DATA SCIENCE & ARTIFICIAL INTELLIGENCE

SCIENTIFIC & DATA-INTENSIVE COMPUTING

UNIVERSITÀ DEGLI STUDI DI TRIESTE

Descriptive Data Analysis

Introductory course on Statistics and Probability

Nicholas A. Pearson

Università degli Studi di Trieste

September 8, 2025

Statistics

Statistics is the science that deals with collecting data and extracting information/knowledge from it.

Data can help us understand a phenomena, but it is necessary to collect the necessary data and to do it well and the data must then be examined to isolate and highlight the information.

Starting from a question about a phenomenon, statistics is involved in determining which data can be used to answer that question, and, if the data is not already available, how it should be collected.

This is followed by a phase in which the data is analyzed to extract the information.

Statistics

Briefly, the steps of a statistical analysis are:

- *formulate a question:* translate a cognitive need into a form that can be answered in statistical terms;
- *identify or collect data:* is a broad topic, which goes by the name of experimental design and sampling; it is based on probability theory;
- *organize and look at the data:* from the collected data, it is not always straightforward to extract the required information; however, you can either synthesize it appropriately and/or represent it graphically, depending on the information sought;
- *formulate a model:* explain the observed data, based on the assumptions made about the phenomenon. The model is then estimated using the available data. This model can be utilized to either confirm or reject hypotheses about the phenomenon or to make predictions for future occurrences.

Population

The final aim of the statistical analysis is to know some aspects of interest for a **population**.

A population is an entire group, a complete set of units. Each component of the whole population is known as **statistical unit**.

- **Examples** of population:
  - the population of Italian men aged 18 years old on 01/01/2012;
  - Italian families at 01/01/2012;
  - 50 major urban centres spread across Europe.
- The population can be *finite* (i.e. Italian population) or *infinite* (i.e. all the people affected by a disease, in the past and the future).

To conduct a statistical analysis we collect **variables** on the statistical units.

Statistical Unit

The same statistical unit can be a part of multiple populations.
A student of a scientific lyceum can be a component of several populations, due to the context in which he/she is observed:

- the student is an element of the population of pupils attending a given school
- the student is part of the population of people living in the same municipality on a given date
- he/she can belong to the population of young people in the given age range
- if we refer to the context of Italian secondary schools, he can no longer be considered a statistical unit. In this case, the statistical units are the individual schools (school population), where the variable can be the number of pupils enrolled in a given year.

Sampling

When we cannot, or do not want to, observe the entire population we can use a sampling procedure.

Sampling and Inference

We observe just a part of the population, the **sample**, and we generalize what was observed on the sample to the whole population.

- Sampling is natural
- Let's imagine that we are cooking soup: to get an idea of the possible success, we take a taste.
- Tasting a spoon of soup and deciding if there is enough salt, you are doing *exploratory-descriptive analysis*.
- If you conclude that the whole soup is tasteless from your small taste, you are doing *inference*.

Sampling and representativeness

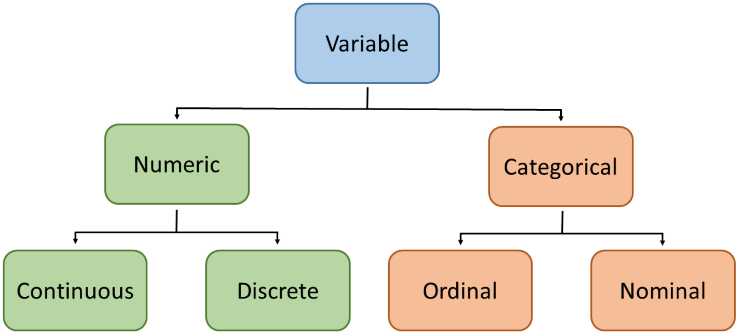
- For the inference to be valid, the spoon we taste must be **representative** of the entire preparation (soup).
- If we add the salt first, then all the ingredients, never mix and taste the soup on the surface, we probably don't have a **representative** sample.
- If we throw in the salt first, then all the ingredients, and mix all the ingredients well before tasting, probably the representativeness of the sample is better.

