Statistics: a gentle introduction and Basic Statistics

Lecture 1

Francesco Dotto
francesco.dotto@uniroma3.it

3/9/2020

Outline of the course:

1 **Introduction**: What is statistics?
Descriptive statistics: Tools for describing and summarizing data

2 **Probability**: Basic notions

3 **Probability**: Discrete distributions

4 **Probability**: Continous distributions

5: **Statistical inference**: From the *sample* to the *population*

Statistics: Quantitative analysis of collective phenomena

The art of learning from data

The information we gather with experiments and surveys is collectively called data.

A course in statistics should teach you how to **make sense** of the data.

Two (+ one) elements of statistics:

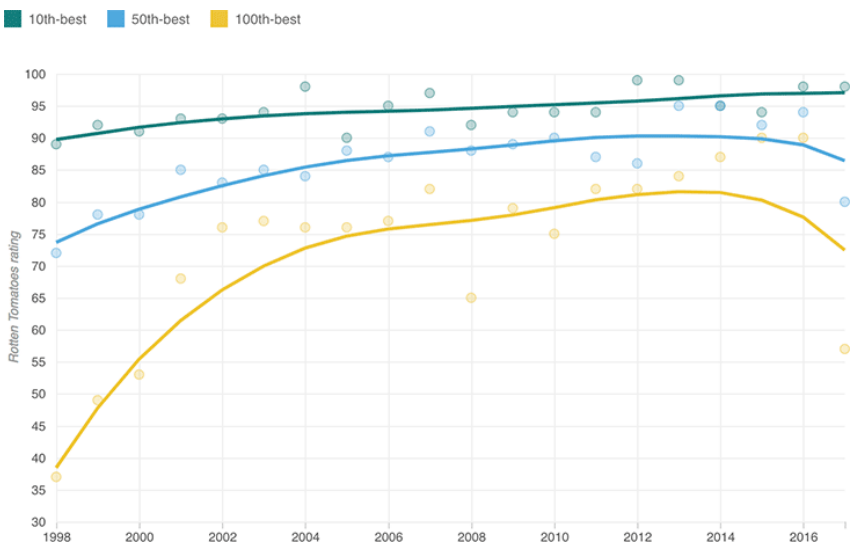
Building blocks of our course

- **Describe**: formulating an hypothesis
  - Descriptive statistics refers to methods for **summarizing the collected data** (where the data constitutes either a sample or a population). The summaries usually consist of plots and numbers such as averages and percentages.
- **Infer**: validating an hypothesis
  - Inferential statistics refers to methods of **making decisions or predictions** about a population, based on data obtained from a sample of that population.

The step from description to inference **probability**.

Toy Example

Are movies getting better?



- **Descriptive Statistics**: there is an increasing trend in the ratings of "good" movies.
- **Inferential Statistics**: can we conclude that movies are getting better or is this trend appearing just by coincidence?

Statistical Variables

A **variable** y is any characteristic observed in a study.

The data values that we observe for a variable are called **observations**.

Statistical Variables

A **variable** y is any characteristic observed in a study.

The data values that we observe for a variable are called **observations**.

- **Unit** - i: A member of a population
- **Population**: The collection of all individuals, families, groups, organizations, and units that we are interested in finding out about
- **Sample** - $y_1, \ldots, y_N$: The subset of the population we observe.
- **Modality** - x_k: The way the variable is presented on a single statistical unit.

Statistical Variables

Variable: Grade in Statistics of Torvergata's Students

Observations: $18, 22, 26, 30, 22, 21, ...$

- **Unit**: the single student
- **Population**: all students of Torvergata
- **Sample**: a class (e.g. you?)
- **Modality**: $18 : 30$

Types of Variables

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

Categorical

A variable is called categorical if **each observation belongs** to one of a **set of distinct categories**.

- **Nominal**: Categories are disconnected.
  - Examples: *Hair Color*, *Religion*,*race*, *McDonald's menu item*, *supported football team*
- **Ordinal**: Categories are ranked.
  - Examples: *Amazon's rating*, *Level of Education*

Examples?

- *Boca Junior players*: Nominal Categorical variable
  - *modalities*: Francesco Totti, Cristiano Ronaldo, ...
- *Basketball roles*: Nominal Categorical variable
  - *modalities*: 1, 2, 3, 4, 5
- *Netflix Membership Level*: Ordinal Categorical variable
  - *modalities*: Base, Standard, Premium

WATCH OUT A variable using numbers as labels for its categories is still a categorical variable and is not quantitative.

Numerical Variable

A variable is called **quantitative** if observations on it **take numerical values** that represent different magnitudes of the variable .

- **Discrete**: the variable assume values in a countable set ("how many")
  - Examples: *Number of episode in a series*, *Grades*
- **Continuous**: the variable assume values in a continuous set ("how much")
  - Examples: *Time*, *most physical measurements*

Exercise

For the following variables, determine type, modalities and the statistical units on which they are observed.

- Number of times an Academy Award winner says “Thanks/Thank you” in its acceptance speech.
- Pro-capita consumption of Alcohol in world countries.
- Names of the dogs died in space.
- Your daily screen-time (i.e. the time you spend on your phone)

What do we do with them?

The objective is to provide insights about the data that **cannot be** quickly **obtained by looking only at the original data**.

What do we do with them?

The objective is to provide insights about the data that **cannot be** quickly **obtained by looking only at the original data**.

- For categorical variables, a key feature to describe is the **relative number of observations in the various categories**.

Example: what percentage of days were sunny in a given year?

What do we do with them?

The objective is to provide insights about the data that **cannot be** quickly **obtained by looking only at the original data**.

- For categorical variables, a key feature to describe is the **relative number of observations in the various categories**.

Example: what percentage of days were sunny in a given year?

- For quantitative variables, key features to describe are the **center and the variability** (sometimes referred to as spread) of the data.

Example: What's a typical annual amount of precipitation? Is there much variation from year to year?

Frequency Distributions

for categorical / numerical discrete variables

Given a sample $y_1, \dots, y_N$ taking values in a set $x_1, \dots, x_K$ we define:

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

$$n_i = \sum_{j=1}^{N} \mathbb{I}(y_j = x_i)$$

- **Relative Frequency** f_i: proportion of the how many $i-th$ modality appears in the sample

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

Frequency Tables

Frequency distribution: tabular summary of data showing the frequency of items in each of several modalities of the variable of interest.

Modalities x	Absolute Frequency n	Relative Frequency f
x_1	n_1	f_1
...	...	...
x_K	n_K	f_K

This is the most general formulation and it applies to both quantitative and qualitative data.

Toy Example

Thanksgiving Preferences

What pie do you eat for thanksgiving?

