Regression and Time Series Models

Manuela Cattelan

manuela.cattelan@unipd.it

Department of Statistical Sciences, University of Padova

Course details

Instructors and lectures

- Proff. Massimiliano Caporin and Manuela Cattelan
- I teach the introductory part (16 h) concerning the basics of statistics
- To arrange a meeting send me an email
  email: `manuela.cattelan@unipd.it`
- The material of the lectures and additional exercises will be available in the Moodle page of the course.

Textbooks

- A free book available online for reviewing all the basics of statistics is A. Holmes, B. Illowsky and S. Dean. (2023) Introductory Business Statistics. Here the link
- A free book available online for implementing basic statistics in Python is Haslwanter T. (2022). An Introduction to Statistics with Python, with Applications in the Life Sciences. Springer, second edition. Here the link
- **Important: I will not deal with Python, during the first part we will be concentrating only on revising statistics.**

Contents

We are going to review topics you should already be familiar with

- Descriptive statistics
  - concepts of population, sample, variable and type of variables
  - main statistics: location, scale, shape
  - graphical representations
  - relation between variables
- Inferential statistics
  - parameter estimation
  - confidence intervals
  - hypothesis testing

Inferential statistics requires some knowledge of probability theory, which I assume you possess.

If I am taking too much for granted, interrupt me!

Do you remember descriptive statistics?

1 Vai a wooclap.com

2 Immettere il codice dell'evento nel banner superiore

Codice evento

MPCQDE

Introduction

What is statistics?

The science of statistics deals with the collection, analysis, interpretation, and presentation of data. We see and use data in our everyday lives.

The world is uncertain, and you should take this into account.

Statistics is used to tackle a variety of problems in finance, including building financial models, estimation and inference for financial models, volatility estimation, risk management, testing financial economics theory, capital asset pricing, derivative pricing, portfolio allocation, risk-adjusted returns, simulating financial systems, hedging strategies, etc..

But before studying those topics, you need the basics!

Descriptive vs inferential statistics

- Descriptive statistics is used for organizing and summarizing data (numbers or graphs)
- Inferential statistics is a way of making inferences about populations based on samples

Some terminology

Population includes all of the elements (persons, things, or objects) under study.

Sample consists of one or more observations from the population. It is a portion of the population.

A **statistical unit** is a unit of observation (an entity) for which data are collected.
A **variable** is a characteristic of interest that is measured, recorded, and analysed.

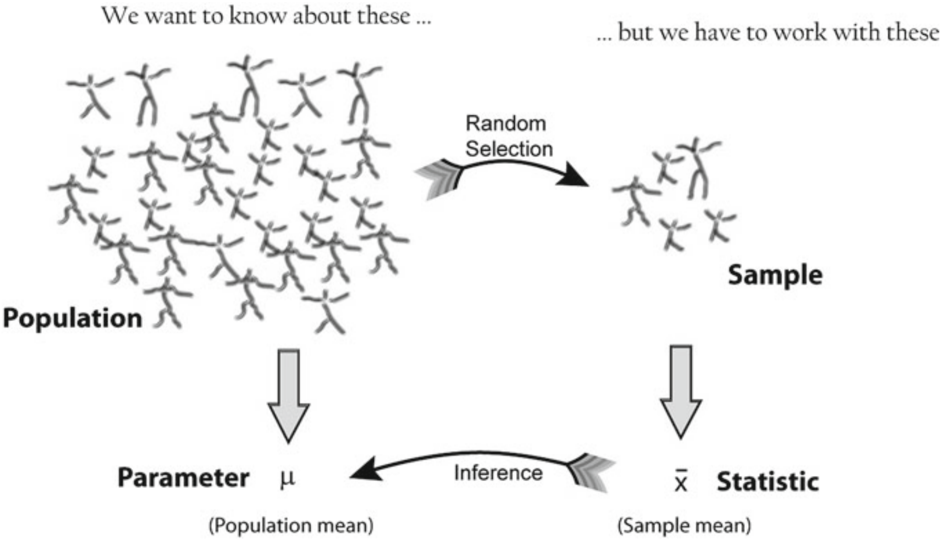
We are typically interested in parameters of the distribution of a variable in a population, but most often we have only sample statistics.

Parameter Characteristic of a distribution describing a population, such as the mean or standard deviation. Often notated using Greek letters (μ, σ^2).

Statistic A function of a sample that does not depend on unknown quantities. Its realization is a numerical value that represents a property of a random sample (e.g. mean, range, standard deviation)

Statistical inference enables you to make an educated guess about a population parameter based on a statistic computed from a representative sample from that population.

Population vs sample



Example

We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first year students at the college. Three of those students spent \$150, \$200, and \$225, respectively.

- The *population* is all first year students attending ABC College this term.
- The *sample* could be the first year students met at the canteen of the ABC College.
- The *variable* is the amount of money spent (excluding books) by one first year student.
- The *parameter* is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term: the population mean.
- The *statistic* is the average (mean) amount of money spent (excluding books) by first year college students in the sample.
- The *data* are the dollar amounts spent by the first year students. Examples of the data are \$150, \$200, and \$225

Data types

Variables can be

- Qualitative (Categorical)
  - Nominal (married/single/divorced; eye colour; gender; nationality)
  - Ordinal (satisfaction rating: dislike/neutral/like; educational level)
- Quantitative (Numerical)
  - Discrete (number of children; people that enter a shop; number of phone calls)
  - Continuous (height; weight; length of phone calls)