Basic terminology

A statistic data is the result of the survey (measurement/observation) of some characteristic (variables) on a statistical unit belonging to a population.

- **Statistical unit**: the individual component of the population;
- **Variable**: the elementary aspect surveyed on the statistical units of the population;
- **Modalities/categories of a variable**: the different ways in which the variable presents itself in the statistical units;
- **Support**: the (theoretical) set of modalities of a variable.

Types of Variables



Categorical variables

A variable is **categorical** if the data that it represents can be divided into specific categories or groups, where each category is distinct and separate from the others.

- a qualitative variable is *nominal* if its modalities do not imply an order;
- a qualitative variable is *ordinal* if its modalities imply an order;

Examples:

- Categorical nominal:
  - Eye colour (brown, blue, green, ...)
  - Car brands (FIAT, Ford, Honda, Toyota, BMW, ...)
- Categorical ordinal:
  - Education level (high school, bachelor's, master's, PhD)
  - Pain level (mild, moderate, severe)

Numerical Variables

A **quantitative** (or numerical) variable is a type of variable that represents numerical data where the numbers have mathematical meaning and can be measured or counted.

Quantitative variables can be:

- *Discrete*: if the set of its modalities is finite or countable
- *Continuous*: if the set of its modalities is a range, limited or unlimited.

We can also distinguish based on the measurement scale:

- *Interval*: ordered values with meaningful differences, but without a true zero
- *Ratio*: ordered values with meaningful differences and a true zero

Examples

- *Discrete ratio*:
  How often have you been to the cinema in the last three months?
- *Continuous ratio*:
  What is your height (cm)?
- *Discrete interval*:
  In what year were you born?
- *Continuous interval*:
  External temperature

Data Matrix

The data is organized in a matrix, for example, the data to construct the gap-minder graph looks like this:

variable ↓

	Country	**time**	**gdppc**	**lifexp**	**Region**	
1	Albania	1980	1061	70	Europe	← statistical units
2	Albania	1990	978	74	Europe	
3	Albania	1991	688	73	Europe	
4	Albania	1992	643	73	Europe	
5	Albania	1993	714	73	Europe	
6	Albania	1994	785	74	Europe	
⋮	⋮	⋮	⋮	⋮	⋮	
N-1	Zimbabwe	2010	323	50	Africa	
N	Zimbabwe	2011	348	52	Africa	

lifexp	Region
70	Europe
74	Europe
73	Europe
73	Europe
73	Europe
74	Europe
⋮	⋮
50	Africa
52	Africa

Sometimes, it is convenient to code categorical variables, such as Region, using 0/1 variables. Here we have 4 possible categories: The Americas, Europe, Africa and Asia.

$\longrightarrow$

Thus, we replace the Region variable with 4 variables: Am (yes/no), Eu (yes/no), Af (yes/no) and As (yes/no).

lifexp	Am	Eu	Af	As
70	0	1	0	0
74	0	1	0	0
73	0	1	0	0
73	0	1	0	0
73	0	1	0	0
74	0	1	0	0
⋮	⋮	⋮	⋮	⋮
50	0	0	1	0
52	0	0	1	0

0/1 variables are usually called indicator variables (because they indicate the presence/absence of a mode), binary variables, or dummy variables (they are not truly variables).

- Looking closer, we note that our last coding is redundant because we know that, in our study, if a country is not in Americas (Am), Europe (Eu) and Africa (Af), can only be in Asia (As).
- Then only 3 of the 4 variables are enough to encode the region.
- Here we code the modality "yes" as 1.

To encode a categorical variable with k categories we need $k-1$ indicator variables. The choice of using 0 and 1 is conventional.

lifexp	Am	Eu	Af
70	0	1	0
74	0	1	0
73	0	1	0
73	0	1	0
73	0	1	0
74	0	1	0
⋮	⋮	⋮	⋮
50	0	0	1
52	0	0	1

Statistical Distributions

Consider a set of N statistical units on which we observe the variable Y. We call **distribution** of the variable Y the set of observations (represented by numbers or verbal expressions) relating to the N units of the set of data. In symbols, the distribution will be denoted as:

$$y_1, y_2, \ldots, y_N$$

where y_1 is the observation related to the unit identified by the number 1, y_2 is the one related to the unit identified by the number 2, and so on.