- **Sample**: 2238 Americans interviewed in 2015
- **Population**: US citizens
- **Variable**: Categorical Nominal
- **Modalities**: Apple, Buttermilk, Cherry, Chocolate, Coconut cream, Key Lime, Peach, Pecan, Pumpkin, Sweet Potato, None, Other

	x	n	f
	Apple	514	0.23
	Buttermilk	35	0.02
	Cherry	113	0.05
	Chocolate	133	0.06
	Coconut cream	36	0.02
	Key Lime	39	0.02
	Peach	34	0.02
	Pecan	342	0.15
	Pumpkin	729	0.33
	Sweet Potato	152	0.07
	None	40	0.02
	Other	71	0.03
Sum		2238	1

Cumulative Relative Frequency

If the variable is categorical ordinal or numerical discrete we can define **Cumulative Relative Frequency**, as the proportion of observation that take a value

Let $x_1, \ldots, x_K$ be the modalities increasingly ordered, then

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

Example:

$x_1 =$ Very Unsatisfied, $x_2 =$ Mildly unsatisfied, $x_3 =$ Neutral, $x_4 =$ Mildly Satisfied, $x_5 =$ Very Satistfied

What is the proportion of customer which is not happy with the service?
F_3

Toy Example

Last year class passed statistics with the following grades:

18 28 26 20 28 18 22 22 18 28 18 18 28 28 28 20 24 22 26 22 30

x	n	f	F
18	5	0.24	0.24
20	2	0.1	0.34
22	4	0.19	0.53
24	1	0.05	0.58
26	2	0.1	0.68
28	6	0.29	0.97
30	1	0.05	1

Exercise

frequency tables

Consider the following absolute frequencies table, where the variable observed is the number of shoes Carlo bought each month of the last year:

	x	n
	0	1
	1	4
	2	3
	3	?
	4	2
Sum		12

- Describe the data. What kind of variable are we dealing with? What are the units?
- If possible, fill-in the missing value in the table.
- Complete the frequency table adding relative and cumulative frequencies.

Frequency table

for continuous numerical variables

When the variable is continuous, we build frequency tables from intervals rather than from the single modalities.

Operationally, given a sample $y_1, \dots, y_N$, we divide all the possible values modalities in a finite number of classes, e.g. $[l_1, u_1), \dots, [l_K, u_K]$.

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

$$n_i = \sum_{j=1}^{N} \mathbb{I}(y_j \in [l_i, u_i))$$

Frequency Tables for Continuous variables

Class $[l, u)$	Absolute Frequency n	Relative Frequency f
$[l_1, u_1)$	n_1	f_1
...	...	...
$[l_K, u_K]$	n_K	f_K

CAVEAT: Classes can have different sizes!

Toy Example

Tarantino's movies

Kill Bill Vol.1

- The observations are the *times* at which a swearword is said in the movie.
- y_i is the time of the $i-th$ swearword / blasphemy in the movie
- Classes are equally sized, 10 minutes interval.

[u,l)	n	f
[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	0.38
[90,100)	1	0.01

Graphical Representation

Frequency tables allow us to summarize the data, but they also allow for graphical representation that makes data even more easy to analyse.

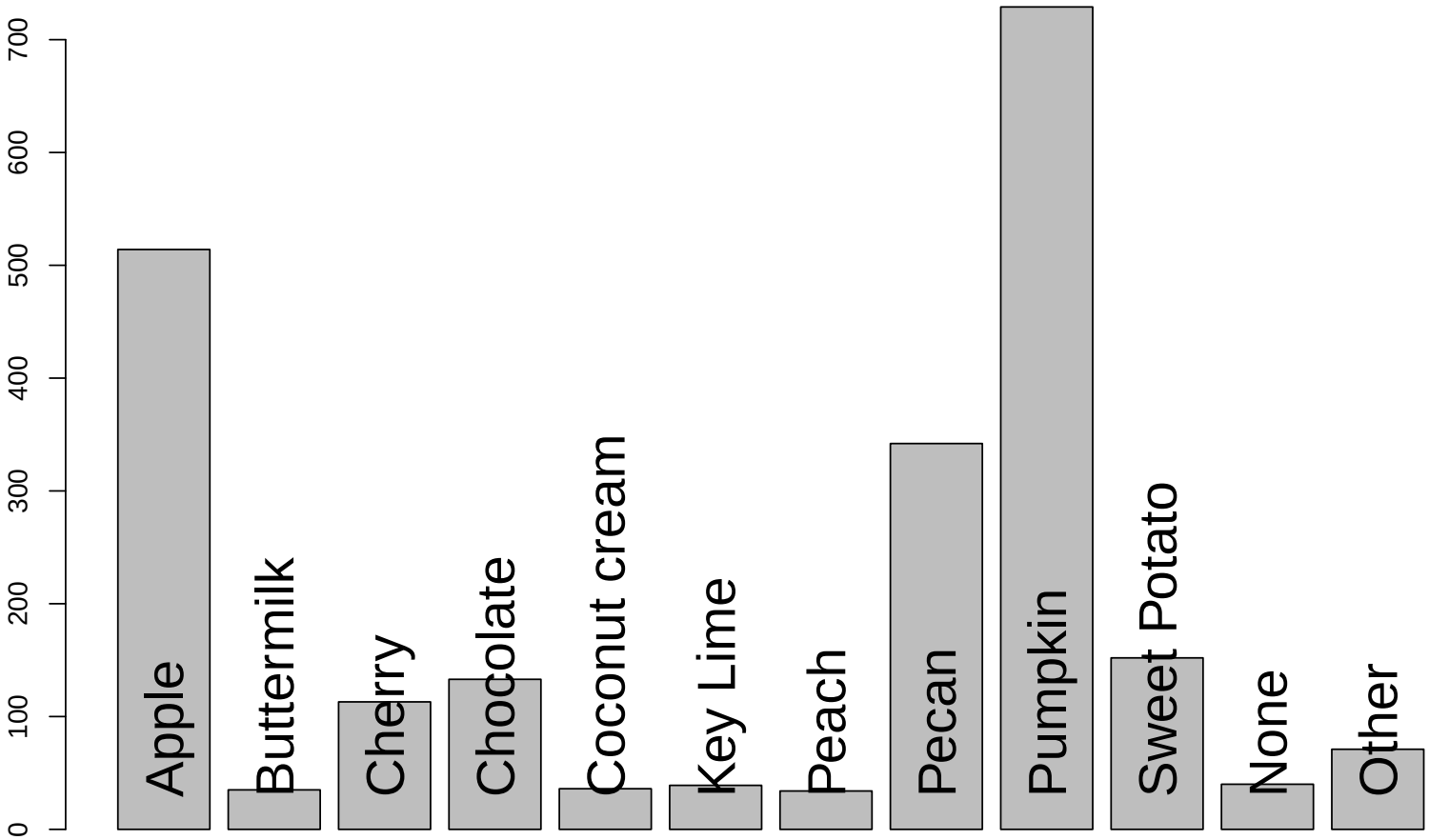
- **Barplot** (Categorical / Discrete variables)
  - each modality is associated to a bar, whose height corresponds to the absolute frequency

- **Histogram** (Continuous variables)
  - each class is associated to a bar whose area correspond to its density

Graphical representation gives us "a feel" for the data, they tell us about the "shape" of the data, their spread and allow us to compare different variables.