The way a set of data is measured is called its level of measurement. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- Nominal scale level (color names; food types)
- Ordinal scale level (level of satisfaction, as excellent, good, satisfactory, unsatisfactory)
- Interval scale level (temperature scales - differences make sense but 0 degrees does not)
- Ratio scale level (weight in kilos; height in meters)

Example 1

Returns of 20 stocks

Symbol	Name	% change	Sector
ROST	Ross Stores Inc	-0.41%	Consumer_discretionary
WBD	Discovery Inc Series A	-0.26%	Communication_services
OTIS	Otis Worldwide Corp	-1.39%	Industrials
FITB	Fifth Third Bancorp	-2.20%	Financial
MS	Morgan Stanley	-1.72%	Financial
ALGN	Align Technology	-0.47%	Health_care
AMT	American Tower Corp	-2.53%	Real_estate
CNC	Centene Corp	-0.60%	Health_care
CAH	Cardinal Health	+0.20%	Health_care
EVRG	Evergy Inc	-0.94%	Utilities
HPQ	HP Inc	-1.83%	Information_technology
EW	Edwards Lifesciences Corp	-1.06%	Health_care
XOM	Exxon Mobil Corp	+1.12%	Energies
COST	Costco Wholesale	-0.44%	Consumer_staples
BMY	Bristol-Myers Squibb Company	-1.59%	Health_care
ZBH	Zimmer Biomet Holdings	-1.30%	Health_care
WTW	Willis Towers Watson Public Ltd	-0.36%	Financial
DXC	Dxc Technology Company	-1.71%	Information_technology
GWW	W.W. Grainger	-0.46%	Industrials
STE	Steris Corp	-0.52%	Health_care

Which are the units? And the variables? Which type of variable is the Sector? And the percentage change in the price of the stock?

Example 2

15 small firms: number of employees, annual revenue and sector of activity

Firm	Employees	Revenue	Sector
A	13	18.48	Materials
B	16	20.58	Financial
C	11	20.84	Health_care
D	16	17.41	Information_technology
E	19	20.14	Communication_services
F	16	18.37	Energies
G	10	23.02	Utilities
H	11	19.46	Health_care
I	12	23.12	Consumer_discretionary
J	12	19.53	Industrials
K	15	22.57	Health_care
L	15	19.98	Financial
M	13	19.20	Health_care
N	20	20.04	Utilities
O	17	23.49	Financial

Frequency distributions

Consider the variable “Sector of activity”, which type of variable is it? How can we summarise the sample?

- X represents the variable “Sector of activity”
- $x_1, \ldots, x_n$ represent the values assumed by the variable for each unit observed
- n is the total number of units ($n = 15$ in this example)
- $x_5 = \textit{Communication\_services}$

Firm	Employees	Revenue	Sector
A	13	18.48	Materials
B	16	20.58	Financial
C	11	20.84	Health_care
D	16	17.41	Information_technology
E	19	20.14	Communication_services
F	16	18.37	Energies
G	10	23.02	Utilities
H	11	19.46	Health_care
I	12	23.12	Consumer_discretionary
J	12	19.53	Industrials
K	15	22.57	Health_care
L	15	19.98	Financial
M	13	19.20	Health_care
N	20	20.04	Utilities
O	17	23.49	Financial

Consider the variable "Sector of activity", which type of variable is it? How can we summarise the data?
Frequency distributions are tables that list the number of occurrences of each value taken by a variable.
Absolute frequency: number of units that have the same value of the variable.

- K is the total number of different values of X observed (here $K = 9$)
- $x_1, \dots, x_K$ represent the K different values assumed by the variable
- n_i, $i = 1, \dots, K$, denotes the absolute frequency for the i-th value taken by the variable (for example $n_3 = n_{Energies} = 1$)
- $\sum_{i=1}^{K} n_i = n$

Sector	n_i
Communication_services	1
Consumer_discretionary	1
Energies	1
Financial	3
Health_care	4
Industrials	1
Information_technology	1
Materials	1
Utilities	2

Consider the variable "Number of employees", which type of variable is it?
How can we summarise the data?
We can compute absolute frequencies, in this case it makes sense to order the different values of the variable

Firm	Employees	Revenue	Sector
A	13	18.48	Materials
B	16	20.58	Financial
C	11	20.84	Health_care
D	16	17.41	Information_technology
E	19	20.14	Communication_services
F	16	18.37	Energies
G	10	23.02	Utilities
H	11	19.46	Health_care
I	12	23.12	Consumer_discretionary
J	12	19.53	Industrials
K	15	22.57	Health_care
L	15	19.98	Financial
M	13	19.20	Health_care
N	20	20.04	Utilities
O	17	23.49	Financial

Number of employees	n_i
10	1
11	2
12	2
13	2
15	2
16	3
17	1
19	1
20	1
	n=15

Another example

Absolute frequencies

From another district, we collect the following data

Firm	Employees	Revenue	Sector
F-1	20	23.38	Real_estate
F-2	18	20.00	Industrials
F-3	19	18.52	Consumer_discretionary
F-4	16	21.22	Health_care
F-5	17	18.02	Real_estate
F-6	15	19.93	Consumer_discretionary
F-7	16	21.69	Information_technology
F-8	12	23.05	Consumer_discretionary
F-9	17	19.87	Consumer_discretionary
F-10	19	20.42	Consumer_staples
F-11	16	19.84	Information_technology
F-12	19	20.03	Information_technology
F-13	11	17.59	Consumer_discretionary
F-14	17	19.48	Real_estate
F-15	17	21.42	Consumer_discretionary
F-16	16	19.74	Information_technology
F-17	15	18.45	Materials
F-18	16	20.17	Utilities
F-19	20	20.13	Communication_services
F-20	15	20.23	Consumer_staples
F-21	11	23.53	Real_estate
F-22	14	18.37	Consumer_discretionary
F-23	18	19.82	Health_care
F-24	11	20.63	Industrials
F-25	19	15.88	Consumer_staples

Consider the variable "Number of employees" for both districts, can we compare their distribution on the basis of the absolute frequencies?

