

Elements of Probability and Statistics
LADE

F. Tusell

Dpto. Economía Aplicada III (Estadística y Econometría)

Academic Year 2009–2010

Index

Introduction

Variables and populations

Numerical and non-numerical variables

Frequencies and distributions

Making tables and graphs

Tables

Pie graphs

Bar charts

Summary statistics of univariate distributions

Moments in \mathcal{R}

Dispersion measures

Central moments and variance

Other dispersion statistics

Shape statistics

Skewness

Kurtosis

Linear transformations and summary statistics

What is Statistics?

- This is the answer (abridged) we get from the Webster dictionary:

What is Statistics?

- This is the answer (abridged) we get from the Webster dictionary:
 1. *The science which has to do with the collection, classification, and analysis of facts of a numerical nature regarding any topic.*
[...]

What is Statistics?

- This is the answer (abridged) we get from the Webster dictionary:
 1. *The science which has to do with the collection, classification, and analysis of facts of a numerical nature regarding any topic.
[...]*
 2. *Classified facts of a numerical nature regarding any topic.
Specifically [...]*

What is Statistics?

- This is the answer (abridged) we get from the Webster dictionary:
 1. *The science which has to do with the collection, classification, and analysis of facts of a numerical nature regarding any topic.
[...]*
 2. *Classified facts of a numerical nature regarding any topic.
Specifically [...]*
 3. *The branch of mathematics which studies methods for the calculation of probabilities.*

What is Statistics?

- This is the answer (abridged) we get from the Webster dictionary:
 1. *The science which has to do with the collection, classification, and analysis of facts of a numerical nature regarding any topic. [...]*
 2. *Classified facts of a numerical nature regarding any topic. Specifically [...]*
 3. *The branch of mathematics which studies methods for the calculation of probabilities.*
 4. *A branch of applied mathematics concerned with the collection and interpretation of quantitative data and the use of probability theory to estimate population parameters [...]*

What is Statistics?

- The dictionary goes on to saying:

What is Statistics?

- The dictionary goes on to saying:

The subject of statistics can be divided into descriptive statistics - describing data, and analytical statistics - drawing conclusions from data.

What is Statistics?

- The dictionary goes on to saying:

The subject of statistics can be divided into descriptive statistics - describing data, and analytical statistics - drawing conclusions from data.

- We will do both.

What is Statistics?

- The dictionary goes on to saying:

The subject of statistics can be divided into descriptive statistics - describing data, and analytical statistics - drawing conclusions from data.

- We will do both.
- In the first half of the course, we will use loosely words such as “statistical variable”, “population”, etc.

What is Statistics?

- The dictionary goes on to saying:

The subject of statistics can be divided into descriptive statistics - describing data, and analytical statistics - drawing conclusions from data.

- We will do both.
- In the first half of the course, we will use loosely words such as “statistical variable”, “population”, etc.
- In the second half, we will lay the foundations of a mathematical model and define things precisely. Everything will acquire new meaning.

What is a population?

- We start with a collection of subjects or entities: the **population**.

What is a population?

- We start with a collection of subjects or entities: the **population**.
- Examples of populations would be:

What is a population?

- We start with a collection of subjects or entities: the **population**.
- Examples of populations would be:
 1. *All the fish in given a lake.*

What is a population?

- We start with a collection of subjects or entities: the **population**.
- Examples of populations would be:
 1. *All the fish in given a lake.*
 2. *All the students at Sarriko.*

What is a population?

- We start with a collection of subjects or entities: the **population**.
- Examples of populations would be:
 1. *All the fish in given a lake.*
 2. *All the students at Sarriko.*
 3. *All female students aged between 18 and 22 in the European Union.*

What is a population?

- We start with a collection of subjects or entities: the **population**.
- Examples of populations would be:
 1. *All the fish in given a lake.*
 2. *All the students at Sarriko.*
 3. *All female students aged between 18 and 22 in the European Union.*
- It is important to unambiguously state what is our target population: “employed people” is not much of a definition, “people employed in the sense of ICSE-93, group (1)” might be better.

What is a population?

- We start with a collection of subjects or entities: the **population**.
- Examples of populations would be:
 1. *All the fish in given a lake.*
 2. *All the students at Sarriko.*
 3. *All female students aged between 18 and 22 in the European Union.*
- It is important to unambiguously state what is our target population: “employed people” is not much of a definition, “people employed in the sense of ICSE-93, group (1)” might be better.
- The examples given were rather loose!

What about statistical variables?

- For the time being, we will call “statistical variable” any characteristic that can be observed or measured, at least in principle, for each member of the population.

What about statistical variables?

- For the time being, we will call “statistical variable” any characteristic that can be observed or measured, at least in principle, for each member of the population.
- Examples would be:

What about statistical variables?

- For the time being, we will call “statistical variable” any characteristic that can be observed or measured, at least in principle, for each member of the population.
- Examples would be:
 1. *The individual weight of each fish in a lake.*

What about statistical variables?

- For the time being, we will call “statistical variable” any characteristic that can be observed or measured, at least in principle, for each member of the population.
- Examples would be:
 1. *The individual weight of each fish in a lake.*
 2. *The number of subjects each student at Sarriko has registered for.*

What about statistical variables?

- For the time being, we will call “statistical variable” any characteristic that can be observed or measured, at least in principle, for each member of the population.
- Examples would be:
 1. *The individual weight of each fish in a lake.*
 2. *The number of subjects each student at Sarriko has registered for.*
 3. *The number of children each woman in the European Union has had.*

What about statistical variables?

- For the time being, we will call “statistical variable” any characteristic that can be observed or measured, at least in principle, for each member of the population.
- Examples would be:
 1. *The individual weight of each fish in a lake.*
 2. *The number of subjects each student at Sarriko has registered for.*
 3. *The number of children each woman in the European Union has had.*
- Again, it is important to unambiguously state what we are measuring.

What about statistical variables?

- For the time being, we will call “statistical variable” any characteristic that can be observed or measured, at least in principle, for each member of the population.
- Examples would be:
 1. *The individual weight of each fish in a lake.*
 2. *The number of subjects each student at Sarriko has registered for.*
 3. *The number of children each woman in the European Union has had.*
- Again, it is important to unambiguously state what we are measuring.
- Examples given can all be coded as numbers. This need not be so!

Numerical statistical variables

- Let's consider numerical statistical variables first. We will distinguish

Numerical statistical variables

- Let's consider numerical statistical variables first. We will distinguish
 - Continuous variables.

Numerical statistical variables

- Let's consider numerical statistical variables first. We will distinguish
 - Continuous variables.
 - Discrete variables.

Numerical statistical variables

- Let's consider numerical statistical variables first. We will distinguish
 - Continuous variables.
 - Discrete variables.
- Some examples first:

Numerical statistical variables

- Let's consider numerical statistical variables first. We will distinguish
 - Continuous variables.
 - Discrete variables.
- Some examples first:
 - Number of brothers a person has \rightarrow *discrete*.

Numerical statistical variables

- Let's consider numerical statistical variables first. We will distinguish
 - Continuous variables.
 - Discrete variables.
- Some examples first:
 - Number of brothers a person has \longrightarrow *discrete*.
 - Voltage of each battery from a population of batteries \longrightarrow *continuous*.

Numerical statistical variables

- Let's consider numerical statistical variables first. We will distinguish
 - Continuous variables.
 - Discrete variables.
- Some examples first:
 - Number of brothers a person has \rightarrow *discrete*.
 - Voltage of each battery from a population of batteries \rightarrow *continuous*.
 - Number of TV sets of each family unit \rightarrow *discrete*.

Numerical statistical variables

- Let's consider numerical statistical variables first. We will distinguish
 - Continuous variables.
 - Discrete variables.
- Some examples first:
 - Number of brothers a person has \rightarrow *discrete*.
 - Voltage of each battery from a population of batteries \rightarrow *continuous*.
 - Number of TV sets of each family unit \rightarrow *discrete*.
- See the pattern?

What makes a variable discrete?

- Discrete variables have a (usually small) set of possible values.

What makes a variable discrete?

- Discrete variables have a (usually small) set of possible values.
- There is “nothing in between” those values: you can have 2 or 3 brothers and sisters, but not 2.74.

What makes a variable discrete?

- Discrete variables have a (usually small) set of possible values.
- There is “nothing in between” those values: you can have 2 or 3 brothers and sisters, but not 2.74.
- Still, they have a well defined numerical meaning: 4 brothers and sisters is exactly twice as many as 2, and four times as many as one.

What makes a variable continuous?

- Basically, any value between two admissible values has to be an admissible value.

What makes a variable continuous?

- Basically, any value between two admissible values has to be an admissible value.
- In X can take the values 4 and 5, 4.7 or 4.12 should also be legitimate values.

What makes a variable continuous?

- Basically, any value between two admissible values has to be an admissible value.
- In X can take the values 4 and 5, 4.7 or 4.12 should also be legitimate values.
- Sometimes, a variable is intrinsically discrete, but the number of possible states is so large, and they are so “dense”, that it is used as if it were continuous.

What makes a variable continuous?

- Basically, any value between two admissible values has to be an admissible value.
- In X can take the values 4 and 5, 4.7 or 4.12 should also be legitimate values.
- Sometimes, a variable is intrinsically discrete, but the number of possible states is so large, and they are so “dense”, that it is used as if it were continuous.
- Examples: Mass of a body, income measured in small units, days elapsed since a distant date. . .

Qualitative variables

- These are variables taking usually one of a (usually small) number of possible states or “values”, that need not be numerical.

Qualitative variables

- These are variables taking usually one of a (usually small) number of possible states or “values”, that need not be numerical.
- These states can sometimes be coded with numbers.

Qualitative variables

- These are variables taking usually one of a (usually small) number of possible states or “values”, that need not be numerical.
- These states can sometimes be coded with numbers.
- **Example 1:** gender (male or female). One can use the convention of coding males as 0 and females as 1.

Qualitative variables

- These are variables taking usually one of a (usually small) number of possible states or “values”, that need not be numerical.
- These states can sometimes be coded with numbers.
- **Example 1:** gender (male or female). One can use the convention of coding males as 0 and females as 1.
- **Example 2:** nationality. One can use the convention France=1, Britain=2, Germany=3, Poland=4.

Qualitative variables

- These are variables taking usually one of a (usually small) number of possible states or “values”, that need not be numerical.
- These states can sometimes be coded with numbers.
- **Example 1:** gender (male or female). One can use the convention of coding males as 0 and females as 1.
- **Example 2:** nationality. One can use the convention France=1, Britain=2, Germany=3, Poland=4.
- **NOTICE:** Polish is not “twice” British! The values of the number codes have no meaning, and cannot be used as the values of a numerical variable.

Nominal and ordinal variables

- Gender (male and female), nationality, religion, etc. are all variables with no natural order. We call them **nominal**.

Nominal and ordinal variables

- Gender (male and female), nationality, religion, etc. are all variables with no natural order. We call them **nominal**.
- Consider a variable taking values: “Never”, “Seldom”, “Sometimes”, “Quite often”, “Always”.

Nominal and ordinal variables

- Gender (male and female), nationality, religion, etc. are all variables with no natural order. We call them **nominal**.
- Consider a variable taking values: “Never”, “Seldom”, “Sometimes”, “Quite often”, “Always”.
- These five “values” or states have a meaningful natural order. “Seldom” seems closer to “Never” than “Always”, for instance.

Nominal and ordinal variables

- Gender (male and female), nationality, religion, etc. are all variables with no natural order. We call them **nominal**.
- Consider a variable taking values: “Never”, “Seldom”, “Sometimes”, “Quite often”, “Always”.
- These five “values” or states have a meaningful natural order. “Seldom” seems closer to “Never” than “Always”, for instance.
- We cannot meaningfully assign values to those states. *However*, there is at least a natural order. We call these variables **ordinal**.

Hierarchy of statistical variables

- There is roughly a ranking of "sophistication" among statistical variables:

Nominal < Ordinal < Discrete < Continuous

Everything we can do with a variable, we can do with another type to the right.

Hierarchy of statistical variables

- There is roughly a ranking of "sophistication" among statistical variables:

Nominal < Ordinal < Discrete < Continuous

Everything we can do with a variable, we can do with another type to the right.

- But not the other way around!

Hierarchy of statistical variables

- There is roughly a ranking of "sophistication" among statistical variables:

Nominal < Ordinal < Discrete < Continuous

Everything we can do with a variable, we can do with another type to the right.

- But not the other way around!
- Coding qualitative variables with number and using those numbers as rightfully numerical variables is about the worst thing a data analyst can do!

Hierarchy of statistical variables

- There is roughly a ranking of "sophistication" among statistical variables:

Nominal < Ordinal < Discrete < Continuous

Everything we can do with a variable, we can do with another type to the right.

- But not the other way around!
- Coding qualitative variables with number and using those numbers as rightfully numerical variables is about the worst thing a data analyst can do!
- *Pleeeeeeease, don't do that yourselves!*

How do we convert among different types?

- Only in one direction!

How do we convert among different types?

- Only in one direction!
- We can construct intervals and make categories.

How do we convert among different types?

- Only in one direction!
- We can construct intervals and make categories.
- For instance, with (continuous) income measured in €, we can construct an (ordinal) variable with values:

Less than 6000€

6000€ - 12000€

Over 12000 €

How do we convert among different types?

- Only in one direction!
- We can construct intervals and make categories.
- For instance, with (continuous) income measured in €, we can construct an (ordinal) variable with values:

Less than 6000€

6000€ - 12000€

Over 12000 €

- There is always some loss of information (but perhaps an increase in interpretability).

Absolute and relative frequency

- We seek information about the whole set of values of the statistical variable.