Graphical Representations

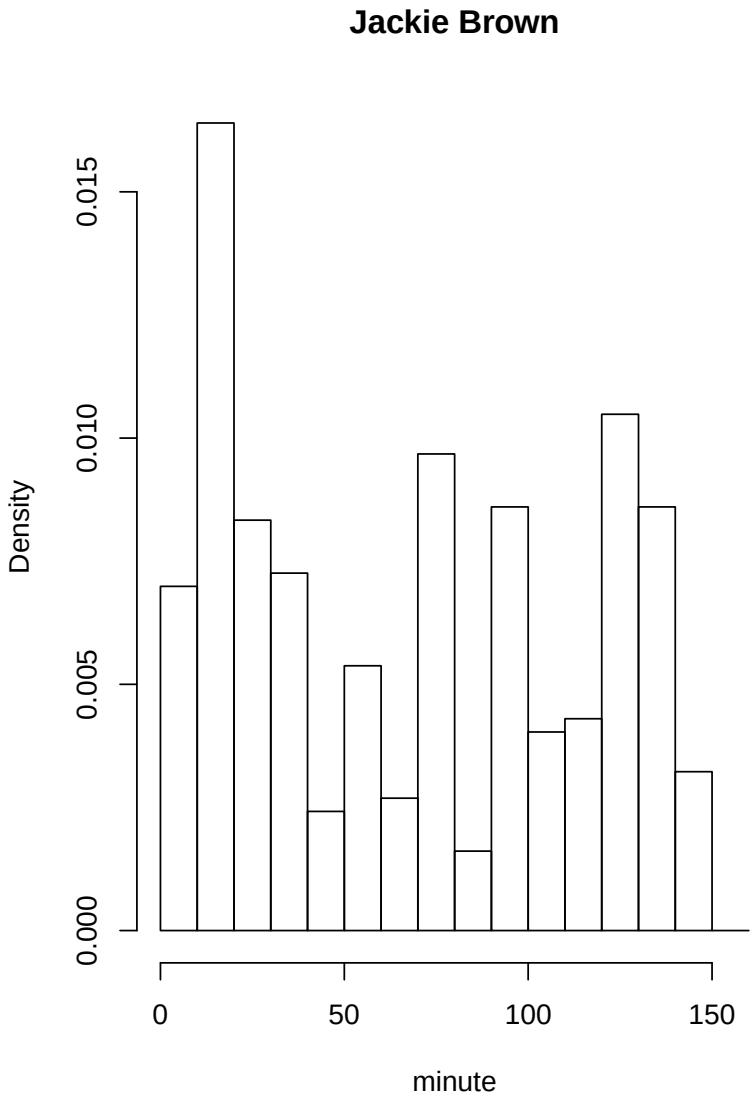
for categorical data / numerical

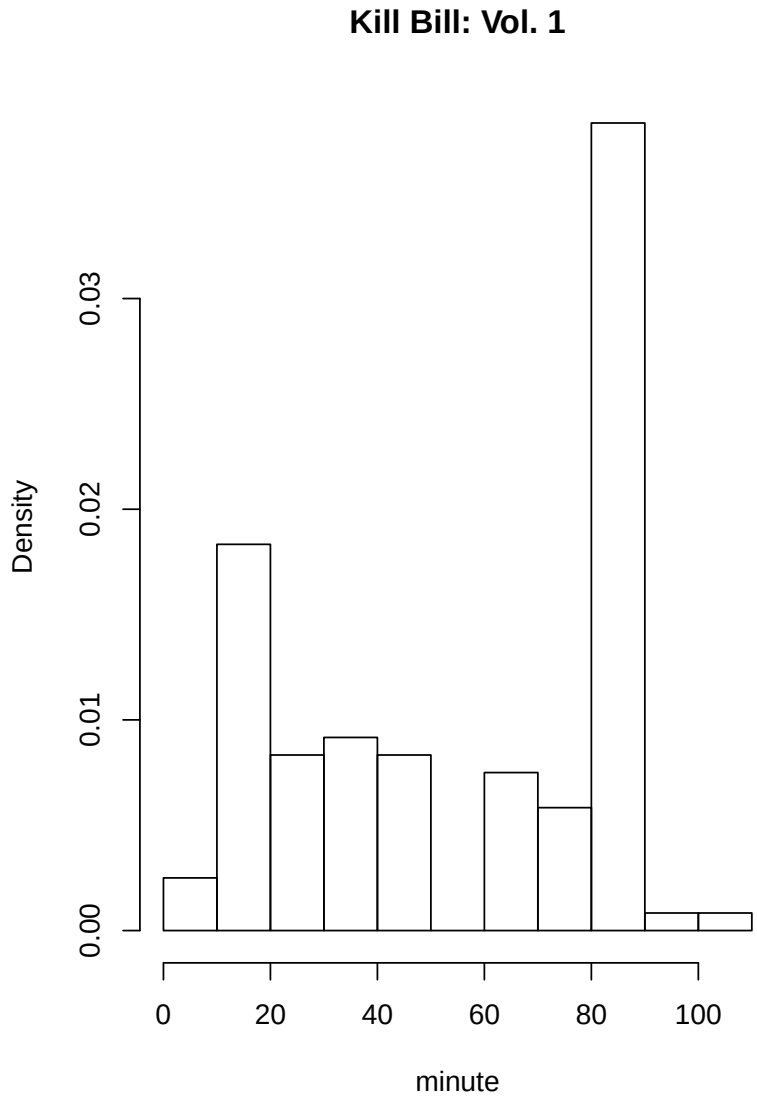


Graphical Representations

for continuous variables

```
## Warning: package 'fivethirtyeight' was built under R version
```





Exercise

Consider the number of shoes that Carlo bought each month of the last year:

	x	n
	0	1
	1	4
	2	3
	3	2
	4	2
Sum		12

- Represent graphically this frequency table, justifying your choice.

Numerical Summaries

If we want a more *concise* summary however we can derive numerical summaries (*statistics*), which describes **a feature** of the variable with one number.

The features we focus on are:

- **centrality**: describing what is a "typical" value for the observations
- **variability**: describing whether the observations take similar values or they are quite different from one another

Measure of Central Tendency

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!

Mode

The **mode** is the modality with the highest observed frequency.