Number of employees	District 1	District 2
10	1	0
11	2	3
12	2	1
13	2	0
14	0	1
15	2	3
16	3	5
17	1	4
18	0	2
19	1	4
20	1	2
	n=15	n=25

Consider the variable "Number of employees" for both districts, can we compare their distribution on the basis of the absolute frequencies?

Number of employees	District 1	District 2
10	1	0
11	2	3
12	2	1
13	2	0
14	0	1
15	2	3
16	3	5
17	1	4
18	0	2
19	1	4
20	1	2
	n=15	n=25

No, they do not have the same size!

We can use relative frequencies that take size into account

$$f_i = \frac{n_i}{n} \text{ and note that } \sum_{i=1}^{K} f_i = \sum_{i=1}^{K} \frac{n_i}{n} = \frac{n}{n} = 1 \text{ and } 0 \leqslant f_i \leqslant 1,\ i = 1, \ldots, K$$

Number of employees	District 1	District 2
10	$1/15 = 0.0\bar{6}$	$0/25 = 0$
11	$2/15 = 0.1\bar{3}$	$3/25 = 0.12$
12	$2/15 = 0.1\bar{3}$	$1/25 = 0.04$
13	$2/15 = 0.1\bar{3}$	$0/25 = 0$
14	$0/15 = 0$	$1/25 = 0.04$
15	$2/15 = 0.1\bar{3}$	$3/25 = 0.12$
16	$3/15 = 0.20$	$5/25 = 0.20$
17	$1/15 = 0.0\bar{6}$	$4/25 = 0.16$
18	$0/15 = 0$	$2/25 = 0.08$
19	$1/15 = 0.0\bar{6}$	$4/25 = 0.16$
20	$1/15 = 0.0\bar{6}$	$2/25 = 0.08$
	1	1

Note that relative frequencies can be computed also for qualitative data.

Cumulative frequencies

Sometimes, we may be interested in the number of units that have a value of the variable not bigger than a specified value x_0

- The cumulative frequencies are the number of units with a value less than or equal to x_0
- Absolute cumulative frequencies. Let $x_1, x_2, \ldots, x_K$ denote the (ordered) different values observed for the variable of interest, and let x_j be the largest value observed such that $x_j \leqslant x_0$, then the absolute cumulative frequency is

$$N_j = \sum_{i=1}^{j} n_i$$

- Relative cumulative frequencies

$$F_j = N_j / n$$

Note that cumulative frequencies can be computed for categorical ordinal data but are meaningless for nominal data.

Example with number of employees in district 1

Number of employees	n_i	f_i	N_i	F_i
10	1	$1/15 = 0.0\bar{6}$	1	$1/15 = 0.0\bar{6}$
11	2	$2/15 = 0.1\bar{3}$	1+2=3	$0.0\bar{6} + 2/15 = 0.20$
12	2	$2/15 = 0.1\bar{3}$	3+2 =5	$0.20 + 2/15 = 0.3\bar{3}$
13	2	$2/15 = 0.1\bar{3}$	5+2 =7	$0.2\bar{6} + 2/15 = 0.4\bar{6}$
15	2	$2/15 = 0.1\bar{3}$	7+2 =9	$0.4\bar{6} + 2/15 = 0.60$
16	3	$3/15 = 0.20$	9+3=12	$0.60 + 3/15 = 0.80$
17	1	$1/15 = 0.0\bar{6}$	12+1 =13	$0.80 + 1/15 = 0.8\bar{6}$
19	1	$1/15 = 0.0\bar{6}$	13+1=14	$0.8\bar{6} + 1/15 = 0.9\bar{3}$
20	1	$1/15 = 0.0\bar{6}$	14+1=15	$0.9\bar{3} + 1/15 = 1.00$
	n=15			

Note that cumulative frequencies can be computed for categorical ordinal data but are meaningless for nominal data.

- How can we summarise continuous variables, that assume most likely a different value for each statistical unit?
- We divide the data into classes and compute absolute, relative and cumulative frequencies for each class.
- To compute the number of observations in each class, we may sort all the observed values.
- Let $x_{(i)}$ denote the i-th smallest value, this is called an *order statistics*.
- Consider the Revenues in district 1, in this case $x_1 = 18.48$ is the value of the variable observed on the first statistical unit, while $x_{(1)} = 17.41$ is the smallest value observed in all the sample.

Consider the “Revenues” of firms in district 1.