Absolute and relative frequency

- We seek information about the whole set of values of the statistical variable.
- The number of times the variable takes a given value i , is called the **absolute frequency**, denoted by n_i .

Absolute and relative frequency

- We seek information about the whole set of values of the statistical variable.
- The number of times the variable takes a given value i , is called the **absolute frequency**, denoted by n_i .
- If we divide n_i by N , the total number of members in the population, we get the **relative frequency**,

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

Absolute and relative frequency

- We seek information about the whole set of values of the statistical variable.
- The number of times the variable takes a given value i , is called the **absolute frequency**, denoted by n_i .
- If we divide n_i by N , the total number of members in the population, we get the **relative frequency**,

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

- Clearly, since $\sum_i n_i = N$, we must have $\sum_i f_i = 1$.

Cumulative frequency

- The **cumulative frequency** F_i is defined as

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

Cumulative frequency

- The **cumulative frequency** F_i is defined as

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

- F_i counts the number of cases whose value of the statistical variable considered is smaller than or equal i .

Cumulative frequency

- The **cumulative frequency** F_i is defined as

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

- F_i counts the number of cases whose value of the statistical variable considered is smaller than or equal i .
- Clearly, for all values i ,

$$0 \leq F_i \leq 1$$

Cumulative frequency

- The **cumulative frequency** F_i is defined as

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

- F_i counts the number of cases whose value of the statistical variable considered is smaller than or equal i .
- Clearly, for all values i ,

$$0 \leq F_i \leq 1$$

- Sometimes also called the "distribution" of the variable.

An example (I)

- Consider a population of $N = 50$ students. After being asked how many languages they speak fluently, their answers are tabulated as follows:

Languages	Students
i	n_i
1	10
2	24
3	11
4	5

An example (II)

- We can compute the relative frequencies and the cumulated frequencies as follows:

Languages	Students	Frequencies:	
i	n_i	Relative (f_i)	Cumulated (F_i)
1	10	0,20	0,20
2	24	0,48	0,68
3	11	0,22	0,90
4	5	0,10	1,00
	50	1,00	

Nominal vs. nominal

- No big deal! We simply arrange the categories of each of two variables in rows and columns.

Nominal vs. nominal

- No big deal! We simply arrange the categories of each of two variables in rows and columns.
- **Example:**

Eye color	Hair color:			
	Blonde	Brown	Black	Red
Blue	40	10	2	3
Grey	24	12	34	29
Dark	1	22	43	12

Nominal vs. ordinal

- Same. It just makes sense that the categories of the ordinal variable appear “in order”.

Nominal vs. ordinal

- Same. It just makes sense that the categories of the ordinal variable appear “in order”.
- **Example:**

TV channel	Watches:			
	Never	Sometimes	Often	Always
TV1-TV2	40	10	2	3
ETB-1	24	12	34	29
ETB-2	1	22	43	12
Others	11	12	3	22

Nominal vs. ordinal

- Same. It just makes sense that the categories of the ordinal variable appear “in order”.
- **Example:**

TV channel	Watches:			
	Never	Sometimes	Often	Always
TV1-TV2	40	10	2	3
ETB-1	24	12	34	29
ETB-2	1	22	43	12
Others	11	12	3	22

- Same with ordinal vs. ordinal or with discrete variables.

One variable continuous

- Now, we would have a very large number of categories in rows or columns.

One variable continuous

- Now, we would have a very large number of categories in rows or columns.
- To present information more clearly, we group the continuous variable in intervals.

Age (years)	Reads local press:			
	Never	Sometimes	Often	Always
[0, 18)	40	8	2	3
[18, 35)	4	22	34	29
[35, 65)	1	22	43	12
[65, ∞)	11	12	3	12

One variable continuous

- Now, we would have a very large number of categories in rows or columns.
- To present information more clearly, we group the continuous variable in intervals.

Age (years)	Reads local press:			
	Never	Sometimes	Often	Always
[0, 18)	40	8	2	3
[18, 35)	4	22	34	29
[35, 65)	1	22	43	12
[65, ∞)	11	12	3	12

- We have to specify unambiguously the interval extremes.

How do we chose the “breaks”?

- Pick boundaries which help interpretation.

How do we chose the “breaks”?

- Pick boundaries which help interpretation.
- In the previous case we used:

$[0, 18)$ $[18, 35)$ $[35, 65)$ $[65, \infty)$

How do we chose the “breaks”?

- Pick boundaries which help interpretation.
- In the previous case we used:

$[0, 18)$ $[18, 35)$ $[35, 65)$ $[65, \infty)$

- 18 is the age majority; 65 the usual limit of working life.

How do we chose the “breaks”?

- Pick boundaries which help interpretation.
- In the previous case we used:

$[0, 18)$ $[18, 35)$ $[35, 65)$ $[65, \infty)$

- 18 is the age majority; 65 the usual limit of working life.
- Pick fine-grained detail where you need it.

How do we chose the “breaks”?

- Pick boundaries which help interpretation.
- In the previous case we used:

$[0, 18)$ $[18, 35)$ $[35, 65)$ $[65, \infty)$

- 18 is the age majority; 65 the usual limit of working life.
- Pick fine-grained detail where you need it.
- If in doubt, pick intervals roughly of equal size.

Class marks

- A value representing the interval or class.

Class marks

- A value representing the interval or class.
- In the previous case, we could pick:

$$[0, 18) \longrightarrow 9;$$

$$[18, 35) \longrightarrow 26;$$

$$[35, 65) \longrightarrow 55;$$

$$[65, \infty) \longrightarrow 75$$

Class marks

- A value representing the interval or class.
- In the previous case, we could pick:

$$[0, 18) \longrightarrow 9;$$

$$[18, 35) \longrightarrow 26;$$

$$[35, 65) \longrightarrow 55;$$

$$[65, \infty) \longrightarrow 75$$

- Usually, half way or close to it.

Class marks

- A value representing the interval or class.
- In the previous case, we could pick:

$$[0, 18) \longrightarrow 9;$$

$$[18, 35) \longrightarrow 26;$$

$$[35, 65) \longrightarrow 55;$$

$$[65, \infty) \longrightarrow 75$$

- Usually, half way or close to it.
- In the old days, a computation saving device. This is no longer needed.

Class marks

- A value representing the interval or class.
- In the previous case, we could pick:

$$[0, 18) \rightarrow 9;$$

$$[18, 35) \rightarrow 26;$$

$$[35, 65) \rightarrow 55;$$

$$[65, \infty) \rightarrow 75$$

- Usually, half way or close to it.
- In the old days, a computation saving device. This is no longer needed.
- Useful when making graphs –more on that later.

Using class marks for quick computations

- The idea is to group cases in intervals and *pretend* that all cases in an interval have the same value —the class mark.

Using class marks for quick computations

- The idea is to group cases in intervals and *pretend* that all cases in an interval have the same value —the class mark.
- In the previous case,

$$\begin{aligned}[0, 18) &\longrightarrow 9; \\ [18, 35) &\longrightarrow 26; \\ [35, 65) &\longrightarrow 55; \\ [65, \infty) &\longrightarrow 75\end{aligned}$$

assume we have 5, 13, 24, and 12 cases in each of the intervals. The approximate “age average” could be computed as

$$\frac{5 \times 9 + 13 \times 26 + 24 \times 55 + 12 \times 75}{5 + 13 + 24 + 12} = \frac{2603}{54} = 48,2$$

Using class marks for quick computations

- The idea is to group cases in intervals and *pretend* that all cases in an interval have the same value —the class mark.
- In the previous case,

$$\begin{aligned} [0, 18) &\longrightarrow 9; \\ [18, 35) &\longrightarrow 26; \\ [35, 65) &\longrightarrow 55; \\ [65, \infty) &\longrightarrow 75 \end{aligned}$$

assume we have 5, 13, 24, and 12 cases in each of the intervals. The approximate “age average” could be computed as

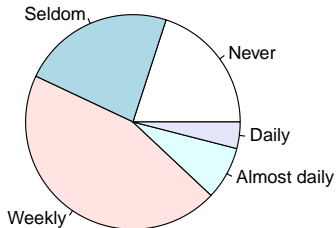
$$\frac{5 \times 9 + 13 \times 26 + 24 \times 55 + 12 \times 75}{5 + 13 + 24 + 12} = \frac{2603}{54} = 48,2$$

- Much faster than considering the exact age of each case.

The infamous pie chart (I)

- When people say "business graph" they usually mean this ugly thing:

How often do you read journals?



The infamous pie chart (II)

- Why people is so fond of this graph, I have no clue.

The infamous pie chart (II)

- Why people is so fond of this graph, I have no clue.
- Probably, they know no better.

The infamous pie chart (II)

- Why people is so fond of this graph, I have no clue.
- Probably, they know no better.
- No particular cognitive advantages.

The infamous pie chart (II)

- Why people is so fond of this graph, I have no clue.
- Probably, they know no better.
- No particular cognitive advantages.
- Lots of ink to display just five percentages!

The infamous pie chart (II)

- Why people is so fond of this graph, I have no clue.
- Probably, they know no better.
- No particular cognitive advantages.
- Lots of ink to display just five percentages!
- Rather than this, use stacked bars to compare categories.

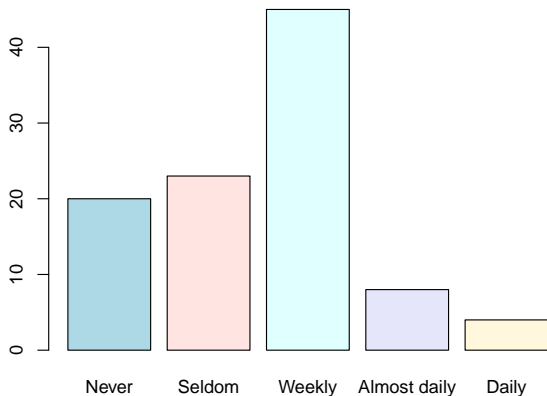
The infamous pie chart (II)

- Why people is so fond of this graph, I have no clue.
- Probably, they know no better.
- No particular cognitive advantages.
- Lots of ink to display just five percentages!
- Rather than this, use stacked bars to compare categories.
- Do yourself a favour, do your readers a favour! No pie charts, please! We can all kill this nasty beast forever!

What to do instead a pie chart (I)

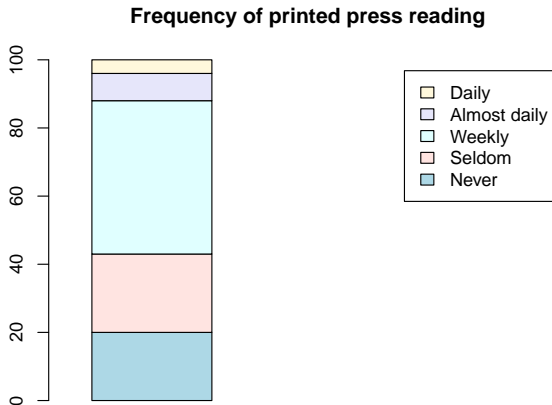
- A bar chart is usually a good idea.

How often do you read journals?



What to do instead a pie chart (II)

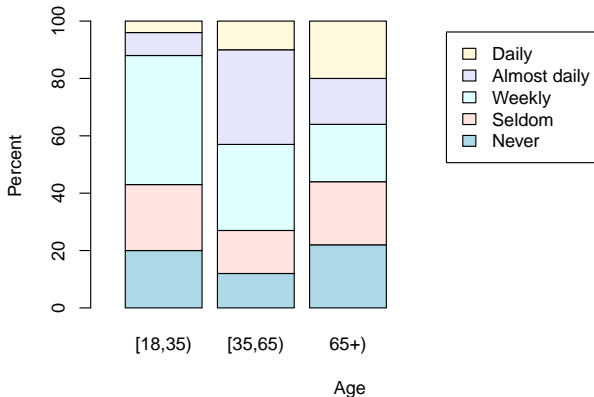
- We can stack the bars:



What to do in place of a pie chart (III)

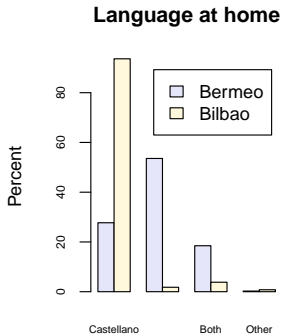
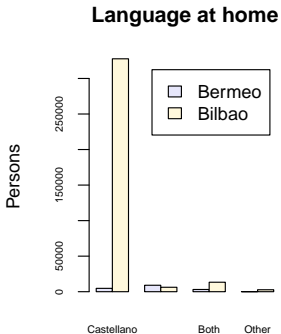
- Seems also a waste for five percentages. The real benefit comes when we plot besides several stacked bars.

How often do you read journals?



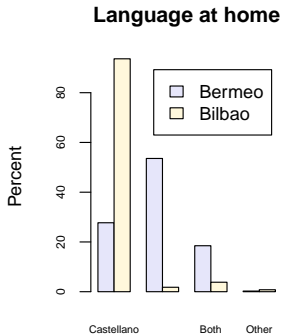
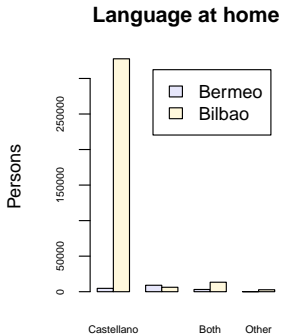
Absolute vs. relative frequencies

- You can use counts or relative frequencies.



Absolute vs. relative frequencies

- You can use counts or relative frequencies.