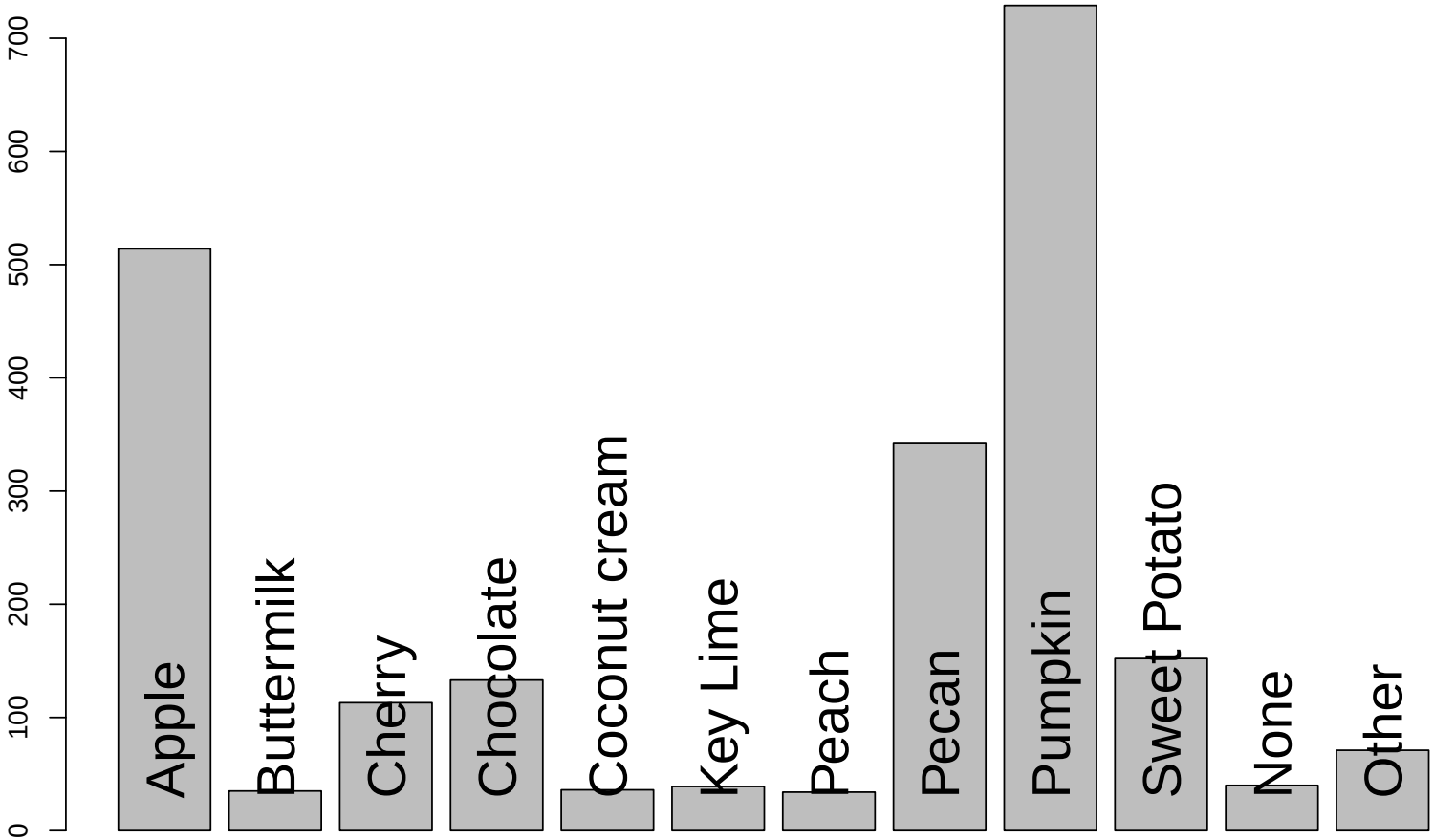
$$y_{MODE} = \{x_j \text{ such that } f_j \geq f_i \ \forall i \neq j\}$$

This is the most general notion of *center* and applies to **all types of data**, both categorical and numerical data.

If data are **grouped** (that is they are divided into classes), then the notion of mode becomes the *modal class*

WATCH OUT: there could be more than one mode

Toy Example


0
100
200
300
400
500
600
700
Apple
Buttermilk
Cherry
Chocolate
Coconut cream
Key Lime
Peach
Pecan
Pumpkin
Sweet Potato
None
Other

Median

The **median** is the middle value of the observations when the observations are ordered (in whichever direction).

Let $y_{(1)}, \ldots, y_{(N)}$ be the *ordered* sample.

- if N is *odd*, the median is the value

$$y_{MED} = y_{(\frac{N+1}{2})}$$

- if N is *even*, the median is the value

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

Median

The median can be computed for all numerical variable and for categorical ordinable variable.

The median is a *center* in the sense that it splits the data in two, half the data below it and half above it

Example:

- Original sample: Good, Good, Bad, Medium, Bad, Bad, Bad
- Ordered sample: Bad, Bad, Bad, **Bad**, Medium, Good, Good

Toy Example

18 28 26 20 28 18 22 22 18 28 18 18 28 28 28 20 24 22 26 22 30

- Ordered sample: 18 18 18 18 18 20 20 22 22 22 **22** 24 26 26 28 28 28 28 28 28 30
- $N = 21$ is odd, take the $(N+1)/2 = 11$-th element
- $y_{MED} = y_{(11)} = 22$

x	n	f	F
18	5	0.24	0.24
20	2	0.1	0.34
22	4	0.19	0.53
24	1	0.05	0.58
26	2	0.1	0.68
28	6	0.29	0.97
30	1	0.05	1

The median is the first modality x_k for which $F_k \geq 0.5$

Quartiles

The median y_{MED} tell us which is the *level* reached by at least 50% of the population.

Its extension, the **percentile of level p**, tell us which is the level reached by at least p*100% of the population, and it is defined as the first modality x_k for which $F_k \geq p$.

The percentiles of level $p = 0.25$ and $p = 0.75$ have a special role and are called **1st** and **3rd quartile** respectively.

Mean

The (arithmetic) **mean** is the sum of the observations divided by the number of observations.

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

It is interpreted as the balance point of the distribution because it minimizes the following quantity

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

Properties

- **internality**:

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

- **linearity**:
  - if $z_i = ay_i + b$ then

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

- **zero-deviation**:

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

- **associativity**:
  - if $\bar{y}, \bar{x}$ are the arithmetic mean of two samples of size N and M, respectively, then the mean of the combined sample is

$$\bar{z} = \frac{N \times \bar{y} + M \times \bar{x}}{N + M}$$

From a Frequency distribution

It is possible to compute the mean **directly from the frequency distribution**, using this alternative definitions:

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

Where x_j are the modalities of the variable.