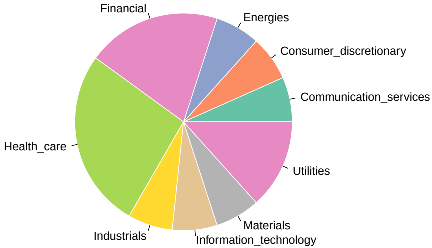
The sorted data are
17.41, 18.37, 18.48, 19.20, 19.46, 19.53, 19.98, 20.04, 20.14, 20.58, 20.84, 22.57, 23.02, 23.12, 23.49

Class or bin	n_i	f_i	N_i	F_i
(17,19]	3	$3/15 = 0.20$	3	0.20
(19,20]	4	$4/15 = 0.2\bar{6}$	$3 + 4 = 7$	$0.20 + 0.2\bar{6} = 0.4\bar{6}$
(20,21]	4	$4/15 = 0.2\bar{6}$	$7 + 4 = 11$	$0.4\bar{6} + 0.2\bar{6} = 0.7\bar{3}$
(21,23]	1	$1/15 = 0.0\bar{6}$	$11 + 1 = 12$	$0.7\bar{3} + 0.0\bar{6} = 0.80$
(23,24]	3	$3/15 = 0.20$	$12 + 3 = 15$	$0.80 + 0.20 = 1.00$

Graphical representations

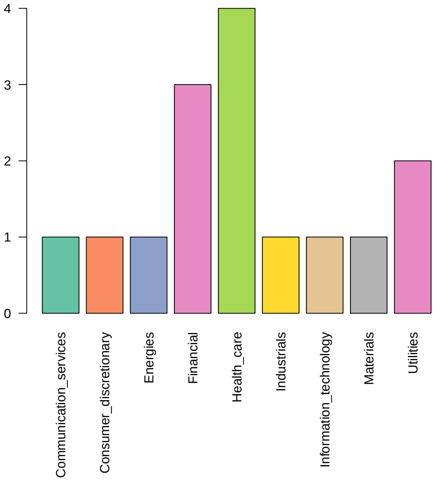
Categorical nominal data

Pie chart

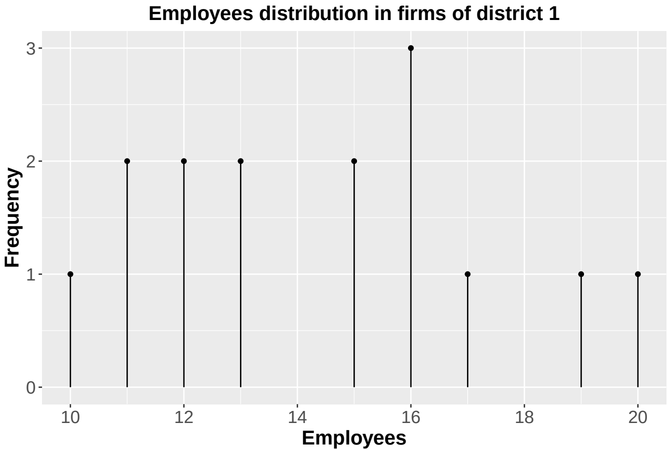


- $\text{angle}_i = n_i/n \times 360 = f_i \times 360$
- Difficult to read (angles??)
- 3D, even worse
- exploded, even worse

Bar plot

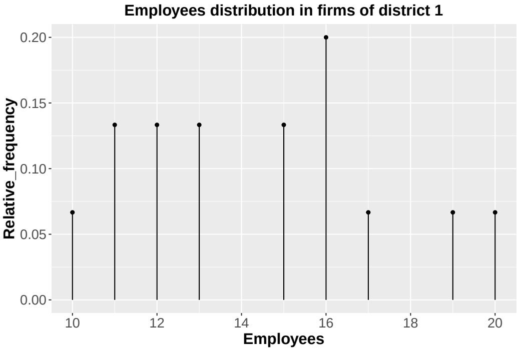


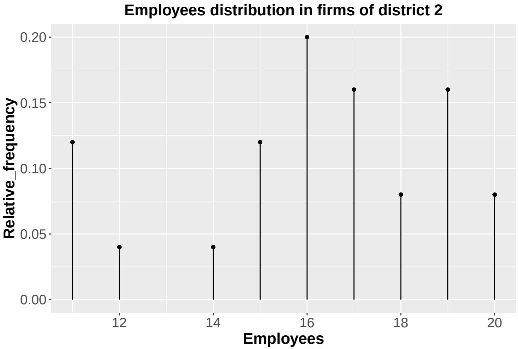
Lollipop plot



Sometimes used also for categorical and even nominal data instead of bar plots.
Pay attention to the x-axis.

In case you need to compare the distribution of a variable in two populations, use relative frequencies!

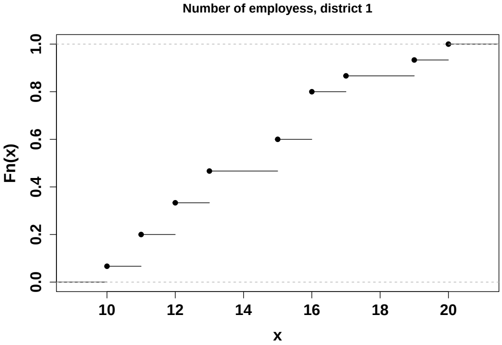




Empirical cumulative distribution functions

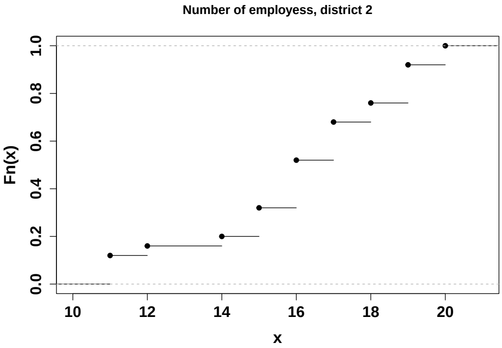
A graphical representation of (relative) cumulative frequencies (the percentage of observations below or equal to a specific value) is the empirical distribution function.