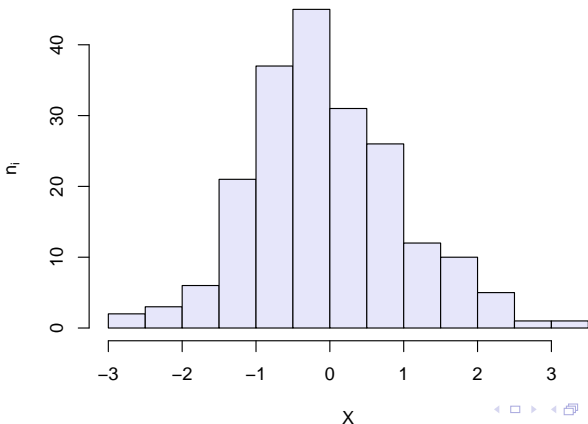


- These two bar charts tell a different story!

Histograms

- A bar chart representing the distribution of a continuous variable.

Histogram of a variable (N=200)



Absolute vs. relative frequencies

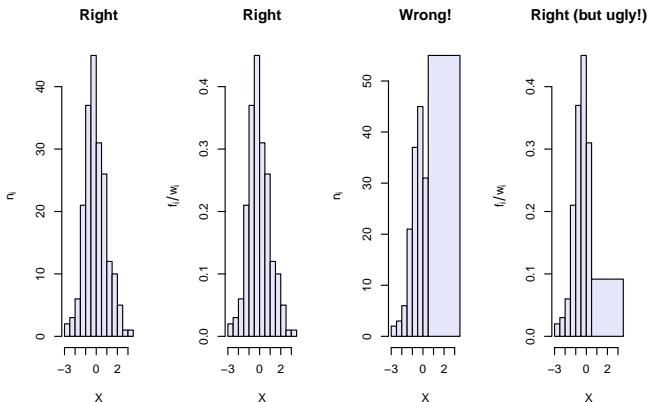
- As with bar charts, you can use relative or absolute frequencies.

Absolute vs. relative frequencies

- As with bar charts, you can use relative or absolute frequencies.
- But with different interval widths, **MUST** use f_i/w_i

Absolute vs. relative frequencies

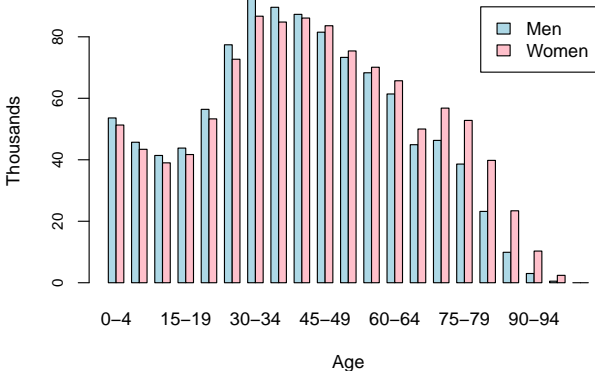
- As with bar charts, you can use relative or absolute frequencies.
- But with different interval widths, **MUST** use f_i/w_i
- Using n_i with varying widths gives the wrong idea.



Adjacent Histograms

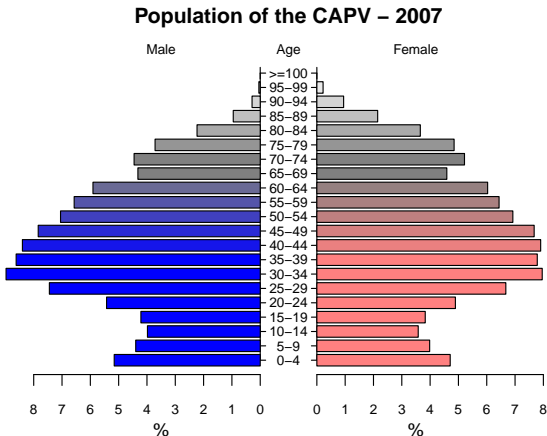
- Useful to show the differences between two groups:

Age distribution in the CAPV – 2007



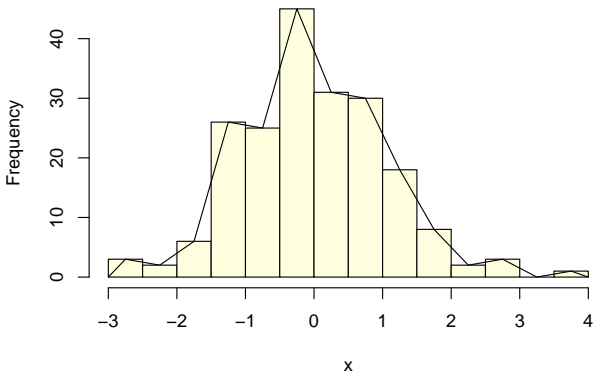
Population Pyramid

- A common alternative to the preceding graph:



Frequency polygon

- Simply draw a broken line between the extremes of the range through the center point of each histogram's box top side.



Moments

- Histograms, tables, etc. give us an idea of the values taken by an statistical variable.

Moments

- Histograms, tables, etc. give us an idea of the values taken by an statistical variable.
- Often, we want summaries that condense in a few numbers properties of the distributions.

Moments

- Histograms, tables, etc. give us an idea of the values taken by an statistical variable.
- Often, we want summaries that condense in a few numbers properties of the distributions.
- Moments are useful for that purpose.

Moments

- Histograms, tables, etc. give us an idea of the values taken by an statistical variable.
- Often, we want summaries that condense in a few numbers properties of the distributions.
- Moments are useful for that purpose.
- They come into two flavours: ordinary moments (or “moments about the origin”) and centered moments.

Moments about the origin

- Assume a statistical variable taking values x_1, x_2, \dots, x_N . We define the h -th moment about the origin a_h as follows:

$$a_h = \frac{1}{N} \sum_{i=1}^N x_i^h$$

Moments about the origin

- Assume a statistical variable taking values x_1, x_2, \dots, x_N . We define the h -th moment about the origin a_h as follows:

$$a_h = \frac{1}{N} \sum_{i=1}^N x_i^h$$

- If there are repeated values, the formula above becomes:

$$a_h = \frac{1}{N} \sum_{i=1}^k n_i x_i^h = \sum_{i=1}^k \frac{n_i}{N} x_i^h = \sum_{i=1}^k f_i x_i^h$$

Moments about the origin

- Assume a statistical variable taking values x_1, x_2, \dots, x_N . We define the h -th moment about the origin a_h as follows:

$$a_h = \frac{1}{N} \sum_{i=1}^N x_i^h$$

- If there are repeated values, the formula above becomes:

$$a_h = \frac{1}{N} \sum_{i=1}^k n_i x_i^h = \sum_{i=1}^k \frac{n_i}{N} x_i^h = \sum_{i=1}^k f_i x_i^h$$

- When $h = 1$, this definition gives us:

$$a_1 = \frac{1}{N} \sum_{i=1}^N x_i = \bar{x};$$

the usual average or arithmetic mean.

a_1 as a location indicator

- This is an indicator of *location*.

a_1 as a location indicator

- This is an indicator of *location*.
- But, beware! The observations need not be anywhere near the mean \bar{x} !

a_1 as a location indicator

- This is an indicator of *location*.
- But, beware! The observations need not be anywhere near the mean \bar{x} !
- For instance, the three values 2, 3, 4 have mean 3; and three is reasonable close to each of them.

a_1 as a location indicator

- This is an indicator of *location*.
- But, beware! The observations need not be anywhere near the mean \bar{x} !
- For instance, the three values 2, 3, 4 have mean 3; and three is reasonable close to each of them.
- But -20, 9 and +20 also have mean 3, which is not close to any of the observed values.

a_1 as a location indicator

- This is an indicator of *location*.
- But, beware! The observations need not be anywhere near the mean \bar{x} !
- For instance, the three values 2, 3, 4 have mean 3; and three is reasonable close to each of them.
- But -20, 9 and +20 also have mean 3, which is not close to any of the observed values.
- Assuming the mean is a “representative observation” can be seriously wrong.

a_1 as a location indicator

- This is an indicator of *location*.
- But, beware! The observations need not be anywhere near the mean \bar{x} !
- For instance, the three values 2, 3, 4 have mean 3; and three is reasonable close to each of them.
- But -20, 9 and +20 also have mean 3, which is not close to any of the observed values.
- Assuming the mean is a “representative observation” can be seriously wrong.
- Would you try to sell cars in a country with average income 30.000€? (Perhaps a single person with an astronomic income while all other people are starving.)

Alternative definitions of “mean”

- There are many, with different properties.

Alternative definitions of “mean”

- There are many, with different properties.
- For instance,

Geometric mean $\left(\prod_{i=1}^N x_i\right)^{\frac{1}{N}}$

Harmonic mean $N \left(\sum_{i=1}^N \frac{1}{x_i}\right)^{-1}$

Generalized mean $\left(\frac{1}{N} \sum_{i=1}^N x_i^m\right)^{\frac{1}{m}} = a_m^{1/m}$

Alternative definitions of “mean”

- There are many, with different properties.
- For instance,

Geometric mean $\left(\prod_{i=1}^N x_i\right)^{\frac{1}{N}}$

Harmonic mean $N \left(\sum_{i=1}^N \frac{1}{x_i}\right)^{-1}$

Generalized mean $\left(\frac{1}{N} \sum_{i=1}^N x_i^m\right)^{\frac{1}{m}} = a_m^{1/m}$

- You may investigate their properties in any book on descriptive statistics...

Alternative definitions of “mean”

- There are many, with different properties.
- For instance,

Geometric mean $\left(\prod_{i=1}^N x_i\right)^{\frac{1}{N}}$

Harmonic mean $N \left(\sum_{i=1}^N \frac{1}{x_i}\right)^{-1}$

Generalized mean $\left(\frac{1}{N} \sum_{i=1}^N x_i^m\right)^{\frac{1}{m}} = a_m^{1/m}$

- You may investigate their properties in any book on descriptive statistics...
- ... or in <http://en.wikipedia.org/wiki/Mean> if you prefer.

Weighted means

- Sometimes we want to give more weight to some elements.

Weighted means

- Sometimes we want to give more weight to some elements.
- We can use a **weighted mean**

$$\bar{x}_p = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i}$$

Weighted means

- Sometimes we want to give more weight to some elements.
- We can use a **weighted mean**

$$\bar{x}_p = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i}$$

- For instance, we may want a grade point average in which Accounting weights twice as much as other subjects and History three times as much. We would compute

$$\bar{x}_p = \frac{2 \times (\text{Accounting}) + 3 \times (\text{History}) + \dots}{2 + 3 + \dots}$$

The mode (I)

- The **mode** is the value which repeats itself most.

The mode (I)

- The **mode** is the value which repeats itself most.
- It gives information on “the most typical” value.

The mode (I)

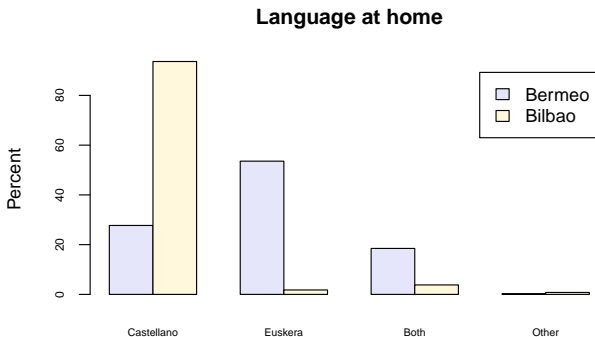
- The **mode** is the value which repeats itself most.
- It gives information on “the most typical” value.
- Valid with both quantitative and qualitative variables.

The mode (I)

- The **mode** is the value which repeats itself most.
- It gives information on “the most typical” value.
- Valid with both quantitative and qualitative variables.
- It need not be unique (and in that case we talk about “multimodal variables” or “multimodal distributions”).

The mode (II)

- Mode for Bilbao is “Castellano”, for Bermeo, “Euskera”.



The median

- The value which leaves one half of the observations to each side.

The median

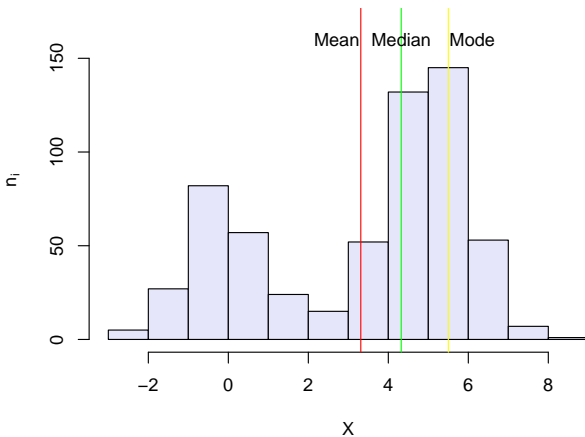
- The value which leaves one half of the observations to each side.
- May not be unique.

The median

- The value which leaves one half of the observations to each side.
- May not be unique.
- If even number of observations, half point of the two central.

Location statistics

Histogram of a variable (N=600)



Central (or centered) moments

- Same definition, but a constant c subtracted from each value x_i (and then we talk of *moments about c*):

$$a_{h,c} = \frac{1}{N} \sum_{i=1}^N (x_i - c)^h$$

Central (or centered) moments

- Same definition, but a constant c subtracted from each value x_i (and then we talk of *moments about c*):

$$a_{h,c} = \frac{1}{N} \sum_{i=1}^N (x_i - c)^h$$

- If c is taken to be the mean, \bar{x} , then we talk of *central moments* or *moments about the mean*.

$$m_h = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^h$$

Central (or centered) moments

- Same definition, but a constant c subtracted from each value x_i (and then we talk of *moments about c*):

$$a_{h,c} = \frac{1}{N} \sum_{i=1}^N (x_i - c)^h$$

- If c is taken to be the mean, \bar{x} , then we talk of *central moments* or *moments about the mean*.

$$m_h = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^h$$

- By far and away, the most important is the *variance*, or second order moment about the mean:

$$m_2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

Other dispersion statistics (I)

- The **standard deviation** S_x , is simply the square root of the variance.