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

$$c_j = (u_j + l_j)/2 \qquad \Rightarrow \bar{y} = \sum_{j=1}^{K} c_j f_j$$

Toy Example

Salaries of NBA players for the season $2017/2018$ in $100.000\$$:

	x	c	n	f	F	c*f
	[0.172, 34.8)	17.50	336	0.59	0.59	10.26
	[34.8, 69.5)	52.17	76	0.13	0.72	6.92
	[69.5, 104)	86.84	41	0.07	0.79	6.21
	[104, 139)	121.50	35	0.06	0.85	7.42
	[139, 173)	156.17	30	0.05	0.9	8.18
	[173, 208)	190.83	18	0.03	0.93	5.99
	[208, 243)	225.50	19	0.03	0.96	7.48
	[243, 277)	260.16	9	0.02	0.98	4.09
	[277, 312)	294.83	6	0.01	0.99	3.09
	[312, 347)	329.49	2	0.00	0.99	1.15
Sum			572	0.99	8.70	**60.79**

WATCH OUT: This is not necessarily equal to the mean of the disaggregated data (in this case 58.58)

Exercise

NBA Salaries

	x	c	n	f	F	c*f
	[0.172, 34.8)	17.50	336	0.59	0.59	10.26
	[34.8, 69.5)	52.17	76	0.13	0.72	6.92
	[69.5, 104)	86.84	41	0.07	0.79	6.21
	[104, 139)	121.50	35	0.06	0.85	7.42
	[139, 173)	156.17	30	0.05	0.9	8.18
	[173, 208)	190.83	18	0.03	0.93	5.99
	[208, 243)	225.50	19	0.03	0.96	7.48
	[243, 277)	260.16	9	0.02	0.98	4.09
	[277, 312)	294.83	6	0.01	0.99	3.09
	[312, 347)	329.49	2	0.00	0.99	1.15
Sum			572	0.99	8.70	60.79

- Compute median and mode, and compare them with the value of the mean. What do you notice?
- Draw the histogram

Mean vs Median

- The mean takes into account all observations, **including possible anomalous ones**!
- The median takes into account only order of the observations, **regardless of their values**.

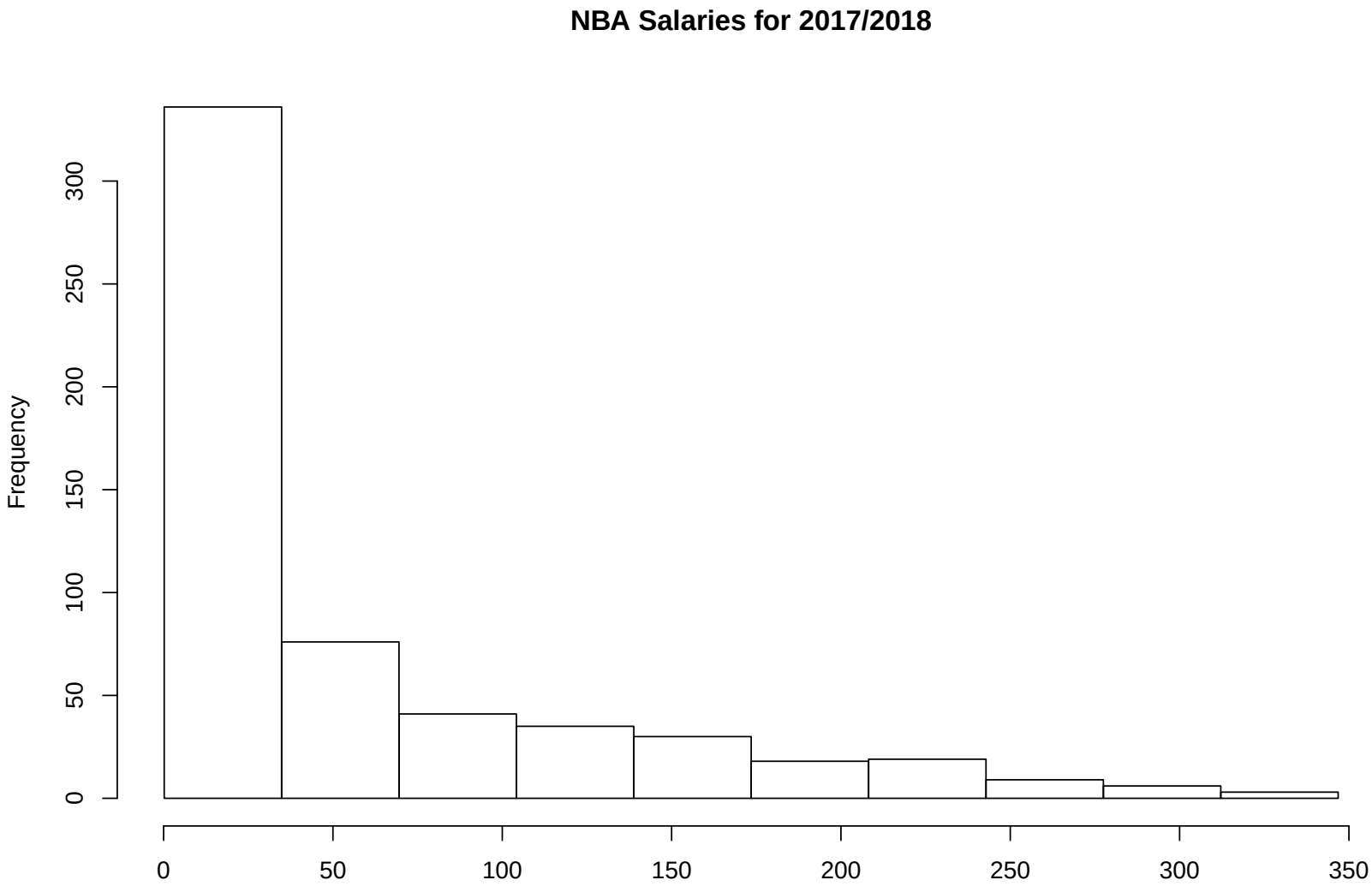
Example:

In the case of the NBA salary data *only very few players* (Stephen Curry, LeBron James) had *very big* pay-checks.

The mean considers them as *regular* players, while the median does not!

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

Toy Example


NBA Salaries for 2017/2018
Frequency
0
50
100
150
200
250
300
0
50
100
150
200
250
300
350

Numerical Summaries

If we want a more *concise* summary however we can derive numerical summaries (*statistics*), which describes **a feature** of the variable with one number.

The features we focus on are:

- **centrality**: describing what is a "typical" value for the observations
- **variability**: describing whether the observations take similar values or they are quite different from one another

Variability

In general

When do we consider observations "similar" between each other?

- **Ranges**:
  - when they varies in a small interval
- **Variance**:
  - when they are all close to the same value (typically the mean)

Ranges

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

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

$$RV = y_{(N)} - y_{(1)}$$

- The **Interquartile Range** is the difference between the $3rd$ and the $1st$ quartile and it gives us the smallest interval in which 50% of the observations lay

$$IQ = y_{Q1} - y_{Q3}$$

This is not affected by anomalous observations.

Variance

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

Problem: To exploit the idea of **deviation from the center** we cannot use just the average of the deviations

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

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

Solution: Square them!

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

Standard deviation

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

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

The standard deviation is on the same scale as the data, making it more interpretable than the variance.