N. employees	n_i	f_i	F_i
10	1	$0.0\overline{6}$	$0.0\overline{6}$
11	2	$0.1\overline{3}$	0.20
12	2	$0.1\overline{3}$	$0.3\overline{3}$
13	2	$0.1\overline{3}$	$0.4\overline{6}$
15	2	$0.1\overline{3}$	0.60
16	3	0.20	0.80
17	1	$0.0\overline{6}$	$0.8\overline{6}$
19	1	$0.0\overline{6}$	$0.9\overline{3}$
20	1	$0.0\overline{6}$	1.00
	n=15		



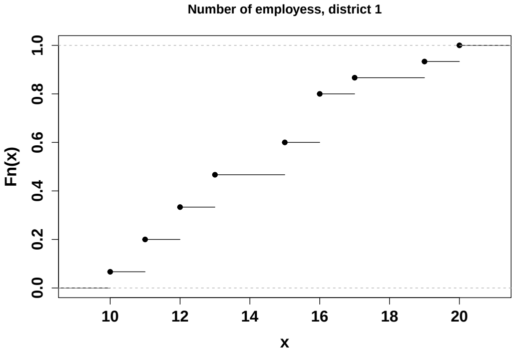
A graphical representation of (relative) cumulative frequencies (the percentage of observations below or equal to a specific value) is the empirical distribution function.

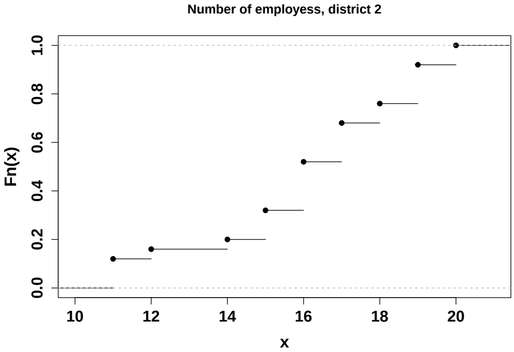
N. employees	n_i	f_i	F_i
11	3	0.12	0.12
12	1	0.04	0.16
14	1	0.04	0.20
15	3	0.12	0.32
16	5	0.20	0.52
17	4	0.16	0.68
18	2	0.08	0.76
19	4	0.16	0.92
20	2	0.08	1.00
	n=25		



Empirical cumulative distribution functions

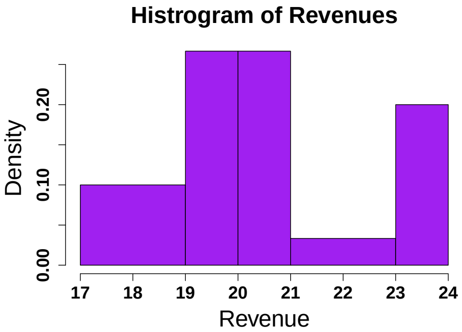
A graphical representation of (relative) cumulative frequencies (the percentage of observations below or equal to a specific value) is the empirical distribution function.





Revenues	n_i	f_i	w_i	d_i
(17,19]	3	$3/15 = 0.20$	19-17=2	$0.20/2 = 0.10$
(19,20]	4	$4/15 = 0.2\bar{6}$	$20 - 19 = 1$	$0.2\bar{6}/1 = 0.2\bar{6}$
(20,21]	4	$4/15 = 0.2\bar{6}$	$21 - 20 = 1$	$0.2\bar{6}/1 = 0.2\bar{6}$
(21,23]	1	$1/15 = 0.0\bar{6}$	$23 - 21 = 2$	$0.0\bar{6}/2 = 0.0\bar{3}$
(23,24]	3	$3/15 = 0.20$	$24 - 23 = 1$	$0.20/1 = 0.20$

w_i is the class width (upper bound-lower bound) and $d_i = f_i/w_i$ is the class density




S&P 500
Index level
4500
4000
3500
3000
2500
Jul 2018
Jan 2019
Jul 2019
Jan 2020
Jul 2020
Jan 2021
Jul 2021
Jan 2022
Jul 2022
Jan 2023
Time

Location measures

Mean

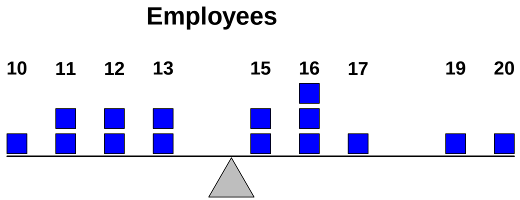
Should you summarise the phenomenon under study with only one number, which one would you use?

The most well-known location measure is the (arithmetic) mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

or, if you have frequency distributions

$$\bar{x} = \frac{\sum_{i=1}^{K} x_i n_i}{\sum_{i=1}^{K} n_i} \text{ or } \bar{x} = \sum_{i=1}^{K} x_i f_i.$$



$$\begin{aligned}\bar{x} &= \frac{1}{15}(13+16+11+16+19+16+10+11+12+12+15+15+13+20+17)\\ &= 10(0.0\bar{6})+11(0.1\bar{3})+12(0.1\bar{3})+13(0.1\bar{3})+15(0.1\bar{3})+16(0.20)+17(0.0\bar{6})+19(0.0\bar{6})+20(0.0\bar{6})\\ &= 14.4\end{aligned}$$

Median

A value of the variable that separates the higher half from the lower half of the data sample

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$	$x_{(15)}$
10	11	11	12	12	13	13	15	15	16	16	16	17	19	20