Sometimes this is referred to as a raw distribution, which does not offer a clear overview.

Distribution

For example, for the variable **Gender** the distribution is:

Male Male Male Male Male Female Female Female Male Male Female Female Male Male Female Female
Female Male Male Female Female Male Female Female Male Female Female Female Female Male Male Male
Male Female Male Male Female Male Male Female Female Male Female Male Female Male Male Female Male
Male Male Male Male Male Male Female Male Male Female Male Female Male Female Female Female Male
Male Female Female Female Male Male Female Male Male Male Male Female Male Female Male Male Female
Male Female Female Female Male Female Female Male Female Male Female Female Male Female Female
Male Male Female Male Female Male Male Male Male Female Male Female Male Male Male Male Female
Male Male Male Male Female Female Male Female Female Female Female Male Male Male Female Female
Female Male Male Male Female Male Female Female Female Female Male Male Male Female Male Female
Male Female Male Male Female Female Male Female Female Female Female Male Female Male Female Male
Female Female Male Male Female Male Male Male Female Female Female Male Female Female Male Male
Male Male Female Male Female Female Male Male Male Female Male Male Female Male Male Male Female
Female Female Female Female Female Male Male Male Male Female Female Male Male Male Female Male
Female Male Male Female Male Female Female Female Female Male Male Male Female Female Male Male
Male Male Female Male Female Male Female Male Male Male Male Male Male Male Female Male Male
Female Female Male Female Male Male Male Female Female Female Male Female Male Male Male Male
Female Female Male Female Male Male Male Male Female Female Female Male Male Male Male Female
Female Male Female Female Male Female Male Female Male Female Female Female Female Male Female
Female Male Female Female Female Female Male Female Female Female Male Male Male Male Male Male
Male Female Male Male Female Male Male Male Male Female Male Male Female Female Female Female
Male Female Female Female Male Male Female Female Female Male Male Male Male Female Female Female
Female Female Female Male Female Female Male Male Male Female Male Male Male Female Female Female
Female Male Female Male Male Male Female Male Female Female Female Female Female Male Female Male
Male Male Male Female Male Male Female Male Male Male Female Male Male Male Female Female Male
Female Male Male Female Female Female Male Female Female Male Male Male Male Male Female Female
Female Male Male Female Male Male Male Female Male Female Male Female Female Male Male Female

Distributions

For the variable **Height** the distribution is:

178 180 175 186 170 164 158 165 182 172 158 158 185 194 170 162 162 169 193 NA 173 186 162 165 175
163 160 170 170 173 186 188 192 165 191 187 162 181 193 173 169 185 170 181 165 181 178 165 178 175
184 192 180 180 188 168 185 181 160 181 160 174 173 168 175 170 179 165 160 158 176 170 158 180 197
181 190 157 180 170 187 182 160 181 163 165 164 187 174 158 180 178 180 169 168 185 147 161 190 170
160 187 167 185 182 173 180 175 188 165 189 187 187 170 170 180 175 175 175 165 162 178 165 159 160
175 178 170 182 169 168 172 175 176 177 176 179 160 170 175 160 178 182 182 165 170 169 185 162 186
180 165 168 185 163 173 160 162 184 166 182 168 177 172 170 175 188 170 178 181 182 165 170 173 175
159 162 188 173 190 186 160 185 165 165 175 177 184 160 187 185 160 181 183 183 158 164 175 158 162
168 195 182 180 170 165 174 184 174 189 157 180 170 180 180 167 185 172 160 170 161 190 175 178 167
168 181 178 180 166 162 180 168 177 164 179 190 179 186 182 180 179 164 180 180 168 158 182 169 180
169 186 158 165 NA 178 160 195 190 183 190 155 175 187 160 180 177 192 179 168 165 170 175 185 171
180 170 170 176 165 161 173 167 185 163 178 163 165 163 160 178 167 175 181 167 170 171 172 172 173
161 174 180 187 182 188 175 187 175 172 179 180 163 180 186 190 184 180 175 183 165 170 170 172 187
165 161 158 175 185 170 168 164 182 178 188 176 170 170 178 164 160 168 176 173 160 180 181 185 170
188 185 185 165 165 168 172 176 175 177 186 188 172 178 167 165 166 173 164 185 174 178 175 172 175
161 179 NA NA 173 172 180 NA 170 185 175 175 165 180 170 188 192 157 160 160 183 168 173 190 175
185 175 180 162 165 159 186 172 173 185 175 165 161 190 166 182 168 178 196 185 165 175 160 165 184
180 171 185 160 163 170 178 172 180 172 184 168 182 168 165 180 172 180 165 175 172 180 185 160 175
173 160 170 180 178 163 183 172 183 167 178 171 182 175 160 175 175 183 184 180 183 170 160 180 182
170 167 178 187 197 184 162 181 170 173 170 176 189 180 160 167 184 170 182 180 179 164 180 180 168

Absolute Frequency Distribution

Let Y be a variable, we call **absolute frequency distribution** the list of observed modalities associated with the number of times they are observed, that is, associated with the respective absolute frequencies.

Most of the time, it is easy to obtain absolute frequency distributions for discrete qualitative and quantitative variables.

On the other hand, handling continuous variables (and discrete ones, if they have many categories) requires some preliminary operations.

Example: frequency distribution of Gender

Modality	Absolute frequency
Female	232
Male	261

Example: frequency distribution of film genre

Modality	Absolute frequency
Other	117
Comedy	9
Science fiction	17
Fantasy	42
Noir/Thriller	103
Horror	12
Psychological	68
Romantic	78
History	37

Example: frequency distribution of sleep hours

Modality	Absolute frequency
5	9
6	49
7	198
8	203
9	23
10	3
12	1

Example: frequency distribution of height

For the height, it is advisable to define classes.

Modality	Absolute frequency
147	1
155	1
157	3
158	11
159	3
160	28
161	7
162	12
163	9
164	8
165	32
166	4
167	10
168	18
169	7
170	39
171	4
172	18
173	17
174	6
175	37
176	8

Modality	Absolute frequency
177	6
178	22
179	8
180	39
181	13
182	17
183	9
184	10
185	23
186	10
187	12
188	10
189	3
190	10
191	1
192	4
193	2
194	1
195	2
196	1
197	2

Intervals of different width

Interval	Absolute frequency
[145,160]	47
(160,165]	68
(165,170]	78
(170,175]	82
(175,180]	83
(180,185]	72
(185,190]	45
(190,195]	10

Interval	Absolute frequency
[145,185]	430
(185,195]	55

The choice of classes is up to the researcher (no fixed rules), it must be done reasonably.

Intervals of different width

By choice (if you wish to provide more detailed information about a portion of the distribution) or by necessity (if the data has already been grouped into classes by someone else), the intervals can be of varying lengths.

In such cases, it is crucial to introduce the **frequency density**, which is defined as:

$$\begin{pmatrix} \text{density} \\ \text{of an interval} \end{pmatrix} = \frac{\text{absolute frequency of } Y \text{ on the interval}}{\text{length of the interval}}$$

Example: Instagram followers

Modality	Frequency
0	1
8	1
10	1
15	1
22	1
30	1
40	1
41	1
50	3
80	3
99	1
100	2
120	2
123	1
124	1
130	1
150	4
173	1
183	1
186	1
192	1
200	5
209	1
215	1
229	1
236	1
240	1
242	1
247	1
250	6
252	1

