 22, 22, 26, 26, 26, 27, 27, 28, 28, 29 \implies y_{MED} = ?$
- Customers' survey data
  - Bad, Bad, Average, Average, Good, Great, Great

Another Example with Grades

- Grade from Statistics' students:

  $24, 22, 23, 23, 23, 24, 22, 30, 28, 28, 27, 26, 27,$

  $26, 28, 19, 18, 18, 18, 19, 20, 20, 21, 21, 22, 23$

- Ordered sample:

  $18, 18, 18, 19, 19, 20, 20, 21, 21, 22, 22, 22, 23,$

  $23, 23, 23, 24, 24, 26, 26, 27, 27, 28, 28, 28, 30$

- $N = 26 \implies y_{MED} = \frac{1}{2}(23 + 23) = 23$

x_k	n_k	f_k	F_k
18	3	0.115	0.115
19	2	0.077	0.192
20	2	0.077	0.269
21	2	0.077	0.346
22	3	0.115	0.461
23	4	0 .154	0.615
24	2	0.077	0.692
26	2	0.077	0.769
27	2	0.077	0.846
28	3	0.115	0.961
30	1	0.039	1
N	26	1	

The median is the first modality x_k for which $F_k \geq 0.50$

- The median tells us what is the *level* reached by at least 50% of the population.
- Quantiles are extensions that tell us what is the level reached by "$q\%$" of the population. They are defined by the first modality x_k for which $F_k \geq p$.
- Quantiles of order $p = 0.25$ and $p = 0.75$ have special roles and are called **1st** and **3d qurtile**, respectively.
- In previous example we have $q_{0.25} = 20$ and $q_{0.75} = 26$.

Exercise

- Let X be the "number of products bought by a customer" from a supermarket in Baker Street. We receive observations on 23 customers that visited the supermarket.

$$(5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9, 9, 10, 10, 11, 11, 11, 12, 12) \tag{7}$$

- Find Median(X), $q_{0.25}(X)$ and $q_{0.75}(X)$. Calculate IQR(X).

- Let X be the "number of products bought by a customer" from a supermarket in Baker Street. We receive observations on 23 customers that visited the supermarket.

$$(5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9, 9, 10, 10, 11, 11, 11, 12, 12) \tag{8}$$

- Find Median(X), $q_{0.25}(X)$ and $q_{0.75}(X)$. Calculate IQR(X).
- Median$(X) = 9$
- $q_{0.25}(X) = 7$, $q_{0.75}(X) = 10 \rightarrow$ IQR$(X) = 3$

- The (arithmetic) mean is the sum of the observations divided by the sample size.

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{9}$$

- It is interpreted as the balance point of the distribution since it solves the following minimisation problem

$$\bar{y} = \arg\min_{c} \sum_{i=1}^{N} (y_i - c)^2 \tag{10}$$

Properties of the Mean

- **Internality**

$$y_{(1)} \leq \bar{y} \leq y_{(N)}$$

- **Linearity**
  - If $z_i = ay_i + b$ then

$$\bar{z} = a\bar{y} + b \tag{11}$$

- **Zero-deviation**

$$\sum_{i=1}^{N}(y_i - \bar{y}) = 0 \tag{12}$$

- **Associativity**
  - Let $\bar{y}$ and $\bar{x}$ be the arithmetic means of two samples of sizes N and M, respectively. The mean of the combined sample is then

$$\bar{z} = \frac{N * \bar{y} + M * \bar{x}}{N + M} \tag{13}$$

- The Mean can be approximated directly from a frequency distribution using the alternative definition

$$\bar{y} = \frac{1}{N}\sum_{j=1}^{K} x_j n_j = \sum_{j=1}^{K} x_j \frac{n_j}{N} = \sum_{j=1}^{K} x_j f_j \tag{14}$$

  where x_j are the modalities of the variable

- This formulation can be used also when observations are grouped in intervals (ℓ_j, u_j): it is enough to replace x_j with the centre of the interval

$$\bar{y} = \sum_{j=1}^{K} c_j f_j \qquad \text{where } c_j = \frac{1}{2}(u_j + \ell_j) \tag{15}$$

Exercise

Mean of grouped data

- Salaries of NBA players for the season 2017/2018 (thousands of U.S. dollars)

x	c	n	f	F	c*f
[0.17, 34.8)	17.50	336	0.59	0.59	
[34.8, 69.5)	52.17	76	0.13	0.72	
[69.5, 104)	86.84	41	0.07	0.79	
[104, 139)	121.50	35	0.06	0.85	
[139, 173)	156.17	30	0.05	0.9	?
[173, 208)	190.83	18	0.03	0.93	
[208, 243)	225.50	19	0.03	0.96	
[243, 277)	260.16	9	0.02	0.98	
[277, 312)	294.83	6	0.01	0.99	
[312, 347)	329.49	2	0.01	1	
Σ		572	1		$\bar{y}$

- Compute the mean, the median and the mode. What do you notice by comparing these measures of centrality?