Ordered value in position

- $\frac{n+1}{2}$ if n is odd, that is $x_{(n+1)/2}$
- any value between those in position $\frac{n}{2}$ and $\frac{n}{2}+1$ if n is even, typically $\left(x_{(n/2)} + x_{(n/2+1)}\right)/2$

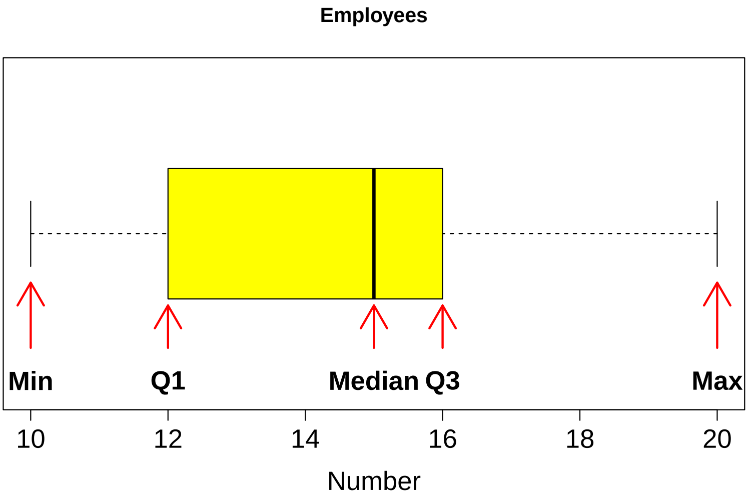
More generally, a p quantile $(0 < p < 1)$ is a value, Q, with the property that at least $100p\%$ of the data are less than Q and at least $100(1-p)\%$ of the data are greater than or equal to Q. When $p = 0.01, \dots, 0.99$, they are also called *percentiles*.

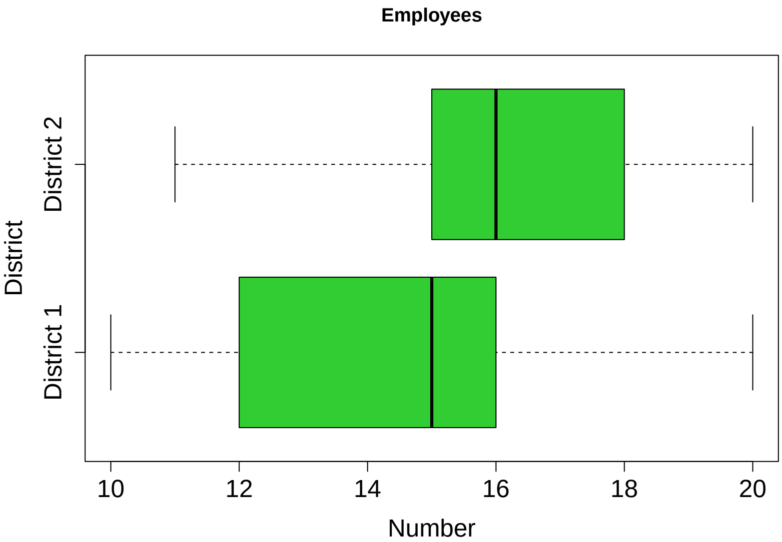
Particularly interesting are the **quartiles**: Q_1, Q_2, Q_3, that divide the distribution into four parts, each containing 25% of the observations.

For example, $0.25 \times 15 = 3.75 \rightarrow x_{(4)}$,
and $0.75 \times 15 = 11.25 \rightarrow x_{(12)}$

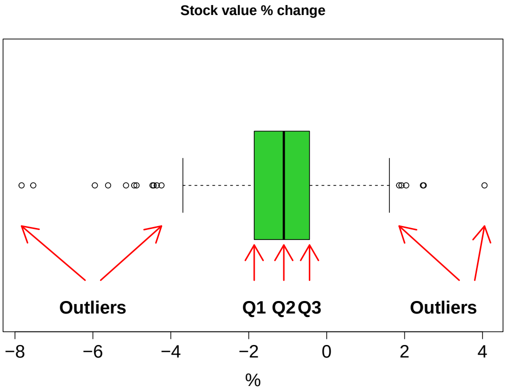
$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$	$x_{(15)}$
10	11	11	12	12	13	13	15	15	16	16	16	17	19	20

Box and whiskers plot




Employees
District 2
District 1
District
10
12
14
16
18
20
Number

Boxplot



Percentage change of the stock market value of the components of S&P500 index (May 26, 2023)

- $Min = -7.83\%$
- $Q1 = -1.86\%$
- $Q2 = -1.10\%$
- $Q3 = -0.44\%$
- $Max = 4.05\%$
- $IQ = Q_3 - Q_1 = -0.44 - (-1.86) = 1.42$

Limits of the whiskers:

$$Q_1 - 1.5 \times IQ = -1.86 - 1.5 \times 1.42 = -3.99$$

and

$$Q_3 + 1.5 \times IQ = -0.44 + 1.5 \times 1.42 = 1.69.$$

Mode

When the variable is nominal, the only location measure we can compute is the **mode**, which is the value of the variable that is assumed most often in the sample.