Computing the Variance

- **Strategy 1** Compute directly all the deviations $(y_i - \bar{y})^2$, then average them

- **Strategy 2**: Exploit the alternative definition of the variance

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

In terms of frequency distributions

$$\frac{1}{N}\sum_{i=1}^{N} y_i^2 = \frac{1}{N}\sum_{j=1}^{K} x_j^2 n_j = \sum_{j=1}^{K} x_j^2 f_j$$

Toy example

	x	n	f	F	xf	x^2	x^2f
	18	5	0.24	0.24	4.29	324	77.14
	20	2	0.1	0.34	1.9	400	38.1
	22	4	0.19	0.53	4.19	484	92.19
	24	1	0.05	0.58	1.14	576	27.43
	26	2	0.1	0.68	2.48	676	64.38
	28	6	0.29	0.97	8	784	224
	30	1	0.05	1	1.43	900	42.86
Sum			1.02	4.34	23.43	4144	566.1

- $S^2 = 566.1 - 23.43^2 = 17.131$
- $S = \sqrt{17.131} = 4.14$

Exercise

Consider (again) the number of shoes that Carlo bought each month of the last year:

	x_i	n_i
	0	1
	1	4
	2	3
	3	2
	4	2
Sum		12

- Determine the mean, the median and the mode of the observed variable. If not possible, explain why.
- Plot the data, motivating your choice of graph.
- Compute interquartile range and sample variance, commenting on the difference between these two measures of variability.

Two-variables

So far we have only dealt with how to describe, represent and summarize **a single** variable.

In real life, however, we usually have more than one variable measured on each unit, that may or may not be related.

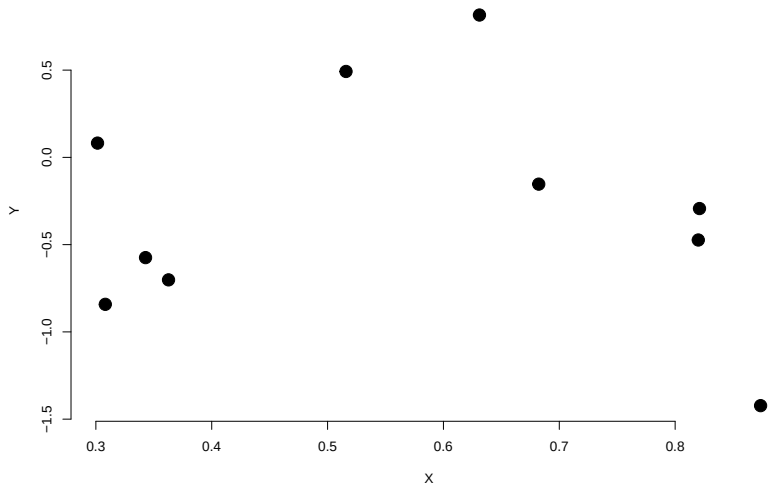
Examples: weight and height of a subject, length and budget of a movie, hair color and first name of a subject.

CAVEAT: We focus on the case of both variables being numerical (although association measures exist for categorical and mixed type variables as well).

Graphical representation

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

- each axis correspond to one of the two variables X, Y
- each point represent a units and its coordinates correspond to values of the two variables observed on it



Association

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

- **Positive** Association: as x goes up, y tends to go up.
- **Negative** Association: as x goes up, y tends to go down.

The **covariance** is an indicator measuring the strength of the association between two variables:

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

Limits of the Covariance:

it is not all fun and games

- the covariance measure linear association, i.e. the case where the relationship between the two variables x and y is of the type

$$y_i = ax_i + b$$

- the covariance depends on the scale of the data. We have no general reference to determine if the observed covariance is *large* or *small*

Correlation

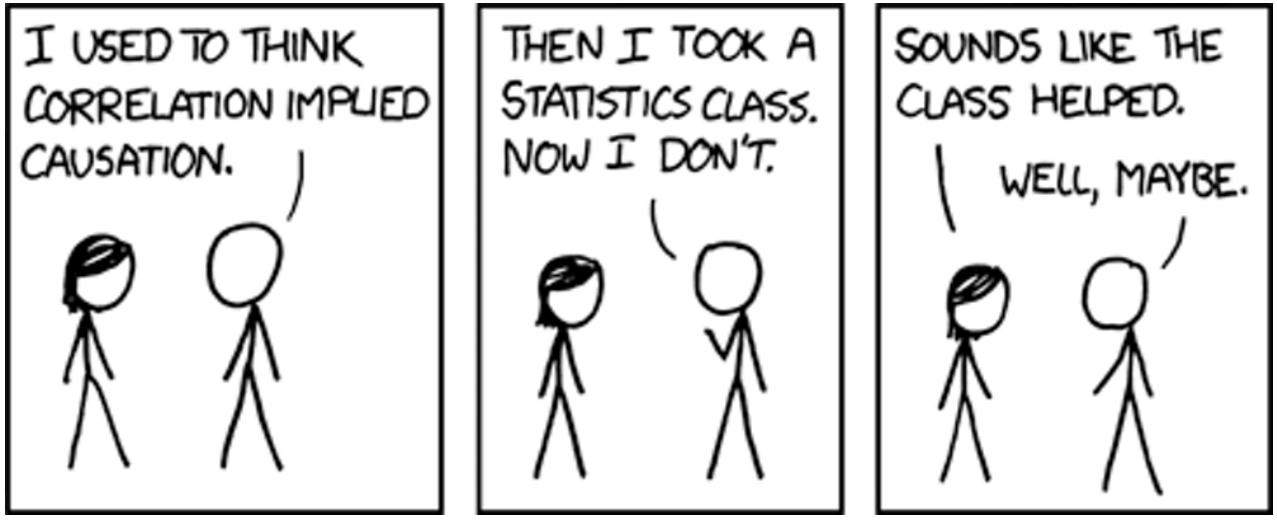
The **correlation** is a rescaled version of the covariance:

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

- the correlation is between -1 and 1; the closer $|r_{x,y}|$ is to 1, the stronger 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
- correlation is **symmetric** with respect to the two variables, i.e. it does not treat favorably one variable over the other

