NOVA IMS INFORMATION MANAGEMENT SCHOOL

Basic concepts
Exploring data

Statistics for Data Science

Bruno Damásio
bdamasio@novaims.unl.pt
@bmpdamasio

2025/2026

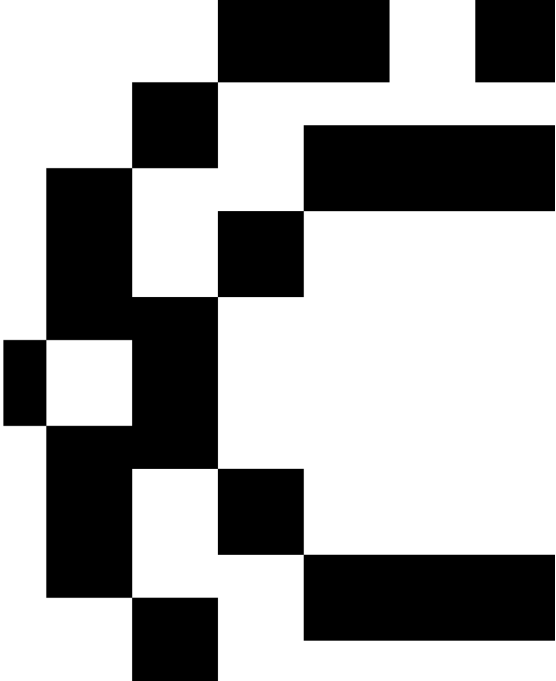


Table of contents

Preliminary concepts

What is Statistics?

The art and science of answering questions and exploring ideas through the processes of gathering data, describing data, and making generalizations about a population on the basis of a smaller sample.

Basic concepts

- When conducting a research study, information is collected concerning **units**;
- A **variable** is a characteristic that is measured and can take on different values. In other words, something that can vary. This is in contrast to a constant which is the same for all units in a study.

Units

The basic objects on which the data is collected.

Variable

Characteristic of units that can take on different values (in other words, something that can vary).

Example: SfDS Student data

- Data are collected from a sample of SfDS students;
- Each student's lab assignment and exam grade are recorded;
- In this example, the units are the students;
- The variables are the lab assignment and exame grades.

Categorical and quantitative variables

Variables can be classified as **categorical** or **quantitative**.

- **Categorical variables** refer to variables with values that can't be quantifiable. These can be classified as:
  - **Nominal**, which describes a name, label or category without a logical order (e.g., gender);
  - **Ordinal**, whose values are defined by an order between the different categories (e.g., education level).

Categorical and quantitative variables

- **Quantitative variables** have numerical values with magnitudes that can be placed in a meaningful order with consistent intervals, also known as numerical. Quantitative variables can be:
  - **Discrete**, which means that can only take on a set number of values (e.g., only whole numbers);
  - **Continuous**, which means that can take on any value and any value between values (e.g., out to an infinite number of decimal places).

Categorical and quantitative variables

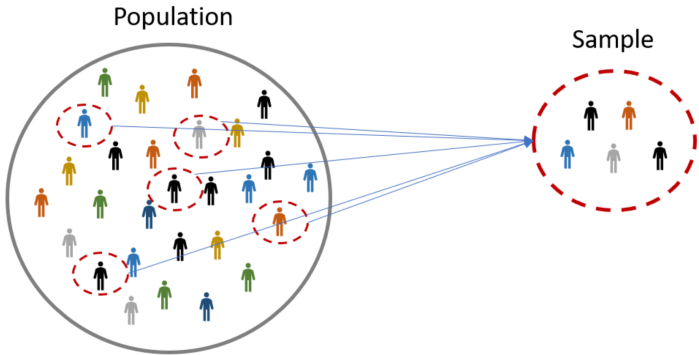
Categorical or Quantitative?

- Weight
- Favorite ice cream flavor
- Children per household
- Running Distance
- Religion
- Satisfaction rating

Samples and Populations

- We often have questions concerning large **populations**;
- Gathering information from the entire population is not always possible due to barriers such as time, accessibility, or cost;
- Instead of gathering information from the whole population, we often gather information from a smaller subset of the population, known as a **sample**;
- Values concerning a sample are referred to as sample **statistics** while values concerning a population are referred to as population **parameters**.

Samples and Populations



Example: Student Housing

- A survey is carried out at NOVA University Campi to estimate the proportion of all undergraduate students living at home during the current semester;
- Of the 3,838 graduate students enrolled at the NOVA University Campi, a random sample of 100 was surveyed;
  - **Population**: all 3,838 graduate students at NOVA University Campi;
  - **Sample**: the 100 undergraduate students surveyed;
- We can use the data collected from the sample of 100 students to make inferences about the population of all 3,838 students.