Other dispersion statistics (I)

- The **standard deviation** S_x , is simply the square root of the variance.
- The **coefficient of variation** is

$$g_0 = \frac{S_x}{\bar{X}};$$

it is similar to S_x but “normalized” to the mean.

Other dispersion statistics (I)

- The **standard deviation** S_x , is simply the square root of the variance.
- The **coefficient of variation** is

$$g_0 = \frac{S_x}{\bar{X}};$$

it is similar to S_x but “normalized” to the mean.

- We usually say that the mean is “representative” if $g_0 < 1$, i.e. the standard deviation small compared to the mean.

Other dispersion statistics (I)

- The **standard deviation** S_x , is simply the square root of the variance.
- The **coefficient of variation** is

$$g_0 = \frac{S_x}{\bar{X}};$$

it is similar to S_x but “normalized” to the mean.

- We usually say that the mean is “representative” if $g_0 < 1$, i.e. the standard deviation small compared to the mean.
- No clear-cut thresholds, each author may use a different criterion.

Other dispersion statistics (II)

- The **mean deviation** with respect to c es defined as:

$$\frac{1}{N} \sum_{i=1}^N |x_i - c|$$

Other dispersion statistics (II)

- The **mean deviation** with respect to c es defined as:

$$\frac{1}{N} \sum_{i=1}^N |x_i - c|$$

- Usually taken with respect to the mean or the median.

Other dispersion statistics (II)

- The **mean deviation** with respect to c es defined as:

$$\frac{1}{N} \sum_{i=1}^N |x_i - c|$$

- Usually taken with respect to the mean or the median.
- When taken with respect to the mean, it is different than the standard deviation. In fact, it can be shown that

$$\frac{1}{N} \sum_{i=1}^N |x_i - c| \leq \sqrt{\frac{1}{N} \sum_{i=1}^N |x_i - c|^2}$$

Other dispersion statistics (III)

- The mean deviation is usually taken w.r.t the median.

Other dispersion statistics (III)

- The mean deviation is usually taken w.r.t the median.
- The median is the value with respect to which the mean deviation is minimal.

Other dispersion statistics (III)

- The mean deviation is usually taken w.r.t the median.
- The median is the value with respect to which the mean deviation is minimal.
- Sketch of the proof. Median as "preferred location" for a facility.

Other dispersion statistics (IV)

- The **range** is simply:

$$R = \max x_i - \min x_i$$

Other dispersion statistics (IV)

- The **range** is simply:

$$R = \max x_i - \min x_i$$

- Problem: quite affected by outliers.

Other dispersion statistics (IV)

- The **range** is simply:

$$R = \max x_i - \min x_i$$

- Problem: quite affected by outliers.
- For that reason, another useful measure is the **interquartile range**.

Other dispersion statistics (IV)

- The **range** is simply:

$$R = \max x_i - \min x_i$$

- Problem: quite affected by outliers.
- For that reason, another useful measure is the **interquartile range**.
- Quartiles: $q_{0.25}, q_{0.50}, q_{0.75}$ are the values dividing the range in four equally populated parts.

Other dispersion statistics (IV)

- The **range** is simply:

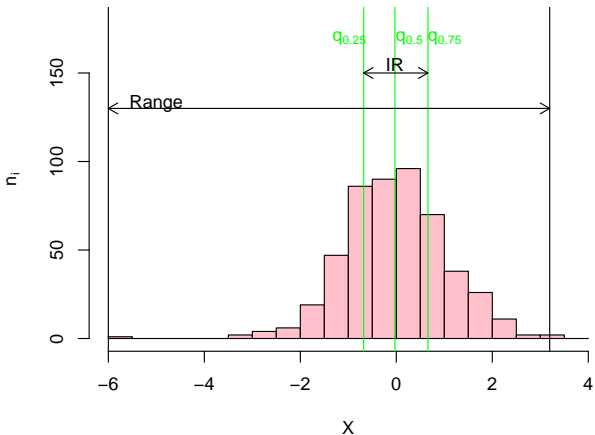
$$R = \max x_i - \min x_i$$

- Problem: quite affected by outliers.
- For that reason, another useful measure is the **interquartile range**.
- Quartiles: $q_{0.25}, q_{0.50}, q_{0.75}$ are the values dividing the range in four equally populated parts.
- $q_{0.50}$ is the same than the median.

Other dispersion statistics (V)

- Notice range is quite affected by an outlier.

Histogram of a variable (N=500)



Other dispersion statistics (VI)

- Quartiles are a special case of quantiles q_α .

Other dispersion statistics (VI)

- Quartiles are a special case of quantiles q_α .
- Other special cases are quintiles, deciles and percentiles.

Other dispersion statistics (VI)

- Quartiles are a special case of quantiles q_α .
- Other special cases are quintiles, deciles and percentiles.
- Deciles divide the range in ten equally populated fragments, quintiles in five equally populated fragments, etc.

Other dispersion statistics (VI)

- Quartiles are a special case of quantiles q_α .
- Other special cases are quintiles, deciles and percentiles.
- Deciles divide the range in ten equally populated fragments, quintiles in five equally populated fragments, etc.
- Some controversy on the definition of quantiles with finite samples.

Other dispersion statistics (VI)

- Quartiles are a special case of quantiles q_α .
- Other special cases are quintiles, deciles and percentiles.
- Deciles divide the range in ten equally populated fragments, quintiles in five equally populated fragments, etc.
- Some controversy on the definition of quantiles with finite samples.
- For instance, in 2, 3, 3, 3, 5, 6, 7, 9, 9 the median is 5. But in 2, 3, 3, 3, 5, 6, 6, 7, 9, 9 the median would be somewhere between 5 and 6.

Other dispersion statistics (VI)

- Quartiles are a special case of quantiles q_α .
- Other special cases are quintiles, deciles and percentiles.
- Deciles divide the range in ten equally populated fragments, quintiles in five equally populated fragments, etc.
- Some controversy on the definition of quantiles with finite samples.
- For instance, in 2, 3, 3, 3, 5, 6, 7, 9, 9 the median is 5. But in 2, 3, 3, 3, 5, 6, 6, 7, 9, 9 the median would be somewhere between 5 and 6.
- If you look at the help page of `quantile` in R you will find... 9 different definitions!

The boxplot (II)

- Instant and very visual information about the main features of the distribution.

The boxplot (II)

- Instant and very visual information about the main features of the distribution.
- Also called “box and whiskers” plot.

The boxplot (II)

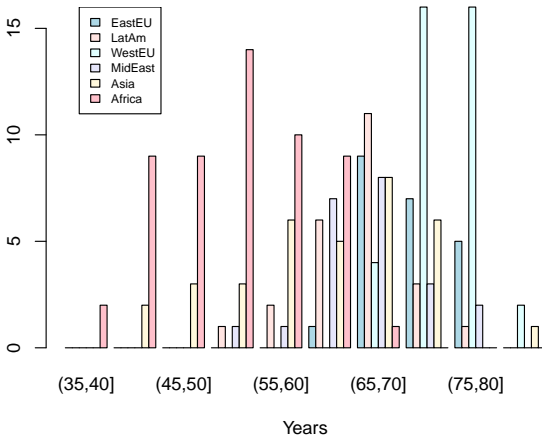
- Instant and very visual information about the main features of the distribution.
- Also called “box and whiskers” plot.
- Particularly nice when it comes to compare subgroups of a population.

The boxplot (II)

- Instant and very visual information about the main features of the distribution.
- Also called “box and whiskers” plot.
- Particularly nice when it comes to compare subgroups of a population.
- Many variants, trying to encode additional information.

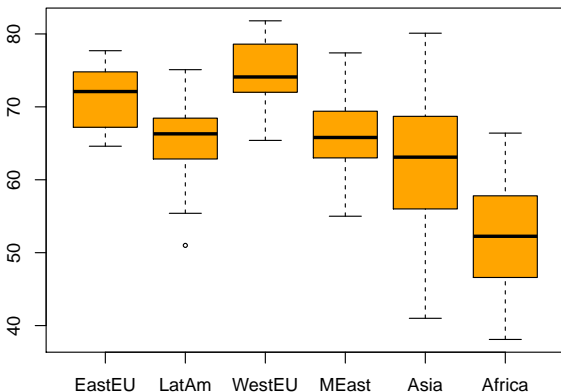
The boxplot (III)

**Countries with a given life expectancy
by geographical area**



The boxplot (IV)

Life expectancy in different countries, grouped by area



Symmetric distributions

- Loosely speaking, the distribution is symmetric when the histogram is symmetric, usually (but not necessarily) around a mode.

Symmetric distributions

- Loosely speaking, the distribution is symmetric when the histogram is symmetric, usually (but not necessarily) around a mode.
- Non-symmetric distributions display a concentration on one side.

Symmetric distributions

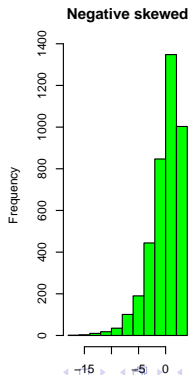
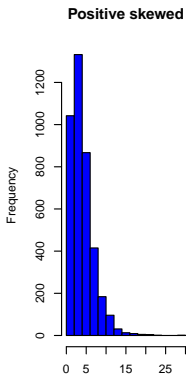
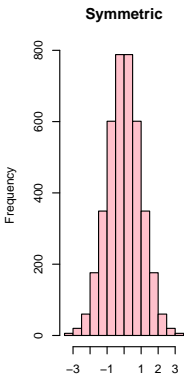
- Loosely speaking, the distribution is symmetric when the histogram is symmetric, usually (but not necessarily) around a mode.
- Non-symmetric distributions display a concentration on one side.
- Both kinds of distributions arise very naturally in applications.

Skew distributions

- If data tends to concentrate on the right side, the distribution is said **negative skewed**. If data concentrates on the left side, the distribution is said **positive skewed**.

Skew distributions

- If data tends to concentrate on the right side, the distribution is said **negative skewed**. If data concentrates on the left side, the distribution is said **positive skewed**.
- In other words, negative skew distributions have long left tails, while positive skew distributions have long right tails.



Symmetry and moments

- If a distribution is symmetric, all central moments of odd order are zero:

$$\begin{aligned}m_{2d+1} &= \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^{2d+1} \\ &= \sum_{j=1}^k f_j (x_j - \bar{x})^{2d+1} \\ &= 0\end{aligned}$$

Symmetry and moments

- If a distribution is symmetric, all central moments of odd order are zero:

$$\begin{aligned}m_{2d+1} &= \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^{2d+1} \\ &= \sum_{j=1}^k f_j (x_j - \bar{x})^{2d+1} \\ &= 0\end{aligned}$$

- Thus, an indicator of skewness would be

$$m_3 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^3,$$

large absolute values being indicative of skewness.

Symmetry and moments

- If a distribution is symmetric, all central moments of odd order are zero:

$$\begin{aligned}m_{2d+1} &= \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^{2d+1} \\ &= \sum_{j=1}^k f_j (x_j - \bar{x})^{2d+1} \\ &= 0\end{aligned}$$

- Thus, an indicator of skewness would be

$$m_3 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^3,$$

large absolute values being indicative of skewness.

Notice...

- The deviations with respect to the mean add up to zero:

$$\sum_{i=1}^N (x_i - \bar{x}) = 0.$$

Notice...

- The deviations with respect to the mean add up to zero:

$$\sum_{i=1}^N (x_i - \bar{x}) = 0.$$

- When raised to a power larger than 1, few “large” deviations weight more than many “small” ones.

Notice...

- The deviations with respect to the mean add up to zero:

$$\sum_{i=1}^N (x_i - \bar{x}) = 0.$$

- When raised to a power larger than 1, few “large” deviations weight more than many “small” ones.
- **Example:** 1, 1, 4. $\bar{x} = 2$. The deviations are -1, -1, +2. When raised to the third power, they are -1, -1 and + 8.

Notice...

- The deviations with respect to the mean add up to zero:
$$\sum_{i=1}^N (x_i - \bar{x}) = 0.$$
- When raised to a power larger than 1, few “large” deviations weight more than many “small” ones.
- **Example:** 1, 1, 4. $\bar{x} = 2$. The deviations are -1, -1, +2. When raised to the third power, they are -1, -1 and + 8.
- Thus, positive skewed distributions will have $m_3 > 0$ while negative skewed distributions will $m_3 < 0$.

Notice...

- The deviations with respect to the mean add up to zero:
$$\sum_{i=1}^N (x_i - \bar{x}) = 0.$$
- When raised to a power larger than 1, few “large” deviations weight more than many “small” ones.
- **Example:** 1, 1, 4. $\bar{x} = 2$. The deviations are -1, -1, +2. When raised to the third power, they are -1, -1 and + 8.
- Thus, positive skewed distributions will have $m_3 > 0$ while negative skewed distributions will $m_3 < 0$.
- m_3 depends on the measurement units (undesirable). Therefore, we seek to normalize it.

Skewness indicators

- The **skewness coefficient** is defined as:

$$\gamma_1 = g_1 = \frac{m_3}{(S_x)^3}$$

Skewness indicators

- The **skewness coefficient** is defined as:

$$\gamma_1 = g_1 = \frac{m_3}{(S_x)^3}$$

- The above definition does not depend on units of measurement (if we multiply all observations by k , both numerator and denominator are multiplied by k^3).