Limit of Correlation

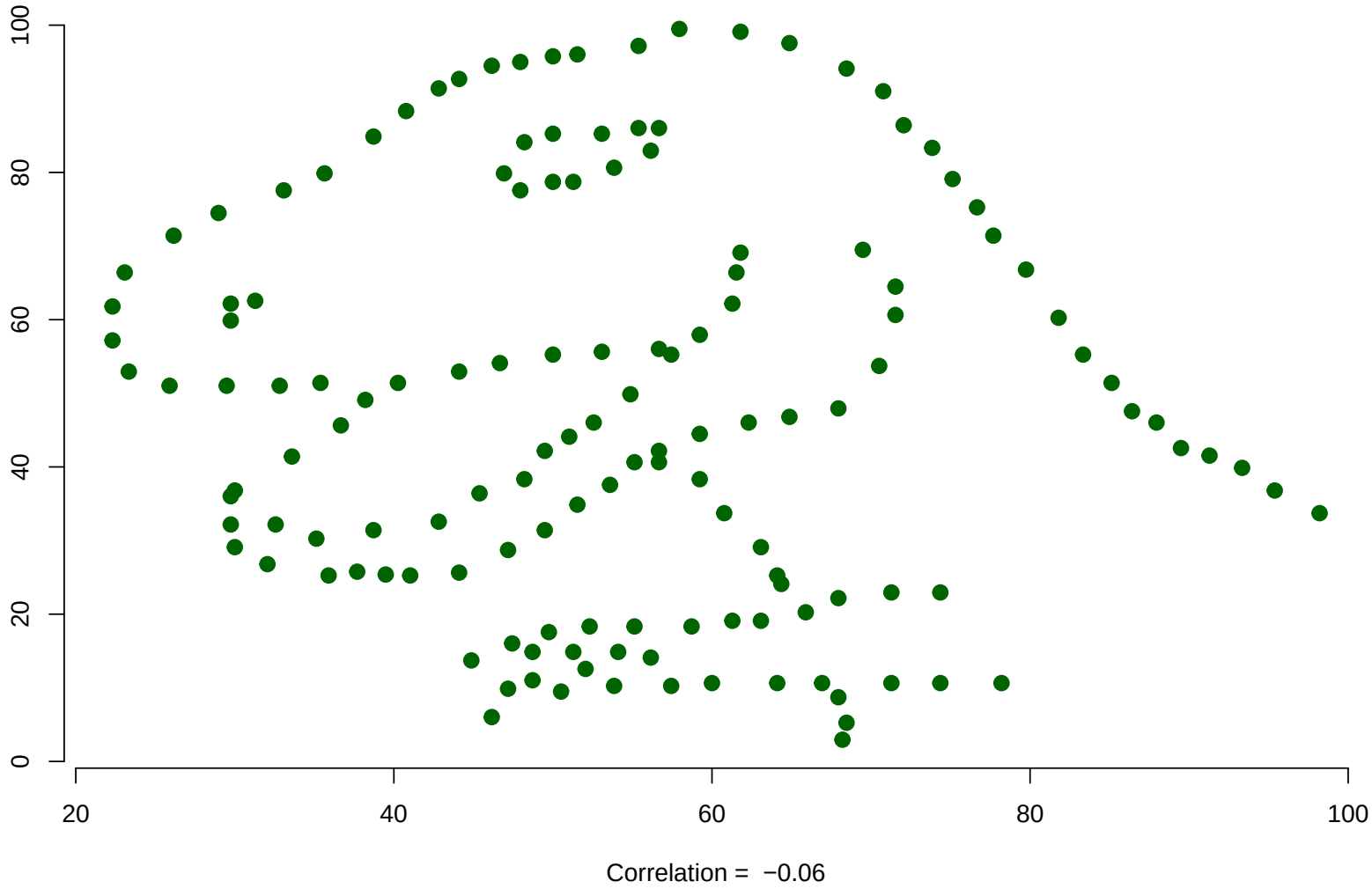
it is *still* not all fun and games



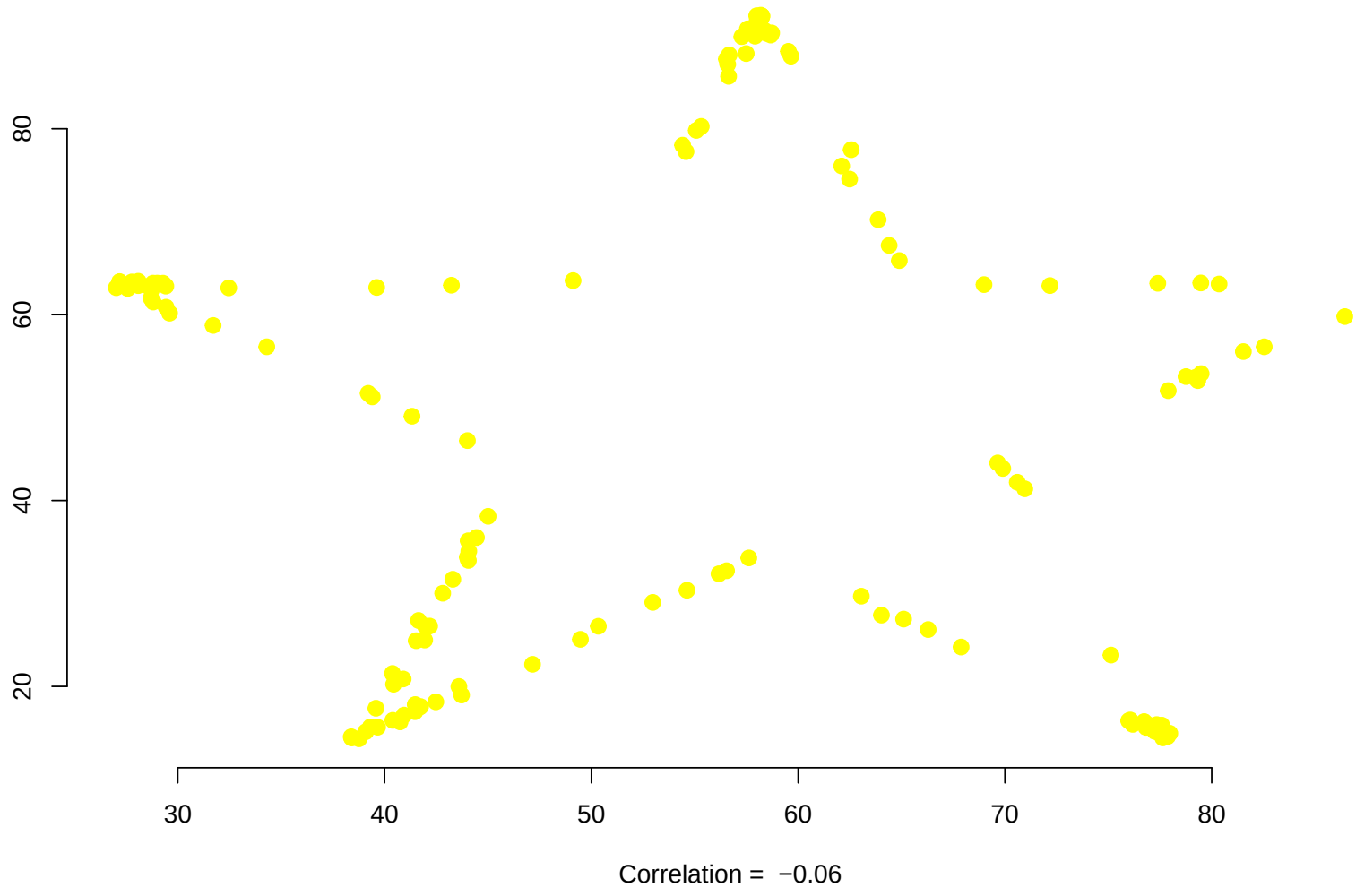
- Correlation still measure **only linear dependence**. That is $|r_{x,y}| = 1$ when data lay on a straight line. However, if there is some more complicated form of dependence, even something simple like $y_i = ax_i^2 + b$, correlation may not capture it.
- **Correlation does not mean causation!** There may be a strong correlation between two variables even when there is no relation between them.