Modality	Frequency
527	1
531	1
548	1
550	1
558	1
561	1
573	1
576	1
580	1
581	1
584	1
586	2
587	1
590	1
598	1
600	17
605	1
615	1
617	1
622	2
626	1
630	1
635	1
639	1
644	1
645	1
650	4
652	1
654	1
655	1
658	2

Modality	Frequency
970	1
976	1
987	1
1000	21
1003	1
1017	1
1018	1
1019	1
1031	1
1032	1
1040	1
1042	1
1043	1
1047	1
1050	1
1067	1
1068	1
1070	1
1075	1
1082	1
1100	4
1110	1
1113	1
1136	1
1164	1
1173	1
1174	1
1200	8
1205	1
1235	1
1259	1

Example: Instagram followers

Interval	Frequency
[0,500]	169
(500,1000]	150
(1000,1500]	59
(1500,2000]	17
(2000,2500]	10
(2500,3000]	6
(3000,100000]	5

Frequency Density

The density tells us the expected number of statistical units for each unit of measurement of the variable. In the first class, for example, we expect to see 17 people in a range of 100 units, in the second last class we expect to see 150 units for every 500.

Interval	Frequency	Interval width	Density
[0,100]	17	100	17/100=0.17
(100,200]	18	100	18/100=0.18
(200,300]	43	100	43/100=0.43
(300,400]	48	100	48/100=0.48
(400,500]	43	100	43/100=0.43
(500,1000]	150	500	150/500=0.3
(1000,10000]	96	9000	96/9000=0.01067

Conditional frequency distributions

Let us call:

- Y the variable we are studying (Ex: number of hours of sleep)
- X the variable through which we extract the statistical units to be considered in the analysis (Ex: gender)

We define the variable Y conditional on $X = x$ and we write $Y|X = x$ to express the restriction of Y to the modelity $X = x$.

The distribution of the variable $Y|X = x$ is usually called the *distribution of Y conditional on $X = x$* or, equivalently, the *distribution of Y given $X = x$*.

There is one of these conditional distribution for each modality of X.

Example: height, males and females

In the case of continuous quantitative variables, which can be divided into intervals, things work in the same way.

Females

Interval	Frequency
[145,160]	47
(160,165]	67
(165,170]	65
(170,175]	40
(175,180]	9
(180,185]	0
(185,190]	0
(190,195]	0

Males

Interval	Frequency
[145,160]	0
(160,165]	1
(165,170]	13
(170,175]	42
(175,180]	74
(180,185]	72
(185,190]	45
(190,195]	10

Relative Frequencies

By dividing an absolute frequency by the total number of statistical units (N in our case) we obtain the so-called relative frequencies, that is:

$$\left(\begin{array}{c}\text{relative}\\\text{frequencies}\end{array}\right)=\frac{\left(\begin{array}{c}\text{absolute}\\\text{frequencies}\end{array}\right)}{\left(\begin{array}{c}\text{total number of}\\\text{observations}\end{array}\right)}$$

They allow comparisons of distributions based on different numbers of statistical units.

Relative frequencies: hours of sleep

Modality	Freq.		
	F	M	Tot
5	4	5	9
6	24	25	49
7	95	103	198
8	98	105	203
9	6	17	23
10	1	2	3
12	0	1	1

Modality	Rel. F.		
	F	M	Tot
5	0.02	0.02	0.02
6	0.11	0.10	0.10
7	0.42	0.40	0.41
8	0.43	0.41	0.42
9	0.03	0.07	0.05
10	0.00	0.01	0.01
12	0.00	0.00	0.00

Example: Instagram Followers

Interval	Freq	Rel. Freq.
[0,100]	17	17/416=0.041
(100,200]	18	18/416=0.043
(200,300]	43	43/416=0.103
(300,400]	48	48/416=0.115
(400,500]	43	43/416=0.103
(500,750]	85	85/416=0.204
(750,1000]	65	65/416=0.156
(1000,2000]	76	76/416=0.183
(2000,5000]	19	19/416=0.046
(5000,100000]	2	2/416=0.005