Sampling Bias

- Recall the entire group of individuals of interest is called the population. It may be unrealistic or even impossible to gather data from the entire population. The subset of the population from which data are actually gathered is the sample;
- A sample should be selected from a population randomly, otherwise it may be prone to **bias**;
- Our goal is to obtain a sample that is **representative** of the population.

Sampling Bias

Systematic favoring of certain outcomes due to the methods employed to obtain the sample.

Example: Weight Loss Study Volunteers

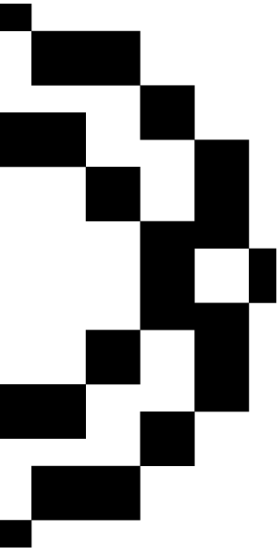
- A medical research center is testing a new weight loss treatment;
- They advertise on a social media site that they are looking for volunteers to participate;
- There is sampling bias because the sample will be limited to people who use the social media site where they advertised:
  - The individuals who choose to participate may be different from the overall population;
  - Volunteers may be individuals who are already actively trying to lose weight;
- This is not a representative sample because the sample may have characteristics that are different from the population of interest.

Simple Random Sampling

- To prevent sampling bias and obtain a representative sample, a sample should be selected using a probability-based sampling design which gives each individual a known chance of being selected.
- The most common probability-based sampling method is the **simple random sampling method**.

Simple random sampling

A method of obtaining a sample from a population in which every member of the population has an equal chance of being selected.



Describing and Visualizing Data

Describing: Proportion and Frequency

Data concerning **one categorical variable** can be summarized using a proportion:

$$p = \frac{\text{Number in the category}}{\text{Total number}} \tag{1}$$

Example: Star Wars Datasets

Consider the `starwars` dataset from the package dplyr:

```
library(dplyr)
data("starwars")
starwars
```

```
## # A tibble: 87 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Luke Sk~    172    77 blond      fair       blue            19   male  mascu~
##  2 C-3PO       167    75 <NA>       gold       yellow         112   none  mascu~
##  3 R2-D2        96    32 <NA>       white, bl~ red             33   none  mascu~
##  4 Darth V~    202   136 none       white      yellow          41.9 male  mascu~
##  5 Leia Or~    150    49 brown      light      brown           19   fema~ femin~
##  6 Owen La~    178   120 brown, gr~ light      blue            52   male  mascu~
##  7 Beru Wh~    165    75 brown      light      blue            47   fema~ femin~
##  8 R5-D4        97    32 <NA>       white, red red             NA   none  mascu~
##  9 Biggs D~    183    84 black      light      brown           24   male  mascu~
## 10 Obi-Wan~    182    77 auburn, w~ fair       blue-gray       57   male  mascu~
## # i 77 more rows
## # i 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

This contains information regarding the star wars characters.

Example: Star Wars Datasets

To obtain the frequency of the gender of all characters:

```
freq_table <- table(starwars$gender)
freq_table
```

```
##
##  feminine masculine
##        17        66
```

To obtain the proportion:

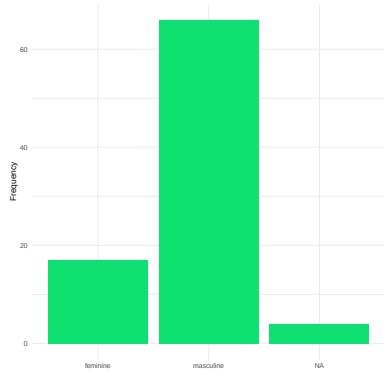
```
prop.table(freq_table)
```

```
##
##  feminine masculine
## 0.2048193 0.7951807
```

Visualize: Bar Chart

To visualize these statistics, we can use a bar chart:

```
library(ggplot2)
ggplot(starwars, aes(x=gender)) +
  geom_bar(fill='#0ee071') +
  labs(x='', y='Frequency') +
  theme_minimal()
```



Summaries of Data

- Central tendency measures describe the "center" around which the data is distributed.
- Variability measures describe "data spread" or how far away the measurements are from the center.
- Relative Standing measures describe the relative position of specific measurements in the data.

Central tendency measures

- **Quantitative variables** are often summarized using numbers to communicate their central tendency.
- The mean and median are two of the most commonly used measures of central tendency.

Central tendency measures

Mean

The numerical average; calculated as the sum of all of the data values divided by the number of values.

The sample mean is represented as $\bar{x}$ ("x-bar") and the population mean is denoted as the Greek letter μ ("mu").

Median

The middle of the distribution that has been ordered from smallest to largest; for distributions with an even number of values, this is the mean of the two middle values.

Importante notes

The median is more resistant to outliers (i.e., extreme odd values) compared to the mean.

Example: Star Wars Datasets

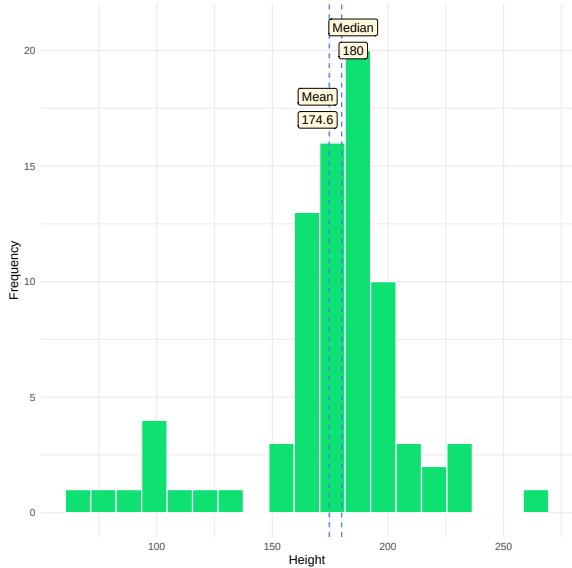
To obtain the mean of the height of the star wars characters:

```
mean(starwars$height, na.rm=TRUE)

## [1] 174.6049
```

To obtain the median:

```
median(starwars$height, na.rm=TRUE)

## [1] 180
```

Visualize: Histogram



Variability measures

- **Variance** and **standard deviation** are measures of variability.
- The standard deviation is the most commonly used measure of variability when data are quantitative and approximately normally distributed.
- When computing the standard deviation by hand, it is necessary to first compute the variance.
- The standard deviation is equal to the square root of the variance.

Standard Deviation

Standard Deviation

Roughly the average difference between individual data values and the mean.

The standard deviation of a sample is denoted as s.

The standard deviation of a population is denoted as σ.

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

Example: Star Wars Datasets

To obtain the variance of the height of the star wars characters:

```
var(starwars$height, na.rm=TRUE)

## [1] 1209.242
```

To obtain the standard deviation:

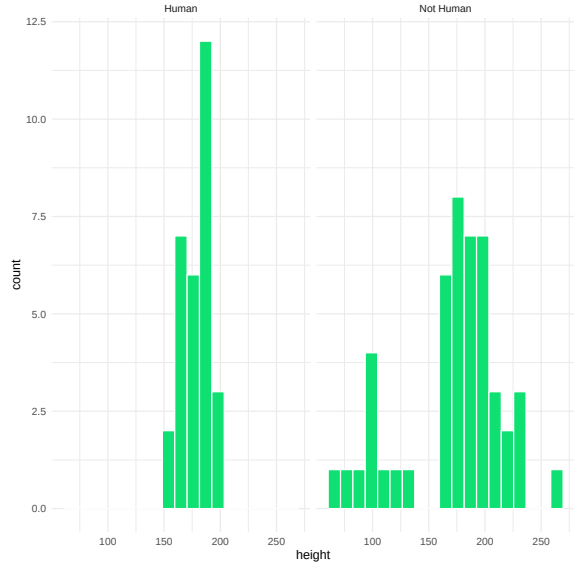
```
sd(starwars$height, na.rm=TRUE)

## [1] 34.77416
```

Example: Star Wars Datasets

The standard deviation according to whether the character is human or not:

```
## # A tibble: 2 x 2
##   human     standard_deviation
##   <chr>                  <dbl>
## 1 Human                   12.0
## 2 Not Human               44.6
```

Visualize: Histogram



Relative standing measures

- Another useful measure is the **percentile**. In general, the n^{th} percentile is a value such that n% of the observations fall at or below it.
- Percentiles divide ordered data into hundredths.
- So, the 25th percentile, corresponds to the value where 25% of the observations fall at or below it.

Relative standing measures

- Instead of dividing into hundredths, we can divide ordered data into quarters. In this setting, we have **quartiles**.
- The first quartile, Q_1 is the value for which 25% of the observations are smaller and 75% are larger.
- Q_2 is the same as the median (50% are smaller, 50% are larger).
- Finally, only 25% of the observations are greater than the third quartile, Q_3.

Example: Star Wars Datasets

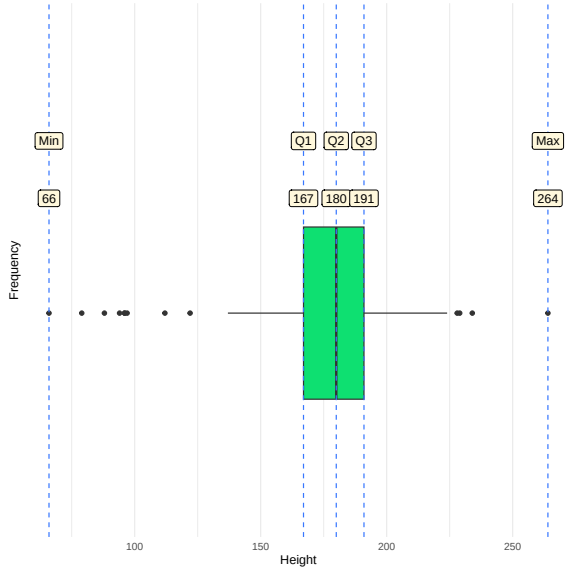
To obtain the percentiles:

```
quantile(starwars$height, probs=seq(0,1,0.1), na.rm=TRUE)

##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##   66  122  165  170  177  180  183  188  193  206  264
```

To obtain quartiles and some summary statistics:

```
summary(starwars$height)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    66.0   167.0   180.0   174.6   191.0   264.0       6
```

Visualize: Boxplots



Correlation

- To describe the strength and direction of association between two quantitative variables, one can use the *correlation*.
- The correlation between two quantitative variables of a *sample* is denoted r.
- The correlation between two quantitative variables of a *population* is denoted ρ, which is the Greek letter "rho".

Correlation

A measure of the direction and strength of the relationship between two variables.

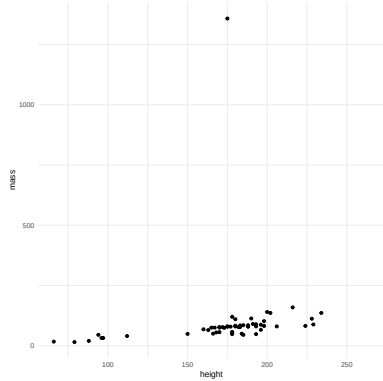
Properties of correlation

1. $-1 \leq \rho \leq +1$.
2. For a positive association, $\rho > 0$, for a negative association $\rho < 0$, if there is no relationship $\rho = 0$.
3. The closer ρ is to 0 the weaker the relationship and the closer to $+1$ or -1 the stronger the relationship (e.g., $\rho = -0.88$ is a stronger relationship than $\rho = +0.60$).
4. The sign of the correlation provides direction only.
5. Correlation is unit free; the x and y variables do NOT need to be on the same scale (e.g., it is possible to compute the correlation between height in centimeters and weight in pounds).
6. It does not matter which variable you label as x and which you label as y. The correlation between x and y is equal to the correlation between y and x.

Visualize: Scatter Plot

Let's start by analysing the scatter plot of height against mass of star wars' characters:

```
ggplot(starwars, aes(x=height, y = mass)) +
  geom_point() +
  theme_minimal()
```



Visualize: Scatter Plot

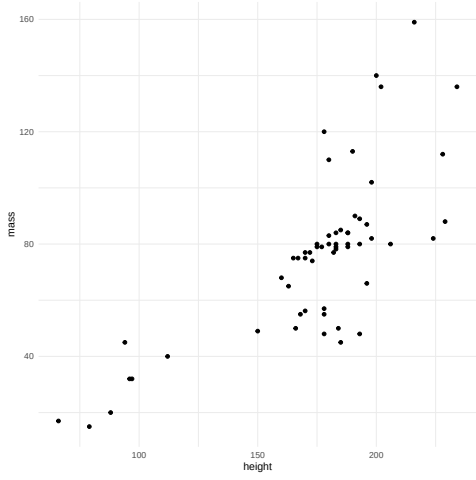
It is a bit difficult to read, due to that extreme outlier:

```
starwars %>%
  filter(mass == max(mass, na.rm=TRUE)) %>%
  select(name, height, mass)


## # A tibble: 1 x 3
##   name                  height  mass
##   <chr>                  <int> <dbl>
## 1 Jabba Desilijic Tiure    175  1358
```

Let's remove it...

Visualize: Scatter Plot

```
ggplot(starwars %>%  filter(mass != max(mass, na.rm = TRUE)), aes(x=height, y = mass)) +
  geom_point() +
  theme_minimal()
```



Example: Star Wars Dataset

To assess the correlation:

```
#Remove NA's
height = starwars$height[!is.na(starwars$height) & !is.na(starwars$mass)]
mass = starwars$mass[!is.na(starwars$height) & !is.na(starwars$mass)]

#Calculate correlation, without outlier removal
cor(height, mass)


## [1] 0.130859


#Calculate correlation, after outlier removal
height = height[mass < max(mass)]
mass = mass[mass < max(mass)]

cor(height, mass)


## [1] 0.7508582
```