Exercise

Mean of grouped data

- Salaries of NBA players for the season 2017/2018 (thousands of U.S. dollars)

x	c	n	f	F	c*f
[0.17, 34.8)	17.50	336	0.59	0.59	10.33
[34.8, 69.5)	52.17	76	0.13	0.72	6.78
[69.5, 104)	86.84	41	0.07	0.79	6.08
[104, 139)	121.50	35	0.06	0.85	7.29
[139, 173)	156.17	30	0.05	0.9	7.81
[173, 208)	190.83	18	0.03	0.93	5.72
[208, 243)	225.50	19	0.03	0.96	6.77
[243, 277)	260.16	9	0.02	0.98	5.20
[277, 312)	294.83	6	0.01	0.99	2.95
[312, 347)	329.49	2	0.01	1	3.29
Σ		572	1		62.22

- Mode $= [0.17, 34.8)$; Median $= [0.17, 34.8)$, Mean $= 62.22$

Mean vs Median

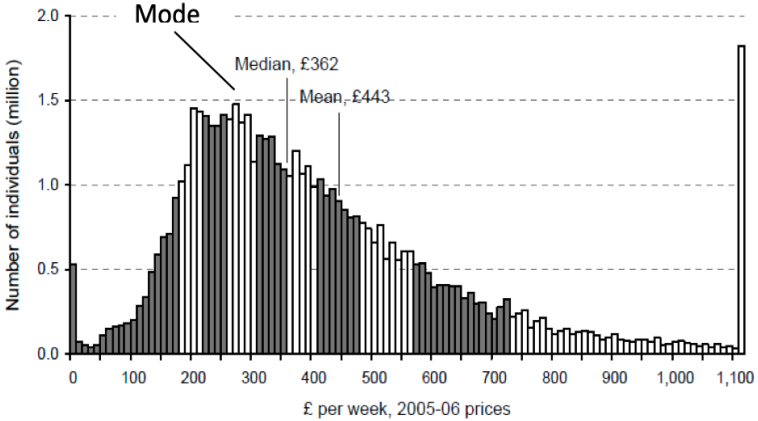
- In the "Mean vs Median" fight we have no winners, they simply measure different things.
- The Mean takes into account all the obsevations, **including the anomalous ones (outliers)**
- The Median takes into account only the order of the observations, **regardless of their values**
- In the NBA example, players with huge pay-checks sensibly affect the value of the mean

The Mean and the Median coincide only when the distribution of the variable is **symmetric**

- Sometimes is hard to select the most appropriate measure of central tendency

Figure 1. The income distribution in 2005–06 (UK)

IFS Briefing Note No 73



Notes: Incomes have been measured before housing costs have been deducted. The right-most bar represents incomes of over £1,100.

Variability

how unpredictable is our variable?

When do we consider observations sufficiently "similar" among each other?

- **Ranges**
  - observations vary in a certain interval
- **Variance**
  - the spread of the observations around a defined value (typically the mean)

Ranges

variability in a set

Based on the concept of **variability** meant as the size of the interval in which the observations fall, we can define two measures of spread:

- **(Global) Range of Variation**: the difference between the maximum and the minimum value observed in the sample

$$RV = y_{(N)} - y_{(1)} \tag{16}$$

- The **Interquartile Range**: the difference between the 3rd and 1st quartile, which gives the smallest interval in which the 50% of the observations falls

$$IQ = y_{Q3} - y_{Q1} \tag{17}$$

Variance

The variance is based on the idea that the larger the deviations from then mean $(y_i - \bar{y})$, regardless of their signs, the larger the variability

NOTE: To exploit the idea of *deviation from the center* we cannot use simply the average of the deviations

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}) \tag{18}$$

because of the **zero-deviation** property of the mean

Solution: square them!

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{19}$$

The **Standard Deviation** is the square root of the variance

$$s = \sqrt{s^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{20}$$

The Standard Deviation is on the same scale of the data, making it more interpretable than the Variance

Your Turn!

Exercise

Back to Tor Vergata Grades.