Skewness indicators

- The **skewness coefficient** is defined as:

$$\gamma_1 = g_1 = \frac{m_3}{(S_x)^3}$$

- The above definition does not depend on units of measurement (if we multiply all observations by k , both numerator and denominator are multiplied by k^3).
- An alternative indicator is **Pearson skewness coefficient**:

$$g_{1,P} = \frac{\bar{x} - Me}{(S_x)}$$

Skewness indicators

- The **skewness coefficient** is defined as:

$$\gamma_1 = g_1 = \frac{m_3}{(S_x)^3}$$

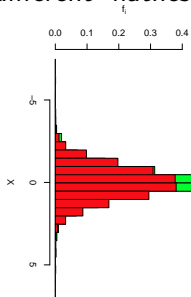
- The above definition does not depend on units of measurement (if we multiply all observations by k , both numerator and denominator are multiplied by k^3).
- An alternative indicator is **Pearson skewness coefficient**:

$$g_{1,P} = \frac{\bar{x} - Me}{(S_x)}$$

- It exploits the idea that long tails tend to dissociate the mean from the median.

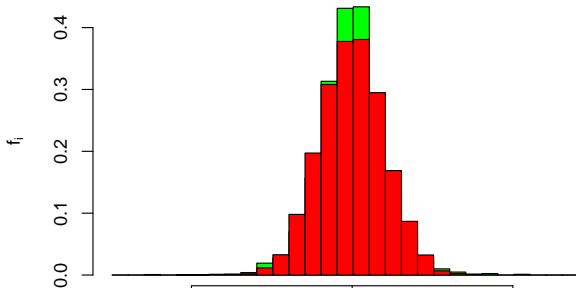
Kurtosis (I)

- It is a measure of “flatness” of a distribution.
- One might think this is already catered for by the variance.
- Not so: the two distributions next have same variance but different “flatness”.



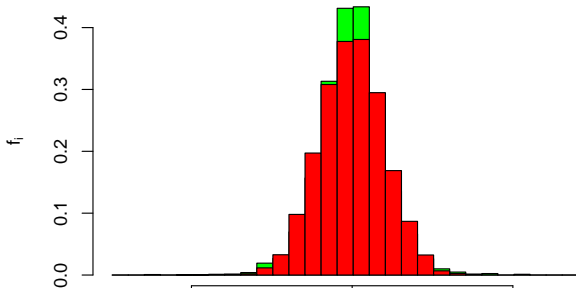
Kurtosis (II)

- The **red** distribution has “average” kurtosis (a Gaussian distribution, that we will study in a few weeks). It is *mesokurtic*.



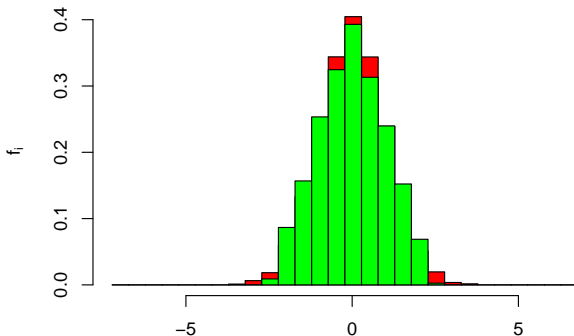
Kurtosis (II)

- The **red** distribution has “average” kurtosis (a Gaussian distribution, that we will study in a few weeks). It is *mesokurtic*.
- The **green** distribution is *leptokurtic*: it has a more acute peak and “fatter” tails.



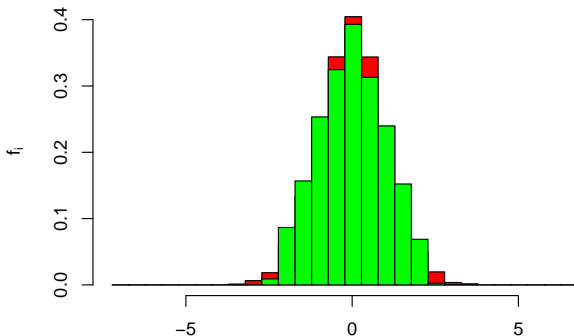
Kurtosis (III)

- Likewise, we might have more observations “centrally” located and less peak and tails than a mesokurtic.



Kurtosis (III)

- Likewise, we might have more observations “centrally” located and less peak and tails than a mesokurtic.
- The green distribution is *platykurtic*: it has less peak and thinner tails.



Kurtosis (III)

- In a nutshell, platykurtic have thin tails and less peak while leptokurtic have higher peak ant fatter tails.

Kurtosis (III)

- In a nutshell, platykurtic have thin tails and less peak while leptokurtic have higher peak and fatter tails.
- The impact of the tails will affect most the high order **even** centered moments,

$$m_{2d} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^{2d}$$

because $(x_i - \bar{x})^{2d}$ can grow very large.

Kurtosis (III)

- In a nutshell, platykurtic have thin tails and less peak while leptokurtic have higher peak and fatter tails.
- The impact of the tails will affect most the high order **even** centered moments,

$$m_{2d} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^{2d}$$

because $(x_i - \bar{x})^{2d}$ can grow very large.

- We can take

$$\frac{m_4}{(S_x)^4}$$

as a suitable scale-free measure of kurtosis.

Kurtosis coefficient

- It turns out that for a very important distribution in practice (the normal or gaussian distribution), $m_4/(S_x)^4$ takes value 3. To have as “origin” zero, we subtract 3.

Kurtosis coefficient

- It turns out that for a very important distribution in practice (the normal or gaussian distribution), $m_4/(S_x)^4$ takes value 3. To have as “origin” zero, we subtract 3.
- Thus, we define the *coefficient of kurtosis* as:

$$g_2 = \frac{m_4}{(S_x)^4} - 3$$

Kurtosis coefficient

- It turns out that for a very important distribution in practice (the normal or gaussian distribution), $m_4/(S_x)^4$ takes value 3. To have as “origin” zero, we subtract 3.
- Thus, we define the *coefficient of kurtosis* as:

$$g_2 = \frac{m_4}{(S_x)^4} - 3$$

- Leptokurtic distributions (= fat tails) have $g_2 > 0$, platykurtic distributions have $g_2 < 0$.

Linear transformations

- Quite often we have a variable X and we would rather use another Y linearly related to X :

$$Y = aX + b$$

Linear transformations

- Quite often we have a variable X and we would rather use another Y linearly related to X :

$$Y = aX + b$$

- For instance, we may have X valued in “pesetas” and want a similar variable Y valued in “euros”. Then, a should be the conversion rate: $a = 1/166,386 = 0,006010121$

Linear transformations

- Quite often we have a variable X and we would rather use another Y linearly related to X :

$$Y = aX + b$$

- For instance, we may have X valued in “pesetas” and want a similar variable Y valued in “euros”. Then, a should be the conversion rate: $a = 1/166,386 = 0,006010121$
- Or, we may have temperatures X in degrees C and want the equivalent Y in Fahrenheit:

$$Y = \left(\frac{9}{5}\right) X + 32$$

Linear transformations

- Quite often we have a variable X and we would rather use another Y linearly related to X :

$$Y = aX + b$$

- For instance, we may have X valued in “pesetas” and want a similar variable Y valued in “euros”. Then, a should be the conversion rate: $a = 1/166,386 = 0,006010121$
- Or, we may have temperatures X in degrees C and want the equivalent Y in Fahrenheit:

$$Y = \left(\frac{9}{5}\right) X + 32$$

- **Question:** What happens to all moments and statistics we have introduced so far?

Linear transformations and location statistics (I)

- A linear transformation is order preserving or order reversing.

Linear transformations and location statistics (I)

- A linear transformation is order preserving or order reversing.
- Clearly, the median of $Y = aX + b$ will be

$$M_e(Y) = aM_e(X) + b.$$

Linear transformations and location statistics (I)

- A linear transformation is order preserving or order reversing.
- Clearly, the median of $Y = aX + b$ will be

$$M_e(Y) = aM_e(X) + b.$$

- With percentiles, we have to be a bit more careful.

Linear transformations and location statistics (I)

- A linear transformation is order preserving or order reversing.
- Clearly, the median of $Y = aX + b$ will be

$$M_e(Y) = aM_e(X) + b.$$

- With percentiles, we have to be a bit more careful.
 - Order preserving: $a > 0 \implies q_\alpha(Y) = aq_\alpha(X) + b.$

Linear transformations and location statistics (I)

- A linear transformation is order preserving or order reversing.
- Clearly, the median of $Y = aX + b$ will be

$$M_e(Y) = aM_e(X) + b.$$

- With percentiles, we have to be a bit more careful.
 - Order preserving: $a > 0 \implies q_\alpha(Y) = aq_\alpha(X) + b.$
 - Order reversing: $a < 0 \implies q_\alpha(Y) = aq_{1-\alpha}(X) + b.$

Linear transformations and location statistics (I)

- A linear transformation is order preserving or order reversing.
- Clearly, the median of $Y = aX + b$ will be

$$M_e(Y) = aM_e(X) + b.$$

- With percentiles, we have to be a bit more careful.
 - Order preserving: $a > 0 \implies q_\alpha(Y) = aq_\alpha(X) + b.$
 - Order reversing: $a < 0 \implies q_\alpha(Y) = aq_{1-\alpha}(X) + b.$
- The case $a = 0$ is trivial!

Linear transformations and location statistics (II)

- Clearly, the mean of $Y = aX + b$ will be $\bar{y} = a\bar{x} + b$.

Linear transformations and location statistics (II)

- Clearly, the mean of $Y = aX + b$ will be $\bar{y} = a\bar{x} + b$.
- Proof:

$$\begin{aligned}\bar{y} &= \frac{1}{N} \sum_{i=1}^N y_i \\ &= \frac{1}{N} \sum_{i=1}^N (ax_i + b) \\ &= a \left(\frac{1}{N} \sum_{i=1}^N x_i \right) + \frac{1}{N} \sum_{i=1}^N b \\ &= a\bar{x} + b\end{aligned}$$

Linear transformations and location statistics (III)

- Likewise, for the mode:

$$\text{Mode}(Y) = a \times \text{Mode}(X) + b.$$

Linear transformations and location statistics (III)

- Likewise, for the mode:

$$\text{Mode}(Y) = a \times \text{Mode}(X) + b.$$

- To summarize:

Linear transformations and location statistics (III)

- Likewise, for the mode:

$$\text{Mode}(Y) = a \times \text{Mode}(X) + b.$$

- To summarize:
 - The location statistics of a linearly transformed variable are the linearly transformed statistics of the original variable.

Linear transformations and location statistics (III)

- Likewise, for the mode:

$$\text{Mode}(Y) = a \times \text{Mode}(X) + b.$$

- To summarize:
 - The location statistics of a linearly transformed variable are the linearly transformed statistics of the original variable.
 - In the case of q_α and an order reversing transformation, we have to replace α by $1 - \alpha$.

Linear transformations and dispersion statistics (I)

- For the h central moment we have:

$$\frac{1}{N} \sum_{i=1}^N y_i^h = \frac{1}{N} \sum_{i=1}^N (ax_i + b)^h$$

Linear transformations and dispersion statistics (I)

- For the h central moment we have:

$$\frac{1}{N} \sum_{i=1}^N y_i^h = \frac{1}{N} \sum_{i=1}^N (ax_i + b)^h$$

- The right hand side can be put in terms of the moments of X using Newton's formula for the expansion of $(w + z)^h$

Linear transformations and dispersion statistics (I)

- For the h central moment we have:

$$\frac{1}{N} \sum_{i=1}^N y_i^h = \frac{1}{N} \sum_{i=1}^N (ax_i + b)^h$$

- The right hand side can be put in terms of the moments of X using Newton's formula for the expansion of $(w + z)^h$
- Low order cases are easy to work with.

Linear transformations and dispersion statistics (II)

- For instance,

$$\begin{aligned} S_y^2 = m_2(Y) &= \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 \\ &= \frac{1}{N} \sum_{i=1}^N (ax_i + b - a\bar{x} - b)^2 \\ &= \frac{1}{N} \sum_{i=1}^N [a(x_i - \bar{x})]^2 \\ &= a^2 m_2(X) = a^2 S_x^2 \end{aligned}$$

Linear transformations and dispersion statistics (II)

- For instance,

$$\begin{aligned} S_y^2 = m_2(Y) &= \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 \\ &= \frac{1}{N} \sum_{i=1}^N (ax_i + b - a\bar{x} - b)^2 \\ &= \frac{1}{N} \sum_{i=1}^N [a(x_i - \bar{x})]^2 \\ &= a^2 m_2(X) = a^2 S_x^2 \end{aligned}$$

- Notice it does not depend on b !

Linear transformations and dispersion statistics (III)

- Other simple cases follow.

Linear transformations and dispersion statistics (III)

- Other simple cases follow.
- $S_y = |a|S_x$.

Linear transformations and dispersion statistics (III)

- Other simple cases follow.
- $S_y = |a|S_x$.
- Mean deviation: $MD(Y) = |a|MD(X)$.

Linear transformations and dispersion statistics (III)

- Other simple cases follow.
- $S_y = |a|S_x$.
- Mean deviation: $MD(Y) = |a|MD(X)$.
- $m_h(Y) = a^h m_h(X)$.