Datasaurus

```
## Warning: package 'datasauRus' was built under R version 3.6.3
```

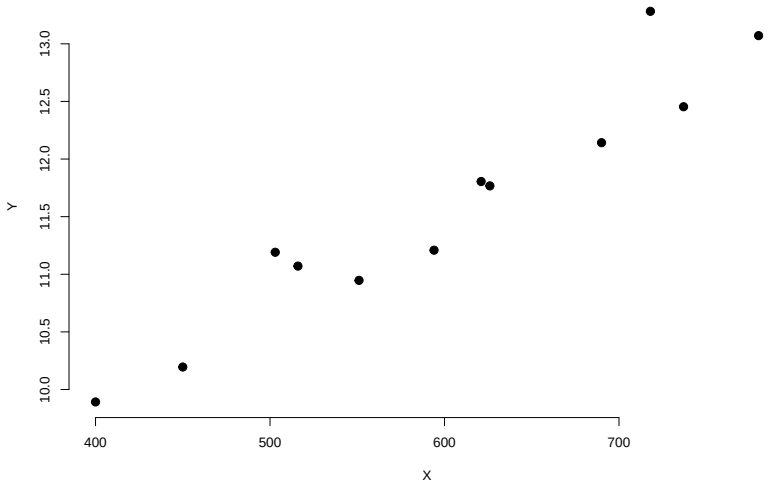


Datasaurus



Spurious Correlation

- X = People who died by falling out of their bed
- Y = Lawyers in Puerto Rico
- $r_{x,y} = 0.957087$



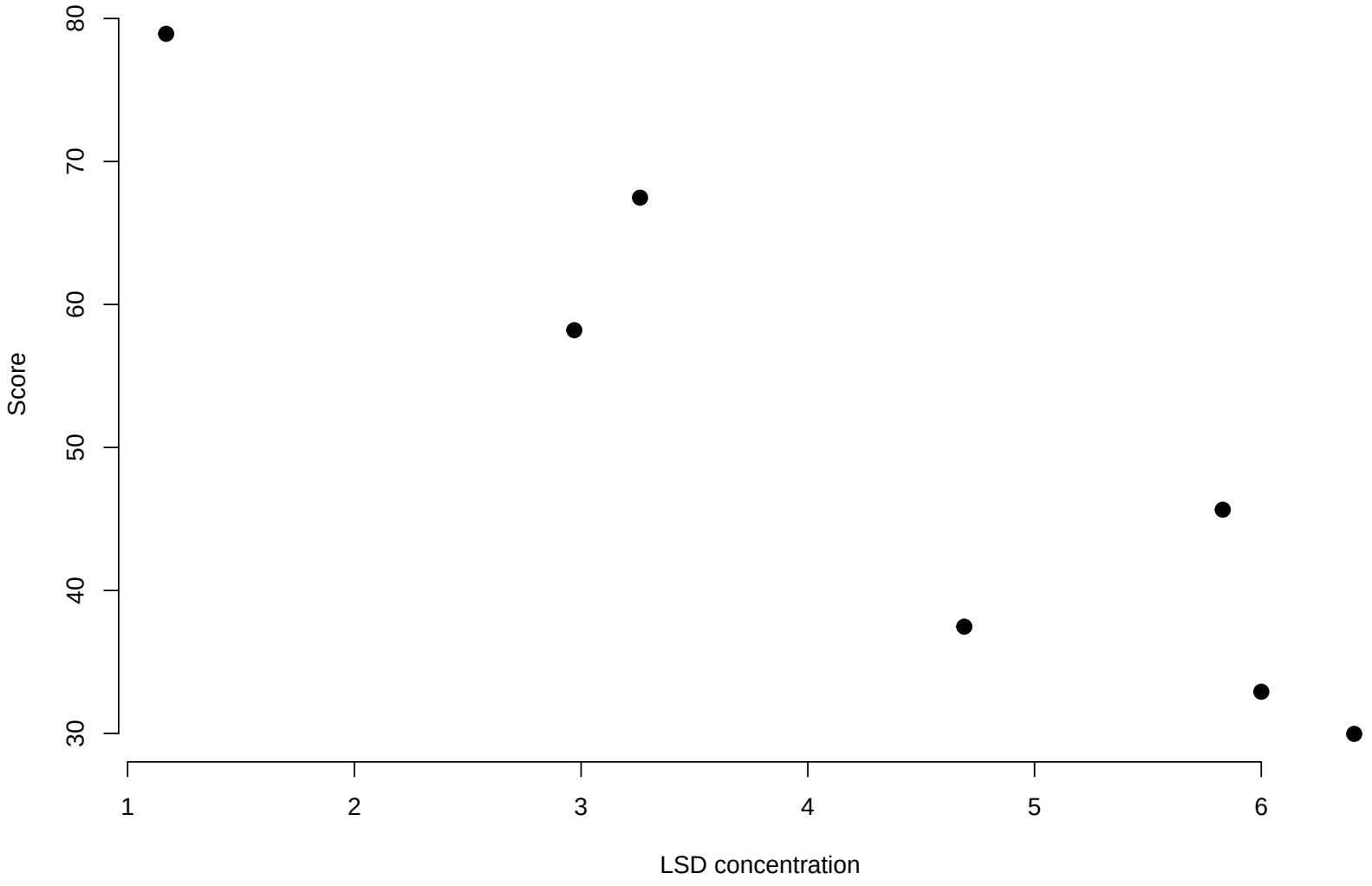
Exercise

Does LSD help with math?

- X: Tissue Concentration of Lysergic Acid Diethylamide (LSD)
- Y: Math Test Scores
- **?** Compute the covariance and the correlation between X and Y

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



Solution

	X	Y	X^2	Y^2	XY
	1.17	78.93	1.37	6229.94	1.37
	2.97	58.20	8.82	3387.24	8.82
	3.26	67.47	10.63	4552.20	10.63
	4.69	37.47	22.00	1404.00	22.00
	5.83	45.65	33.99	2083.92	33.99
	6.00	32.92	36.00	1083.73	36.00
	6.41	29.97	41.09	898.20	41.09
Sum	30.33	350.61	153.89	19639.24	153.89

$$\bar{X} = 4.3 \qquad \bar{Y} = 50.1$$

$$s_X = 1.66 \qquad s_X = 15.95$$

$$cov_{x,y} = -28.92 \qquad r_{x,y} = -0.936$$