- The sample: $24, 20, 23, 20, 23, 24, 22, 30, 29, 28, 27, 26, 27, 22, 24, 21, 18,$ $26, 28, 19, 18, 18, 18, 19, 20, 20, 21, 21, 22, 23, 24, 29, 30, 18, 19, 21, 20, 30$
- Compute the Mean, Median and Mode. Comment on findings.
- Plot the data, justifying your graphical choice
- Compute the Interquartile Range, Variance and Standard Deviation. Again, comment on findings.

Solution

x_i	n_i	f_i	F_i
18	5	0.132	0.132
19	3	0.079	0.211
20	5	0.132	0.343
21	4	0.105	0.448
22	3	0.079	0.527
23	3	0.079	0.606
24	4	0.105	0.711
26	2	0.053	0.764
27	2	0.053	0.815
28	2	0.051	0.868
29	2	0.053	0.921
30	3	0.079	1
Σ	38	1	-

- $\mathsf{Mode}(X) = (18, 20), \mathsf{Median}(X) = 22, \bar{X} = 22.89$
- $q_{0.25}(X) = 20, q_{0.75}(X) = 26, \mathsf{IQR}(X) = 6$

- So far we have only dealt with describing, representing and summarising a signle variable
- However, in reality we usually have to analyse more than one variable that **may or may not** be related
- *Examples:* weight and height of a subject, length and budget of a movie, age and hair colour etc

CAVEAT: we will focus only on the case of two numerical variables, albeit association measures also for categorical and mixed-type variables as well

Conditional Mean and Variance

- A two way table for the *Number of Houses* (Y) and *Number of Cars* (X) owned by a sample of 68 families.

		Houses		
		1	2	3
	1	21	8	0
Cars	2	12	11	1
	3	7	6	2

- Since both characters are quantitative, it is possible to calculate their **conditional mean** and **variance**.

$$\bar{y}_{x=x_i} = \frac{1}{n_i}\sum_{j=1}^{K} y_j n_{ij} \qquad \sigma^2_{y|x=x_i} = \frac{1}{n_i}\sum_{j=1}^{K}(y_j - \bar{y}_{x=x_i})^2 n_{ij} \tag{21}$$

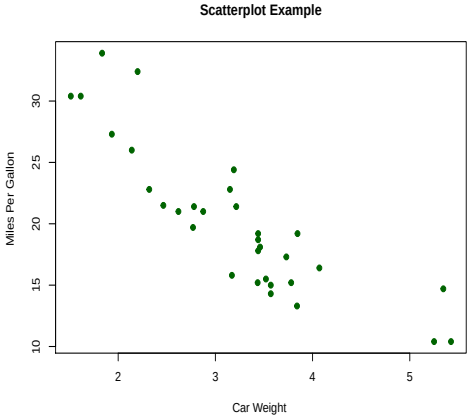
- For example, the average number of Houses **for families who own one Car** is $\bar{y}_{x=1} = (1 \cdot 21/29 + 2 \cdot 8/29 + 3 \cdot 0/29) = 1.28$. Calculate all the remaining conditional means and variances.

Scatterplot

graphical representation of two dimensional variables

A two dimensional variable (X_1, X_2) is usually represented through a scatterplot

- Each axis is related to one of the variables of interest
- Each point represent a unit and its coordinates correspond to the values of the variables observed on it



When we have more than one variable at hand, we want to know whether there is a relationship among them

- **Positive:** when x goes up, y tends to go up
- **Negative:** when x goes up, y tends to go down

The **Covariance** is an indicator that measures the strength of the association between two variables:

$$cov_{x,y} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{22}$$

- The covariance is a measure of linear association; i.e. the case where the relationship between two variables x and y is of the form

$$y_i = ax_i + b \tag{23}$$

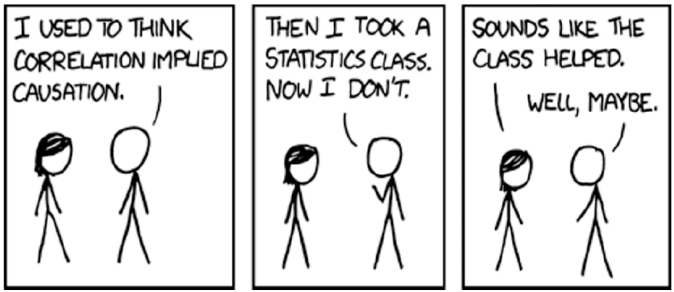
- The value of the covariance depends on the scale of the data. We have no general reference to assess if the sample covariance is large or small