For example, the mode of the distribution of the variable "Sector" of the firms in district 1 is *Health_care*

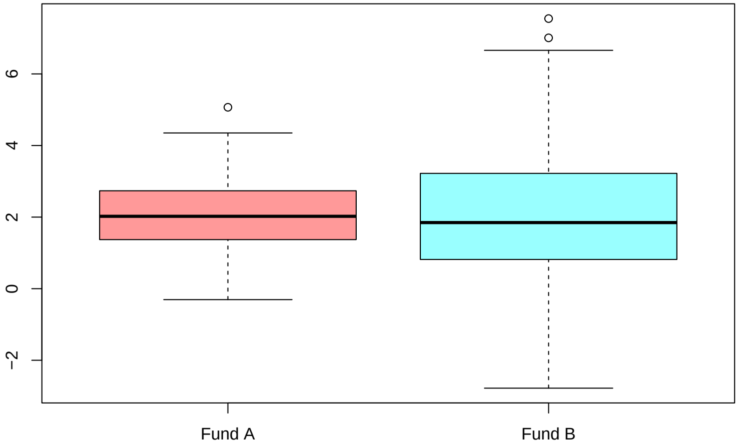
Sector	n_i
Communication_services	1
Consumer_discretionary	1
Energies	1
Financial	3
Health_care	4
Industrials	1
Information_technology	1
Materials	1
Utilities	2

- Can we compute the mode of the variable “Number of Employees”?
- Can we compute the median of the variable “Sector”?
- Consider the variable “Education level” of a person, can we compute its mean? And its median? And its mode?

Scale measures

Variability

Consider the boxplots of the performances of two funds, fund A and fund B, in the last 6 months. Their means are equal, but is there one you would prefer?



$$\text{range} = \max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n) = x_{(n)} - x_{(1)}$$

$$\text{interquartile range} = (\text{third quartile}) - (\text{first quartile}) = Q_3 - Q_1$$

The range is very sensitive to outliers.

$$\begin{aligned}
\text{range(fund A)} &= 5.07 - (-0.310) = 5.38 \\
\text{range(fund B)} &= 7.54 - (-2.78) = 10.32 \\
\text{IQ (fund A)} &= 2.74 - 1.375 = 1.365 \\
\text{IQ (fund B)} &= 3.195 - 0.8175 = 2.3775
\end{aligned}$$

Variance

The variance is one of the most employed measures of variability

$$\text{variance}(x_1, \dots, x_n) = \text{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. There is a formula that simplifies computations

$$\begin{aligned}
\text{var}(x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2 - \frac{1}{n}\sum_{i=1}^{n} 2\bar{x}x_i \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 + \frac{n\bar{x}^2}{n} - \frac{2\bar{x}}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 + \bar{x}^2 - 2\bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2
\end{aligned}$$

And the standard deviation is its square root

$$sd(x) = \sqrt{\text{var}(x)}.$$

We can compute var(fund A) $= 1.069$ and var(fund b) $= 3.558$.

Let's compute the variance for a small example

- Data: $1, 3, 2, 5$
- Mean: $\bar{x} = \frac{1+3+2+5}{4} = 2.75$
- Mean of the squares: $\frac{1^2+3^2+2^2+5^2}{4} = 9.75$
- Variance: $\text{var}(x) = 9.75 - 2.75^2 = 2.19$

Let's compute the variance for another small example

- Data: $1, 3, 2, 5$, with absolute frequencies $3, 1, 4, 2$, respectively
- Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i n_i = \frac{1(3)+3(1)+2(4)+5(2)}{10} = 2.4$
- Mean of the squares: $\frac{1^2(3)+3^2(1)+2^2(4)+5^2(2)}{10} = 7.8$
- Variance: $\text{var}(x) = 7.8 - 2.4^2 = 2.04$

Try to obtain the same results using relative frequencies.

The practical importance of variability may depend on the level of the phenomenon.

If one compares the variability of two samples, the different location should be taken into account. For this reason, one should consider the **coefficient of variation**

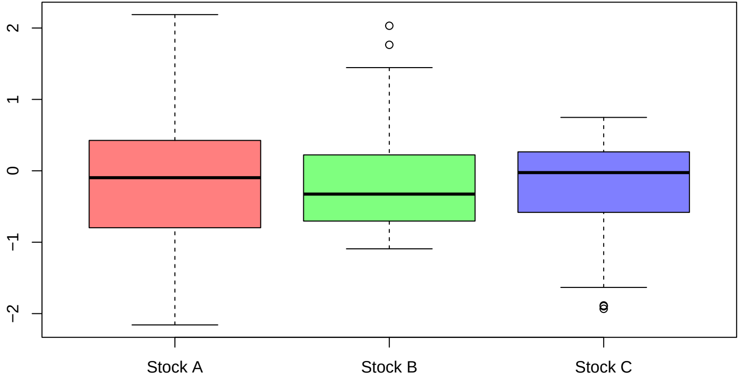
$$\mathrm{CV}(x) = \frac{\mathrm{sd}(x)}{|\bar{x}|}$$

The coefficients of variation for the two funds are

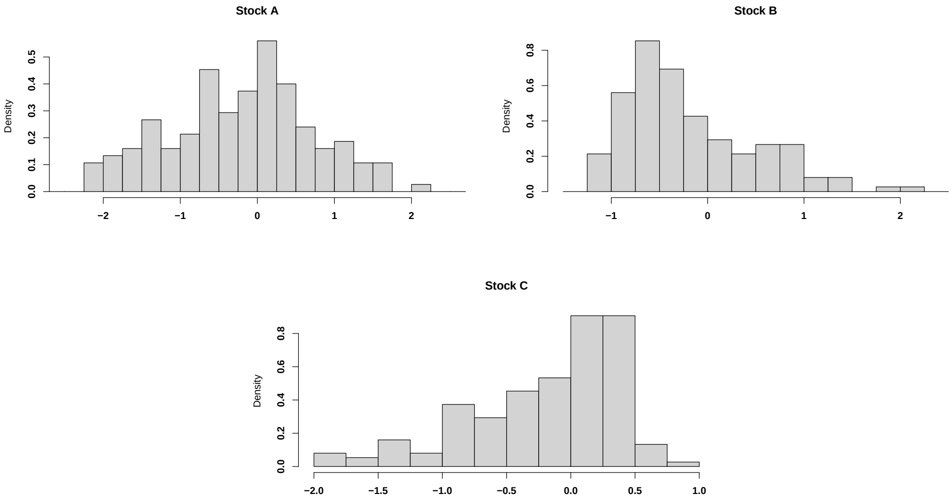
$$\mathrm{CV}(\text{fund A}) = \frac{1.034}{2.022} = 0.511$$
$$\mathrm{CV}(\text{fund B}) = \frac{1.886}{2.0179} = 0.935$$