Linear transformations and shape statistics (I)

- Skewness:

$$g_1(Y) = \frac{m_3(Y)}{(S_Y)^3} = \frac{a^3 m_3(X)}{|a|^3 (S_X)^3} = \frac{a}{|a|} \frac{m_3(X)}{(S_X)^3}$$

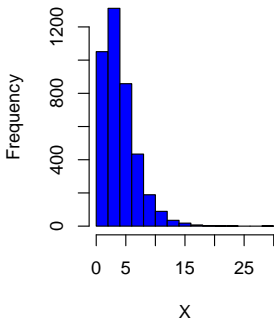
Linear transformations and shape statistics (I)

- Skewness:

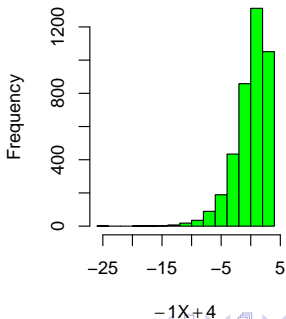
$$g_1(Y) = \frac{m_3(Y)}{(S_Y)^3} = \frac{a^3 m_3(X)}{|a|^3 (S_X)^3} = \frac{a}{|a|} \frac{m_3(X)}{(S_X)^3}$$

- g_1 is either unchanged or changes sign.

Original



Transformed



Linear transformations and shape statistics (II)

- Similarly

$$g_2(Y) = \frac{m_4(Y)}{(S_Y)^4} = \frac{a^4 m_4(X)}{|a|^4 (S_X)^3} = \frac{m_4(X)}{(S_X)^4} = g_2(X)$$

Linear transformations and shape statistics (II)

- Similarly

$$g_2(Y) = \frac{m_4(Y)}{(S_Y)^4} = \frac{a^4 m_4(X)}{|a|^4 (S_X)^3} = \frac{m_4(X)}{(S_X)^4} = g_2(X)$$

- g_2 is invariant under linear transformations.

So...

- Shape is preserved by linear transformations (except for the orientation of skewness).

So...

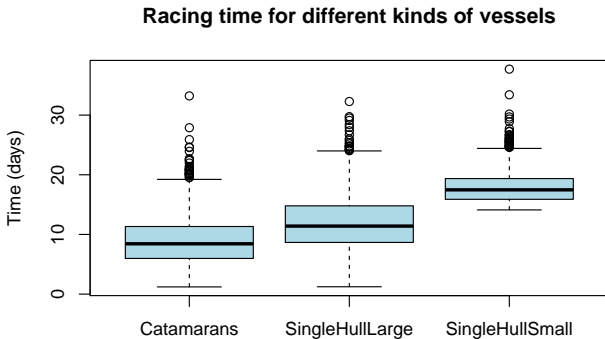
- Shape is preserved by linear transformations (except for the orientation of skewness).
- Moments, quartiles, etc. of X and $Y = aX + b$ are easily related.

So...

- Shape is preserved by linear transformations (except for the orientation of skewness).
- Moments, quartiles, etc. of X and $Y = aX + b$ are easily related.
- For many purposes (not least for comparison purposes) it is handy to reduce distributions to common mean and variance. For instance, $m = 0$ and $S^2 = 1$.

Making fair comparisons (I)

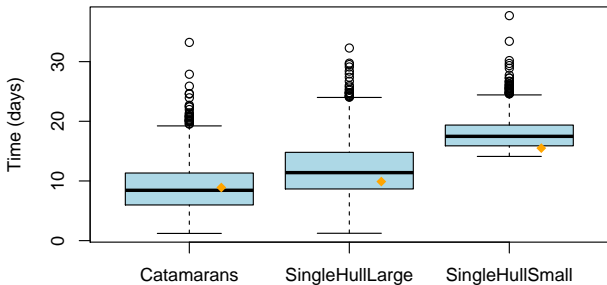
- Assume the distribution of days for a transoceanic race is the following:



Making fair comparisons (II)

- Now suppose the times of the first arriving vessels in each category are 8.9, 9.9 and 16 days (orange dots).

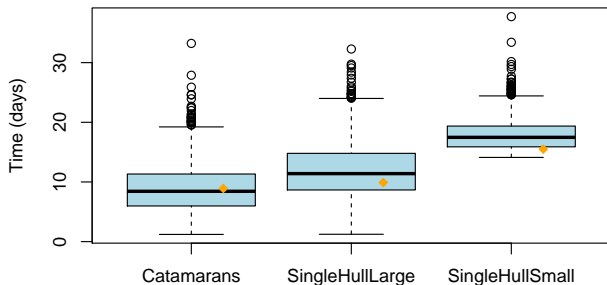
Racing time for different kinds of vessels



Making fair comparisons (II)

- Now suppose the times of the first arriving vessels in each category are 8.9, 9.9 and 16 days (orange dots).

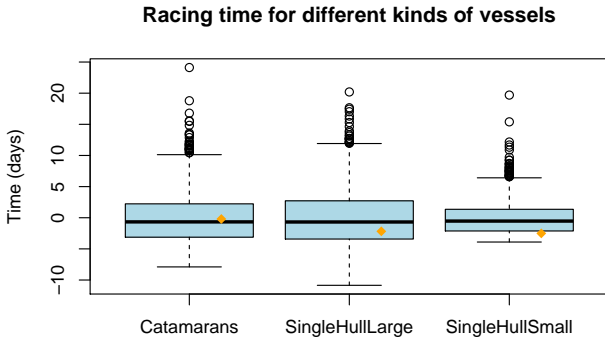
Racing time for different kinds of vessels



- The first catamaran arrived first, but the pilot does not seem particularly able!

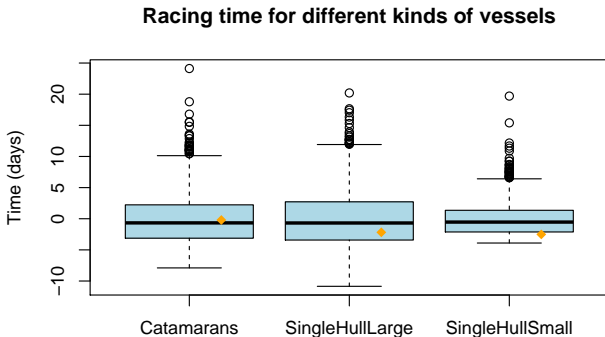
Making fair comparisons (III)

- We can try to align subtracting the mean of each category:



Making fair comparisons (III)

- We can try to align subtracting the mean of each category:

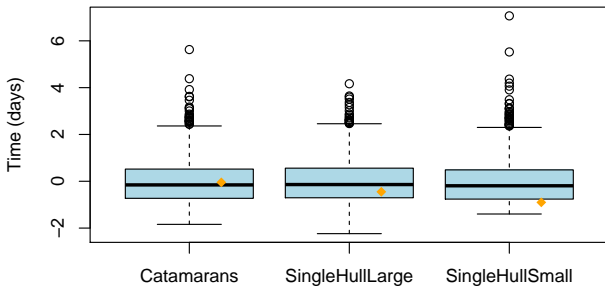


- Not quite what we want yet.

Making fair comparisons (IV)

- We can now divide all times by the standard deviation.

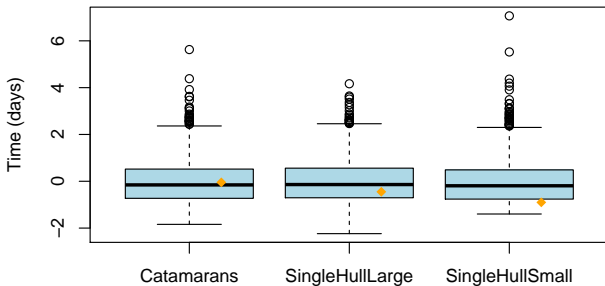
Racing time for different kinds of vessels



Making fair comparisons (IV)

- We can now divide all times by the standard deviation.

Racing time for different kinds of vessels



Making fair comparisons (V)

- Raw times: 8.9, 9.9 and 15.5 days.

Making fair comparisons (V)

- Raw times: 8.9, 9.9 and 15.5 days.
- Category means: 9.09, 12.08 and 18.012.

Making fair comparisons (V)

- Raw times: 8.9, 9.9 and 15.5 days.
- Category means: 9.09, 12.08 and 18.012.
- Category standard deviations: 4.287, 4.847 and 2.787.

Making fair comparisons (V)

- Raw times: 8.9, 9.9 and 15.5 days.
- Category means: 9.09, 12.08 and 18.012.
- Category standard deviations: 4.287, 4.847 and 2.787.
- Standardized times:

$$-0.0443 = \frac{8.9 - 9.09}{4.2867}$$

$$-0.4497 = \frac{9.9 - 12.080}{4.8479}$$

$$-0.9012 = \frac{15.5 - 18.012}{2.7873}$$

Making fair comparisons (V)

- Raw times: 8.9, 9.9 and 15.5 days.
- Category means: 9.09, 12.08 and 18.012.
- Category standard deviations: 4.287, 4.847 and 2.787.
- Standardized times:

$$-0.0443 = \frac{8.9 - 9.09}{4.2867}$$

$$-0.4497 = \frac{9.9 - 12.080}{4.8479}$$

$$-0.9012 = \frac{15.5 - 18.012}{2.7873}$$

- Now, the small single hull first vessel looks best!

Variance and concentration

- One might think that the variance is a good indicator of concentration, and to a certain extent it is.

Variance and concentration

- One might think that the variance is a good indicator of concentration, and to a certain extent it is.
- However

Variance and concentration

- One might think that the variance is a good indicator of concentration, and to a certain extent it is.
- However
 1. It is not scale independent.

Variance and concentration

- One might think that the variance is a good indicator of concentration, and to a certain extent it is.
- However
 1. It is not scale independent.
 2. It hides much information that a histogram, for instance, would give.

Variance and concentration

- One might think that the variance is a good indicator of concentration, and to a certain extent it is.
- However
 1. It is not scale independent.
 2. It hides much information that a histogram, for instance, would give.
- The Lorenz curve, introduced next, offers a suitable alternative.

The Lorenz curve (I)

- Consider observations on income for N subjects:
 x_1, x_2, \dots, x_N .

The Lorenz curve (I)

- Consider observations on income for N subjects:
 x_1, x_2, \dots, x_N .
- Order these observations to get: $x_{(1)}, x_{(2)}, \dots, x_{(N)}$.

The Lorenz curve (I)

- Consider observations on income for N subjects:
 x_1, x_2, \dots, x_N .
- Order these observations to get: $x_{(1)}, x_{(2)}, \dots, x_{(N)}$.
- **Aside question:** If $N = 101$, what would $x_{(51)}$ be? What about $x_{(76)}$?

The Lorenz curve (I)

- Consider observations on income for N subjects:
 x_1, x_2, \dots, x_N .
- Order these observations to get: $x_{(1)}, x_{(2)}, \dots, x_{(N)}$.
- **Aside question:** If $N = 101$, what would $x_{(51)}$ be? What about $x_{(76)}$?
- Now compute

$$M = x_{(1)} + x_{(2)} + \dots + x_{(N)}$$

(the cumulated income).

The Lorenz curve (I)

- Consider observations on income for N subjects:
 x_1, x_2, \dots, x_N .
- Order these observations to get: $x_{(1)}, x_{(2)}, \dots, x_{(N)}$.
- **Aside question:** If $N = 101$, what would $x_{(51)}$ be? What about $x_{(76)}$?
- Now compute

$$M = x_{(1)} + x_{(2)} + \dots + x_{(N)}$$

(the cumulated income).

- or, if data are grouped,

$$M = n_1 x_{(1)} + n_2 x_{(2)} + \dots + n_k x_{(k)}$$

The Lorenz curve (II)

- Now compute the fraction of income each value of $x_{(i)}$ represents:

$$q_i = \frac{n_i x_{(i)}}{M}$$

The Lorenz curve (II)

- Now compute the fraction of income each value of $x_{(i)}$ represents:

$$q_i = \frac{n_i x_{(i)}}{M}$$

- Finally, compute

$$Q_i = \sum_{j=1}^i q_j$$

The Lorenz curve (II)

- Now compute the fraction of income each value of $x_{(i)}$ represents:

$$q_i = \frac{n_i x_{(i)}}{M}$$

- Finally, compute

$$Q_i = \sum_{j=1}^i q_j$$

- If data are grouped, compute also $f_i = n_i/N$ (if data are not grouped, these are just $1/N$).

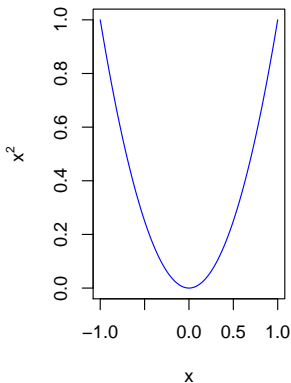
The Lorenz curve (II)

- Now, we will plot a curve through the pairs (F_i, Q_i) for $i = 1, 2, \dots, k$.

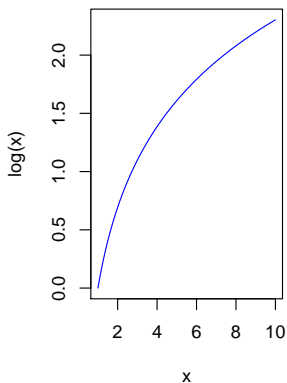
Convex and concave functions

- Convex functions have non decreasing slope, concave functions non increasing slope.

Convex

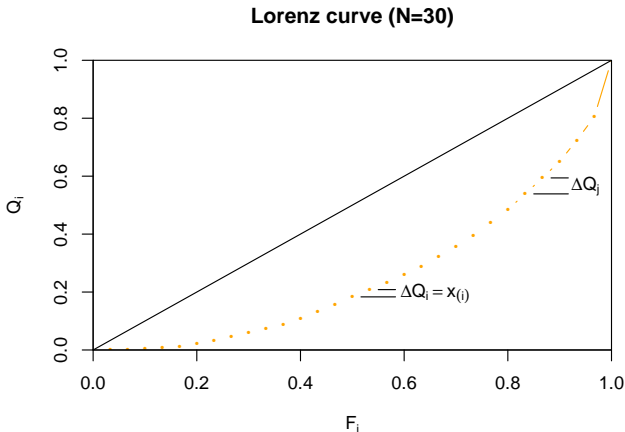


Concave



Lorenz curves are convex

- Horizontal steps are equal, vertical steps grow in size, so slope never decreases.