- The **Correlation** is a rescaled version of the Covariance

$$r_{x,y} = \frac{cov_{x,y}}{s_x \, s_y} \tag{24}$$

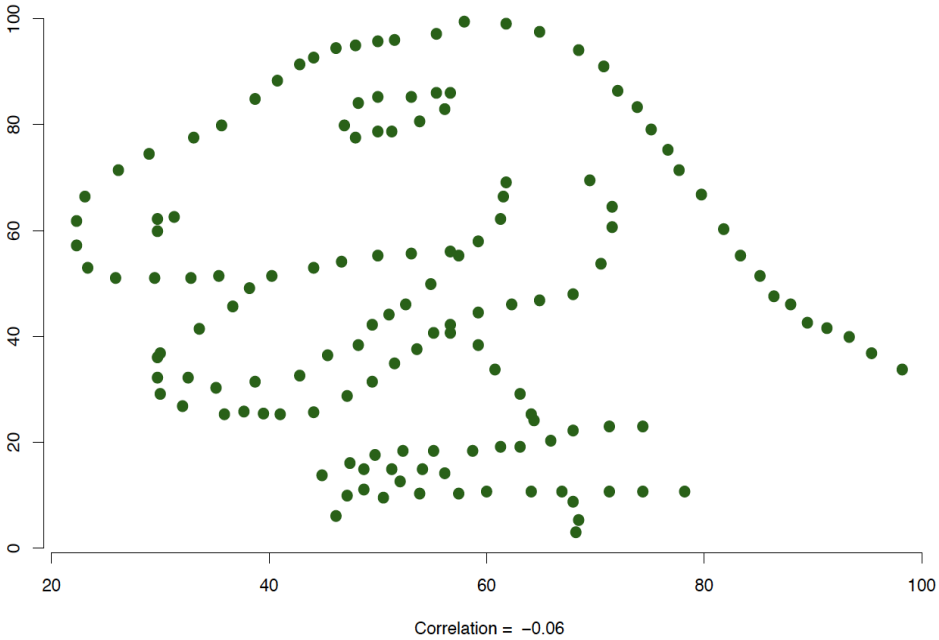
- We have that $r_{x,y} \in [-1, 1]$. The closer $|r_{x,y}|$ to 1, the stronger is the linear association in the observations
- Correlation does not depend on the variables' unit; i.e. it is not affected by the scale of the observations
- The correlation is symmetric with respect to the two variable ($r_{x,y} = r_{y,x}$)

A tattoo you should have on your forehead



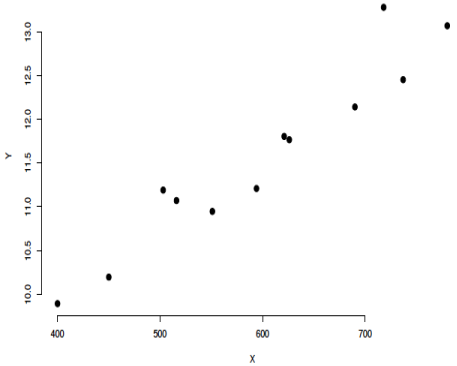
- Correlation is **only a measure of linear dependence**. More complicated forms of relationship, even something basic as $y_i = ax_i^2 + b$, may not be correctly captured by correlation.
- **CORRELATION IS NOT CAUSALITY!** There may be a strong correlation between two variable but it does not directly mean that changes in one are causing changes in the other.

Datasaurus



Spurious relationships

- X = number of people who died falling out of their bed
- Y = number of lawyers in Puerto Rico
- $r_{x,y} = 0.957087$



Exercise

Does LSD help with Math?

- X = Tissue concentration of Lysergic Acid Diethylamide (LSD)
- Y = Math Test score
- For the two variables X and Y computer Mean, Variance, St. Dev, Covariance and Correlation

X	Y
1.17	78.93
2.97	58.20
3.26	67.47
4.69	37.47
5.83	45.65
6.00	32.92
6.41	29.97

Exercise

Does LSD help with Math?

- X = Tissue concentration of Lysergic Acid Diethylamide (LSD)
- Y = Math Test score
- For the two variables X and Y computer Mean, Variance, St. Dev, Covariance and Correlation

X	Y
1.17	78.93
2.97	58.20
3.26	67.47
4.69	37.47
5.83	45.65
6.00	32.92
6.41	29.97

Solution

$\bar{X} = 4.33$; $\bar{Y} = 50.09$; $s_X = 1.79$; $s_Y = 17.23$; $Cov_{X,Y} = -28.92$; $corr_{X,Y} = -0.936$