Shape measures

Skewness



Histograms



Stock A
Density
0.0
0.1
0.2
0.3
0.4
0.5
−2
−1
0
1
2
Stock B
Density
0.0
0.2
0.4
0.6
0.8
−1
0
1
2
Stock C
Density
0.0
0.2
0.4
0.6
0.8
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0

The most common coefficient of skewness was introduced by Karl Pearson

$$\frac{1}{n\,\mathrm{sd}(x)^3}\sum_{i=1}^{n}(x_i-\bar{x})^3,$$

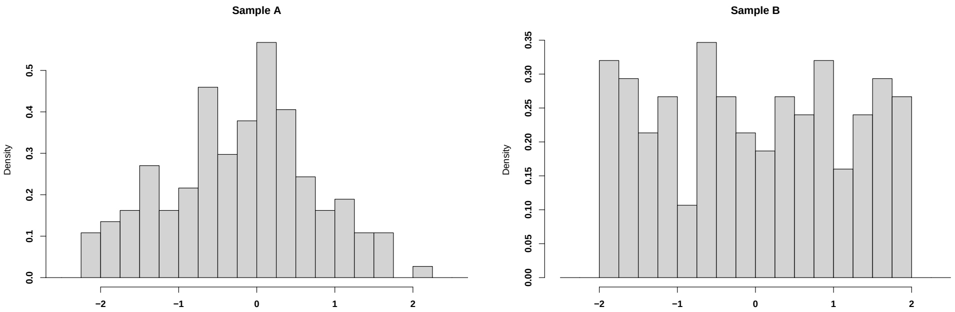
where n is the sample size, $\mathrm{sd}(x)$ the standard deviation of the sample and $\bar{x}$ its mean.

Symmetric data $\rightarrow$ positive and negative terms cancel out, so we expect its value to be circa 0

Skew data $\rightarrow$ either positive or negative terms will dominate the sum and the index will assume values different from 0

- Coefficient of skewness of stock A: -0.055
- Coefficient of skewness of stock B: 0.856
- Coefficient of skewness of stock C: -0.928

Kurtosis



Which is the main difference you see in these distributions?

Kurtosis

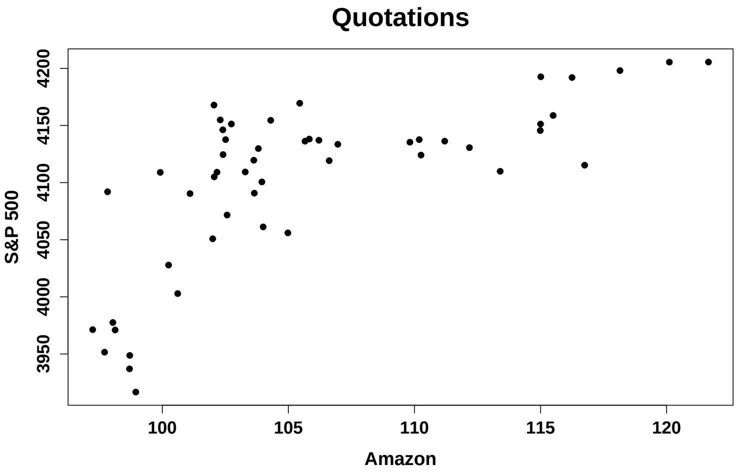
The kurtosis describes the behaviour of the tails of a distribution. The coefficient of kurtosis is

$$\frac{1}{n\ \mathrm{sd}(x)^4}\sum_{i=1}^{n}(x_i-\bar{x})^4,$$

where n is the sample size, $\mathrm{sd}(x)$ the standard deviation of the sample and $\bar{x}$ its mean.

- Coefficient of kurtosis of sample A: 2.45
- Coefficient of kurtosis of sample B: 1.77

Relationship among variables



Amazon stock prices and S&P 500 value between March 17, 2023 and May 31, 2023

The strength of the linear relationship between two quantitative variables can be quantified using the covariance.

Let $x_1, \dots, x_n$ be the value assumed by the first variable in the sample and $y_1, \dots, y_n$ those of the second variable, the covariance is

$$\text{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y},$$

where $\bar{x}$ is the mean of variable x and $\bar{y}$ is the mean of variable y.

The covariance between Amazon and S&P500 quotations is 332.36. How can we interpret this number?

The correlation is easily interpretable

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{\text{sd}(x)\text{sd}(y)},$$

where $\text{sd}(x)$ is the standard deviation of x and $\text{sd}(y)$ is the standard deviation of y. The correlation is limited

$$-1 \leqslant \text{cor}(x, y) \leqslant 1.$$

so its interpretation is easier.

The correlation between Amazon and S&P500 quotations is 0.70.

Correlation

Some examples