Frequency distribution: notation

- y_i i-th modelity / interval $(c_{i-1}, c_i]$ of the variable Y for $i = 1, 2, \ldots, k$ (k modalities / intervals);
- n_i absolute frequency: number of statistical units that present the modality/interval y_i;
- N total number of observations

$$N = \sum_{i=1}^{k} n_i = n_1 + n_2 + \cdots + n_k$$

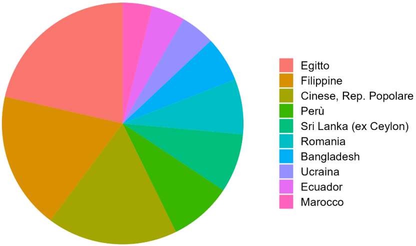
- f_i relative frequency ($f_i = n_i/N$).

modality/interval	Freq.	Rel.F.
y_1	n_1	$f_1 = n_1/N$
y_2	n_2	$f_2 = n_2/N$
$\vdots$	$\vdots$	$\vdots$
y_k	n_k	$f_k = n_k/N$
Total	N	1

Plot

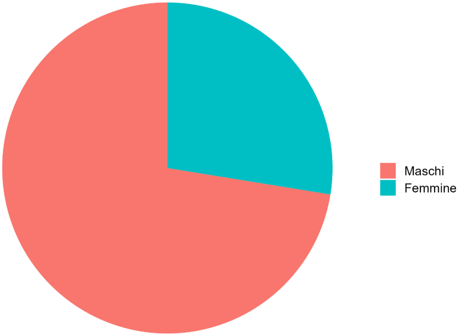
We can visualize the frequency distributions of a variable representing each modality with the relative frequency.
Example: Relative frequency distribution of the top 10 nationalities in Milan in 2023

Nationality	Inhabitants
Egitto	45457
Filippine	38942
Cinese, Rep. Popolare	37041
Perù	17799
Sri Lanka (ex Ceylon)	16724
Romania	15673
Bangladesh	12802
Ucraina	9704
Ecuador	9513
Marocco	8351



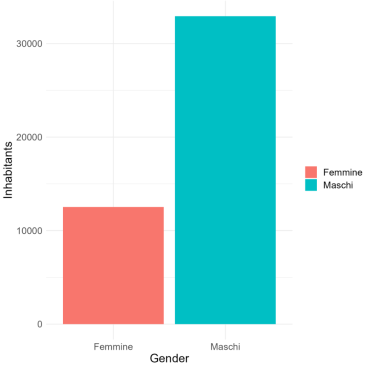
Another plot...

Example: Relative frequency distribution of Egyptians by gender in Milan in 2023



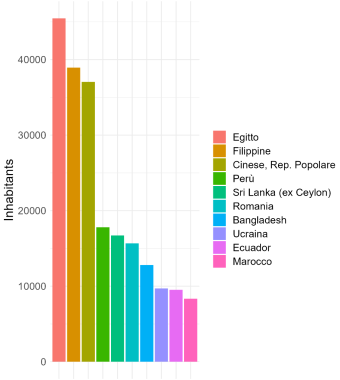
Another plot...

Example: Absolute frequency distribution of Egyptians by gender in Milan in 2023



When the variable is nominal (as in this case), the position of the modalities is arbitrary

Barplot

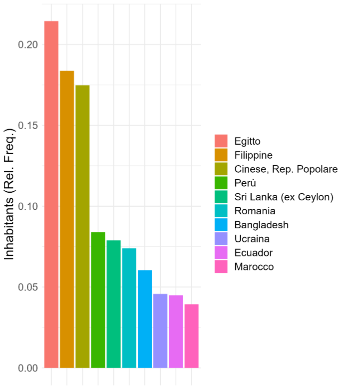


The plot shows:

$$x\text{-axis} = \begin{pmatrix} \text{modalities} \\ \text{reported in the} \\ \text{frequency} \\ \text{distribution} \end{pmatrix}$$

$$(\text{bars height}) = (\text{absolute frequencies})$$

Barplot



The plot shows:

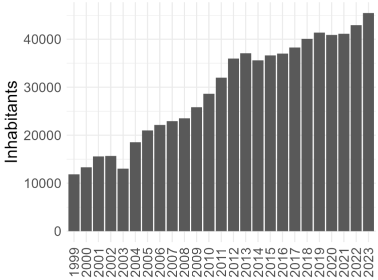
$$x\text{-axis} = \left(\begin{array}{c} \text{modalities} \\ \text{reported in the} \\ \text{frequency} \\ \text{distribution} \end{array} \right)$$

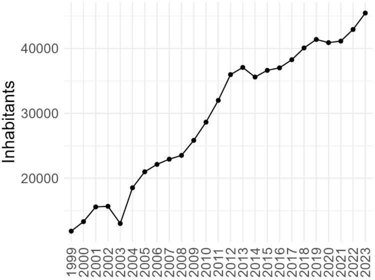
$$(\text{bars height}) = (\text{absolute frequencies})$$

Using the relative frequencies
we obtain the same plot

$$(\text{bars height}) = (\text{relative frequencies})$$

Barplot: Frequency by year

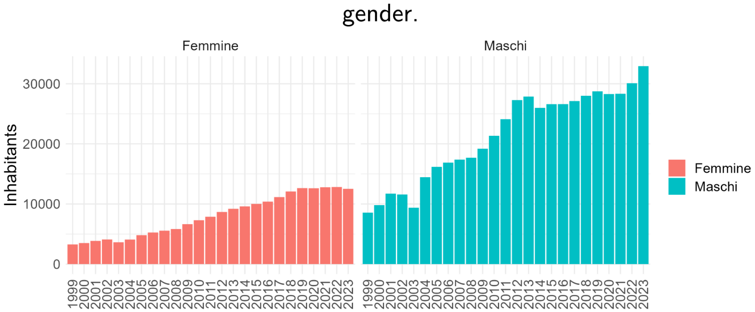




When the variable is ordinal or quantitative discrete, the position of the modalities is meaningful and follows a specific order.

Barplot and comparisons: year and gender

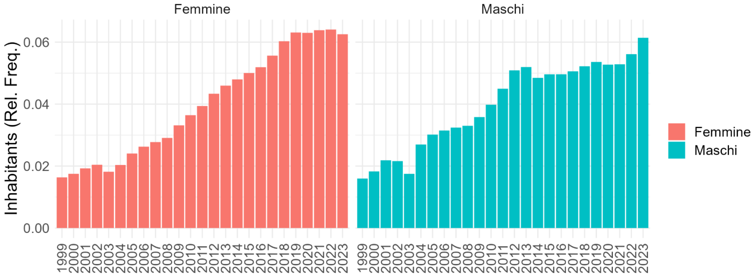
We can compare the population distribution in Milan in the last years by gender.



What type of frequencies are these? Are they suitable for this task?

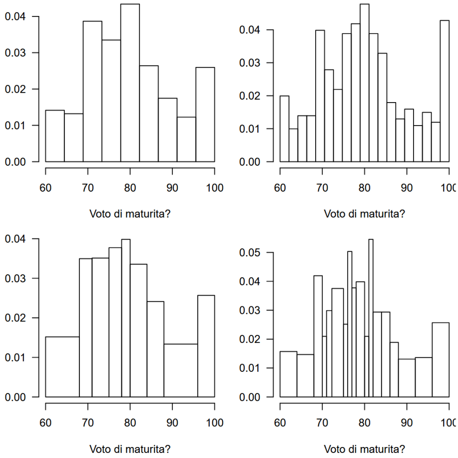
Barplot and comparisons: year and gender

No! This is exactly the reason why we need to compute relative frequencies.



High school grade

Which one of these histograms is useful? Which one gives too many details? Which one hides too much of the data?



Histograms and intervals

- The fewer the number of intervals, the fewer the details,
- However, too many intervals means too much detail. This might lead to sample bias: we depict characteristics that are related to our specific sample.
- A good strategy is to make several graphs from the same data, changing the interval widths, and then choose the best one;
- The interval number is related to the data size!