Invariance under scale changes

- Suppose that income is measured in units k times smaller than the original ones. Then,

Invariance under scale changes

- Suppose that income is measured in units k times smaller than the original ones. Then,
 1. All x_i are multiplied by k .

Invariance under scale changes

- Suppose that income is measured in units k times smaller than the original ones. Then,
 1. All x_i are multiplied by k .
 2. All $x_{(i)}$ are multiplied by k .

Invariance under scale changes

- Suppose that income is measured in units k times smaller than the original ones. Then,
 1. All x_i are multiplied by k .
 2. All $x_{(i)}$ are multiplied by k .
 3. All $q_i = x_{(i)}/M$ are invariant!.

Invariance under scale changes

- Suppose that income is measured in units k times smaller than the original ones. Then,
 1. All x_i are multiplied by k .
 2. All $x_{(i)}$ are multiplied by k .
 3. All $q_i = x_{(i)}/M$ are invariant!.
- The Lorenz curve is unaffected by changes in the scale of measure.

The Gini index (I)

- For a perfectly egalitarian distribution of income, the Lorenz curve would overlap the slope 1 line.

The Gini index (I)

- For a perfectly egalitarian distribution of income, the Lorenz curve would overlap the slope 1 line.
- It is apparent that the area between that line and the Lorenz curve is a sensible measure of inequality in the distribution.

The Gini index (I)

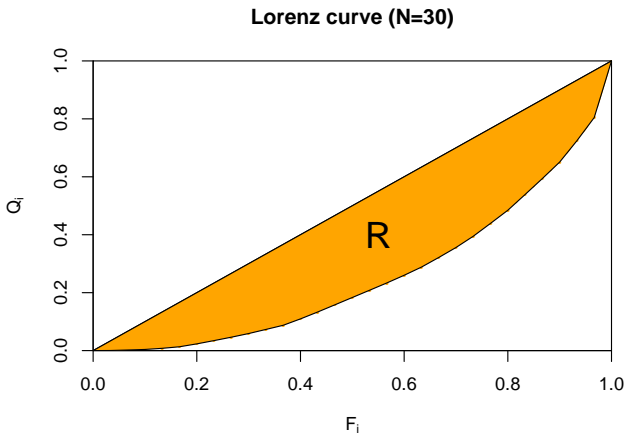
- For a perfectly egalitarian distribution of income, the Lorenz curve would overlap the slope 1 line.
- It is apparent that the area between that line and the Lorenz curve is a sensible measure of inequality in the distribution.
- An elaboration of that idea is the **Gini index**. It is defined as

$$I_G = \frac{R}{\Delta}$$

where R is the area mentioned and $\Delta = \frac{1}{2}$ the total area under the equal distribution line.

The Gini index (II)

- Graphically,



The Gini index (III)

- There is no need to measure the orange surface, the Gini index can be readily computed by:

$$I_G = 2 \sum_{i=1}^k q_i F_i^* - 1$$

where $F_i^* = F_i - f_i/2$.

The Gini index (III)

- There is no need to measure the orange surface, the Gini index can be readily computed by:

$$I_G = 2 \sum_{i=1}^k q_i F_i^* - 1$$

where $F_i^* = F_i - f_i/2$.

- Can be easily proved.

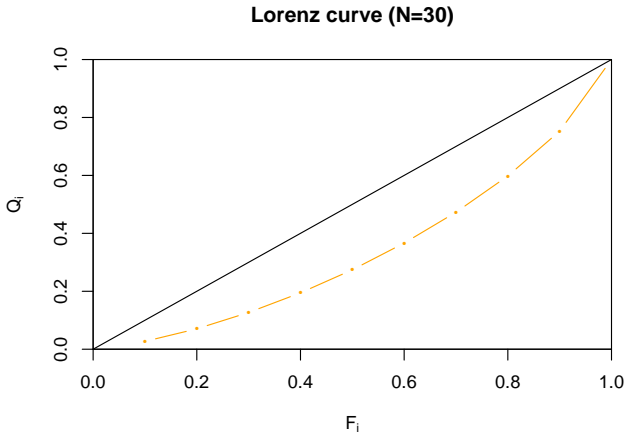
The Gini index: an example (I)

- This is the average income for each of the ten deciles in Spain, 1990-1991 (Encuesta de Presupuestos Familiares, INE).

Decile	Income	Decile	Income
First	601431	Sixth	2049420
Second	962087	Seventh	2365199
Third	1251174	Eight	2777410
Fourth	1506535	Ninth	3436151
Fifth	1764340	Tenth	5489341

The Gini index: an example (II)

- The Lorenz curve is:



The Gini index: an example (III)

Decile	Income	F_i	q_i	$q_i(F_i - f_i/2)$
First	601431	0.10	0.0271	0.00135
Second	962087	0.20	0.0433	0.00659
Third	1251174	0.30	0.0563	0.01409
Fourth	1506535	0.40	0.0678	0.02375
Fifth	1764340	0.50	0.0794	0.03576
Sixth	2049420	0.60	0.0923	0.05077
Seventh	2365199	0.70	0.1065	0.06924
Eight	2777410	0.80	0.1250	0.09381
Ninth	3436151	0.90	0.1547	0.13154
Tenth	5489341	1.00	0.2472	0.23487
			Sum	0.66169

Thus, $I_G = 2 \times 0.66169 - 1 = 0.3234$.

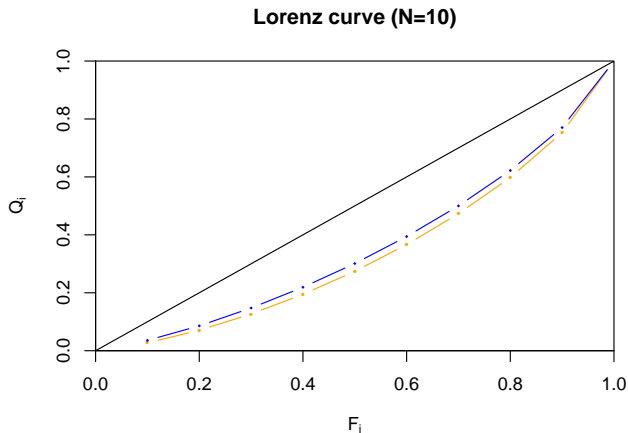
Adding a constant amount to each group (I)

- The effect is always to reduce inequality.

Adding a constant amount to each group (I)

- The effect is always to reduce inequality.
- The reason is simple: you increase proportionally more the smaller income brackets.

Adding a constant amount to each group (II)



Two dimensional frequency distributions

- It happens oftentimes that we record **two** values for each member of the population.

Two dimensional frequency distributions

- It happens oftentimes that we record **two** values for each member of the population.
- For instance, for the students in this class we could measure $X =$ height in cm and $Y =$ weight in Kg.

Two dimensional frequency distributions

- It happens oftentimes that we record **two** values for each member of the population.
- For instance, for the students in this class we could measure $X =$ height in cm and $Y =$ weight in Kg.
- Then, for each student we have a pair (X, Y) .

Two dimensional frequency distributions

- It happens oftentimes that we record **two** values for each member of the population.
- For instance, for the students in this class we could measure $X =$ height in cm and $Y =$ weight in Kg.
- Then, for each student we have a pair (X, Y) .
- Everything we said about cualitative, discrete, ordinal, nominal etc. variables remains applicable.

Contingency tables (I)

- We can set up a table in which we count how many cases have each combination of values of X and Y .

Contingency tables (I)

- We can set up a table in which we count how many cases have each combination of values of X and Y .
- For instance,

Eye color (X)	Hair color (Y)			
	Black	Brown	Blonde	Red
Black	n_{11}	n_{12}	n_{13}	n_{14}
Brown	n_{21}	n_{22}	n_{23}	n_{24}
Green	n_{31}	n_{32}	n_{33}	n_{34}
Blue	n_{41}	n_{42}	n_{43}	n_{44}
Other	n_{51}	n_{52}	n_{53}	n_{54}

Contingency tables (II)

- If the variable(s) are continuous, we can discretize by taking intervals.

Contingency tables (II)

- If the variable(s) are continuous, we can discretize by taking intervals.
- For instance,

Height (cm) (X)	Weight (Kg) (Y)			
	[35,50)	[50,65)	[65,80)	≥ 80
< 135	n_{11}	n_{12}	n_{13}	n_{14}
[135, 145)	n_{21}	n_{22}	n_{23}	n_{24}
[145, 165)	n_{31}	n_{32}	n_{33}	n_{34}
[165, 185)	n_{41}	n_{42}	n_{43}	n_{44}
≥ 185	n_{51}	n_{52}	n_{53}	n_{54}

Contingency tables (III)

- The (integer) numbers n_{ij} are called **absolute frequencies**.

Contingency tables (III)

- The (integer) numbers n_{ij} are called **absolute frequencies**.
- The total number of cases in the population is

$$\sum_{i=1}^I \sum_{j=1}^J n_{ij} = N$$

Contingency tables (III)

- The (integer) numbers n_{ij} are called **absolute frequencies**.
- The total number of cases in the population is

$$\sum_{i=1}^I \sum_{j=1}^J n_{ij} = N$$

- The (fractional) numbers $f_{ij} = \frac{n_{ij}}{N}$ are called **relative frequencies**.

Contingency tables (III)

- The (integer) numbers n_{ij} are called **absolute frequencies**.
- The total number of cases in the population is

$$\sum_{i=1}^I \sum_{j=1}^J n_{ij} = N$$

- The (fractional) numbers $f_{ij} = \frac{n_{ij}}{N}$ are called **relative frequencies**.
- Clearly,

$$\sum_{i=1}^I \sum_{j=1}^J f_{ij} = 1$$

Contingency tables (IV)

- A contingency table can be equivalently set up a table in terms of absolute or (like in the following) relative frequencies.

Eye color (X)	Hair color (Y)				Row marginal
	Black	Brown	Blonde	Red	
Black	f_{11}	f_{12}	f_{13}	f_{14}	$f_{1.}$
Brown	f_{21}	f_{22}	f_{23}	f_{24}	$f_{2.}$
Green	f_{31}	f_{32}	f_{33}	f_{34}	$f_{3.}$
Blue	f_{41}	f_{42}	f_{43}	f_{44}	$f_{4.}$
Other	f_{51}	f_{52}	f_{53}	f_{54}	$f_{5.}$
Column marginal	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	1

Joint, conditional, marginal... (I)

- The marginals are defined

$$f_{i.} = \sum_{j=1}^J f_{ij} \quad \forall i$$

$$f_{.j} = \sum_{i=1}^I f_{ij} \quad \forall j$$

Joint, conditional, marginal... (I)

- The marginals are defined

$$f_{i.} = \sum_{j=1}^J f_{ij} \quad \forall i$$

$$f_{.j} = \sum_{i=1}^I f_{ij} \quad \forall j$$

- Clearly,

$$\sum_{i=1}^I f_{i.} = 1 \quad \sum_{j=1}^J f_{.j} = 1$$

Joint, conditional, marginal... (II)

- The marginal distribution is the distribution of one of the characters, *irrespective of the values taken by the other*.

Joint, conditional, marginal... (II)

- The marginal distribution is the distribution of one of the characters, *irrespective of the values taken by the other*.
- Clearly,

$$\sum_{i=1}^I f_{i.} = 1 \quad \sum_{j=1}^J f_{.j} = 1$$

Joint, conditional, marginal... (II)

- The marginal distribution is the distribution of one of the characters, *irrespective of the values taken by the other*.
- Clearly,

$$\sum_{i=1}^I f_{i.} = 1 \quad \sum_{j=1}^J f_{.j} = 1$$

- Similar definitions apply to absolute frequencies, e.g.
 $n_{i.} = \sum_{j=1}^J n_{ij}$.

Joint, conditional, marginal... (III)

- Sometimes, we are interested in the distribution of one of the characters restricting the other to take a certain value.

Joint, conditional, marginal... (III)

- Sometimes, we are interested in the distribution of one of the characters restricting the other to take a certain value.
- For instance, we might be interested in the distribution of eye color but only among those whose hair is black.

Joint, conditional, marginal... (III)

- Sometimes, we are interested in the distribution of one of the characters restricting the other to take a certain value.
- For instance, we might be interested in the distribution of eye color but only among those whose hair is black.
- This is the so-called **conditional distribution** $X|Y = \text{"black"}$.

Joint, conditional, marginal... (IV)

- If we go back to our table, it is clear that *among people with black hair*, the relative frequencies of each eye color are:

$$n_{11}/n_{.1}, n_{21}/n_{.1}, n_{31}/n_{.1}, n_{41}/n_{.1}, n_{51}/n_{.1}$$

Eye color (X)	Hair color (Y)			
	Black	Brown	Blonde	Red
Black	n_{11}	n_{12}	n_{13}	n_{14}
Brown	n_{21}	n_{22}	n_{23}	n_{24}
Green	n_{31}	n_{32}	n_{33}	n_{34}
Blue	n_{41}	n_{42}	n_{43}	n_{44}
Other	n_{51}	n_{52}	n_{53}	n_{54}
	$n_{.1}$			

Joint, conditional, marginal... (V)

- Equivalently, it is clear that *among people with black hair*, the relative frequencies of each eye color are:

$$f_{11}/f_{.1}, f_{21}/f_{.1}, f_{31}/f_{.1}, f_{41}/f_{.1}, f_{51}/f_{.1}$$

Eye color (X)	Hair color (Y)			
	Black	Brown	Blonde	Red
Black	f_{11}	f_{12}	f_{13}	f_{14}
Brown	f_{21}	f_{22}	f_{23}	f_{24}
Green	f_{31}	f_{32}	f_{33}	f_{34}
Blue	f_{41}	f_{42}	f_{43}	f_{44}
Other	f_{51}	f_{52}	f_{53}	f_{54}
	$f_{.1}$			

Independence (I)

- We now come to a crucial concept in Statistics.

Independence (I)

- We now come to a crucial concept in Statistics.
- Very loosely stated now: it will be defined formally in the second half of the course.

Independence (I)

- We now come to a crucial concept in Statistics.
- Very loosely stated now: it will be defined formally in the second half of the course.
- However, it is important that we gain intuition from now.

Independence (I)

- We now come to a crucial concept in Statistics.
- Very loosely stated now: it will be defined formally in the second half of the course.
- However, it is important that we gain intuition from now.
- Consider

Treatment (X)	Outcome (Y)		Row marginal
	Gets sick	Stays healthy	
Vaccine	f_{11}	f_{12}	$f_{1.}$
Placebo	f_{21}	f_{22}	$f_{2.}$
Column marginal	$f_{.1}$	$f_{.2}$	1

Independence (II)



Treatment (X)	Outcome (Y)		Row marginal
	Gets sick	Stays healthy	
Vaccine	f_{11}	f_{12}	$f_{1.}$
Placebo	f_{21}	f_{22}	$f_{2.}$
Column marginal	$f_{.1}$	$f_{.2}$	1

Independence (II)

- The conditional distribution of sick/healthy among people treated with the vaccine is $f_{11}/f_{1.}$, $f_{12}/f_{1.}$.

-

Outcome (Y)			
Treatment (X)	Gets sick	Stays healthy	Row marginal
Vaccine	f_{11}	f_{12}	$f_{1.}$
Placebo	f_{21}	f_{22}	$f_{2.}$
Column marginal	$f_{.1}$	$f_{.2}$	1

Independence (II)

- The conditional distribution of sick/healthy among people treated with the vaccine is $f_{11}/f_{1.}$, $f_{12}/f_{1.}$.
- The conditional distribution of sick/healthy among people non treated is $f_{21}/f_{2.}$, $f_{22}/f_{2.}$.

•

Treatment (X)	Outcome (Y)		Row marginal
	Gets sick	Stays healthy	
Vaccine	f_{11}	f_{12}	$f_{1.}$
Placebo	f_{21}	f_{22}	$f_{2.}$
Column marginal	$f_{.1}$	$f_{.2}$	1

Independence (II)

- The conditional distribution of sick/healthy among people treated with the vaccine is $f_{11}/f_{1.}$, $f_{12}/f_{1.}$.
- The conditional distribution of sick/healthy among people non treated is $f_{21}/f_{2.}$, $f_{22}/f_{2.}$.
- If both distributions are the same, there would be strong indication that the vaccine has no effect or Treatment and Outcome are **independent**.

Treatment (X)	Outcome (Y)		Row marginal
	Gets sick	Stays healthy	
Vaccine	f_{11}	f_{12}	$f_{1.}$
Placebo	f_{21}	f_{22}	$f_{2.}$
Column marginal	$f_{.1}$	$f_{.2}$	1

Independence (II)

- The conditional distribution of sick/healthy among people treated with the vaccine is $f_{11}/f_{1.}$, $f_{12}/f_{1.}$.
- The conditional distribution of sick/healthy among people non treated is $f_{21}/f_{2.}$, $f_{22}/f_{2.}$.
- If both distributions are the same, there would be strong indication that the vaccine has no effect or Treatment and Outcome are **independent**.

Treatment (X)	Outcome (Y)		Row marginal
	Gets sick	Stays healthy	
Vaccine	f_{11}	f_{12}	$f_{1.}$
Placebo	f_{21}	f_{22}	$f_{2.}$
Column marginal	$f_{.1}$	$f_{.2}$	1

Independence (III)



Treatment (X)	Outcome (Y)		Row marginal
	Gets sick	Stays healthy	
Vaccine	f_{11}	f_{12}	$f_{1.}$
Placebo	f_{21}	f_{22}	$f_{2.}$
Column marginal	$f_{.1}$	$f_{.2}$	1

Independence (III)

- Notice: from $f_{11}/f_{1.} = f_{21}/f_{2.}$, we get:

$$\begin{aligned}
 f_{.1} = f_{11} + f_{21} &= f_{11} + f_{2.} f_{11}/f_{1.} \\
 &= f_{11} (1 + f_{2.}/f_{1.}) \\
 &= f_{11} \frac{f_{1.} + f_{2.}}{f_{1.}} \\
 &= f_{11}/f_{1.} = f_{21}/f_{2.}
 \end{aligned}$$

-

Outcome (Y)			
Treatment (X)	Gets sick	Stays healthy	Row marginal
Vaccine	f_{11}	f_{12}	$f_{1.}$
Placebo	f_{21}	f_{22}	$f_{2.}$
Column marginal	$f_{.1}$	$f_{.2}$	1

Independence (IV)

- The previous argument is completely general: independence implies equality of all conditionals among them and hence to the marginals.

•

Treatment (X)	Outcome (Y)		Row marginal
	Gets sick	Stays healthy	
Vaccine	f_{11}	f_{12}	$f_{1.}$
Placebo	f_{21}	f_{22}	$f_{2.}$
Column marginal	$f_{.1}$	$f_{.2}$	1

Independence (IV)

- The previous argument is completely general: independence implies equality of all conditionals among them and hence to the marginals.
- As a consequence, $f_{11}/f_{1.} = f_{.1} \implies f_{11} = f_{1.} \times f_{.1}$

•

Treatment (X)	Outcome (Y)		Row marginal
	Gets sick	Stays healthy	
Vaccine	f_{11}	f_{12}	$f_{1.}$
Placebo	f_{21}	f_{22}	$f_{2.}$
Column marginal	$f_{.1}$	$f_{.2}$	1

Independence (IV)

- The previous argument is completely general: independence implies equality of all conditionals among them and hence to the marginals.
- As a consequence, $f_{11}/f_{1.} = f_{.1} \implies f_{11} = f_{1.} \times f_{.1}$
- In general, independence means $f_{ij} = f_{i.} \times f_{.j}$ for all i, j .

•

Treatment (X)	Outcome (Y)		Row marginal
	Gets sick	Stays healthy	
Vaccine	f_{11}	f_{12}	$f_{1.}$
Placebo	f_{21}	f_{22}	$f_{2.}$
Column marginal	$f_{.1}$	$f_{.2}$	1

Independence (IV)

- The previous argument is completely general: independence implies equality of all conditionals among them and hence to the marginals.
- As a consequence, $f_{11}/f_{1.} = f_{.1} \implies f_{11} = f_{1.} \times f_{.1}$
- In general, independence means $f_{ij} = f_{i.} \times f_{.j}$ for all i, j .
- In practice, the two members of the equality will rarely be identical; but when they are close enough, we have strong indication of independence.

	Outcome (Y)		
Treatment (X)	Gets sick	Stays healthy	Row marginal
Vaccine	f_{11}	f_{12}	$f_{1.}$
Placebo	f_{21}	f_{22}	$f_{2.}$
Column marginal	$f_{.1}$	$f_{.2}$	1

Graphing two dimensional observations

- Which graph you choose depends on what is interesting to you.

Graphing two dimensional observations

- Which graph you choose depends on what is interesting to you.
- You might want to look at the *relationship between X and Y values*.

Graphing two dimensional observations

- Which graph you choose depends on what is interesting to you.
- You might want to look at the *relationship between X and Y values*.
- When X and Y are discrete (or qualitative) you might be more interested in the *frequency of different combinations of X and Y values*.

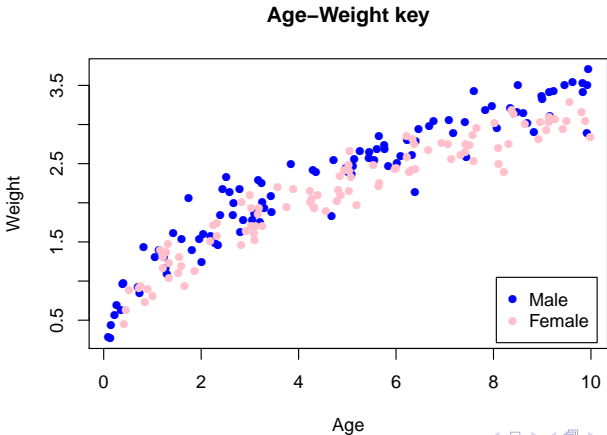
Scatter plot

- If interest centers on the X - Y relationship, a simple scatter plot might be the best.



Scatter plot

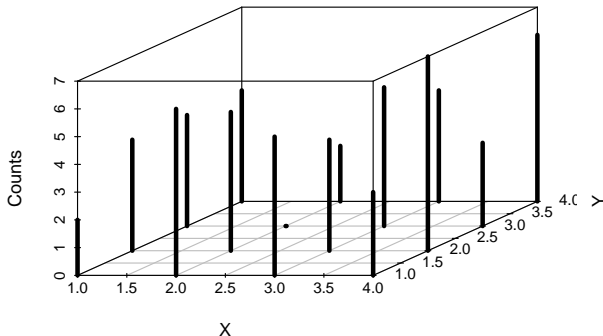
- If interest centers on the X - Y but you want also differentiate between groups, you may use color or different plotting characters:



Scatter plot

- If interest centers on the relative frequency of different combinations of categories, one can draw a two-dimensional histogram:

Two dimensional histogram



Moments

- Now that we have (X, Y) ,

Moments

- Now that we have (X, Y) ,
 - We continue to have all moments of X , central or not:

$$a_h(X) = \frac{1}{N} \sum_{i=1}^N X_i^h \quad m_h(X) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})^h$$

Moments

- Now that we have (X, Y) ,
 - We continue to have all moments of X , central or not:

$$a_h(X) = \frac{1}{N} \sum_{i=1}^N X_i^h \quad m_h(X) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})^h$$

- Similarly for Y

Moments

- Now that we have (X, Y) ,
 - We continue to have all moments of X , central or not:

$$a_h(X) = \frac{1}{N} \sum_{i=1}^N X_i^h \quad m_h(X) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})^h$$

- Similarly for Y
- ...and new **cross-moments** defined as:

$$a_{h_1, h_2} = \frac{1}{N} \sum_{i=1}^N X_i^{h_1} Y_i^{h_2}$$

$$m_{h_1, h_2} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})^{h_1} (Y_i - \bar{y})^{h_2}$$

Moments

- Now that we have (X, Y) ,
 - We continue to have all moments of X , central or not:

$$a_h(X) = \frac{1}{N} \sum_{i=1}^N X_i^h \quad m_h(X) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})^h$$

- Similarly for Y
- ...and new **cross-moments** defined as:

$$a_{h_1, h_2} = \frac{1}{N} \sum_{i=1}^N X_i^{h_1} Y_i^{h_2}$$

$$m_{h_1, h_2} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})^{h_1} (Y_i - \bar{y})^{h_2}$$

- Of particular interest is the case $h_1 = 1, h_2 = 1$.

The covariance

- When $h_1 = 1$, $h_2 = 1$, we have the **covariance** between X and Y .

$$m_{1,1} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})^{h_1} (Y_i - \bar{y})^{h_2}$$

The covariance

- When $h_1 = 1$, $h_2 = 1$, we have the **covariance** between X and Y .

$$m_{1,1} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})^{h_1} (Y_i - \bar{y})^{h_2}$$

- It has a nice interpretation.

The covariance

- When $h_1 = 1$, $h_2 = 1$, we have the **covariance** between X and Y .

$$m_{1,1} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})^{h_1} (Y_i - \bar{y})^{h_2}$$

- It has a nice interpretation.
 1. If $m_{1,1} > 0$, values of X above the mean tend to be associated to values of Y above their mean, and same for values below their means. There is positive covariation.

The covariance

- When $h_1 = 1$, $h_2 = 1$, we have the **covariance** between X and Y .

$$m_{1,1} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})^{h_1} (Y_i - \bar{y})^{h_2}$$

- It has a nice interpretation.
 1. If $m_{1,1} > 0$, values of X above the mean tend to be associated to values of Y above their mean, and same for values below their means. There is positive covariation.
 2. If $m_{1,1} < 0$, values of X above the mean tend to be associated to values of Y below their mean, and viceversa. X and Y evolve in opposite directions.

The covariance

- When $h_1 = 1$, $h_2 = 1$, we have the **covariance** between X and Y .

$$m_{1,1} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})^{h_1} (Y_i - \bar{y})^{h_2}$$

- It has a nice interpretation.
 1. If $m_{1,1} > 0$, values of X above the mean tend to be associated to values of Y above their mean, and same for values below their means. There is positive covariation.
 2. If $m_{1,1} < 0$, values of X above the mean tend to be associated to values of Y below their mean, and viceversa. X and Y evolve in opposite directions.
 3. If $m_{1,1} = 0$, values of X and Y appear unrelated.

The correlation coefficient

- Notice: $m_{1,1}$ depends on the scales of both X and Y :

$$S_{xy} = m_{1,1} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})(Y_i - \bar{y})$$

The correlation coefficient

- Notice: $m_{1,1}$ depends on the scales of both X and Y :

$$S_{xy} = m_{1,1} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})(Y_i - \bar{y})$$

- We can obtain a measure of association which is scale-independent dividing by the standard deviations.

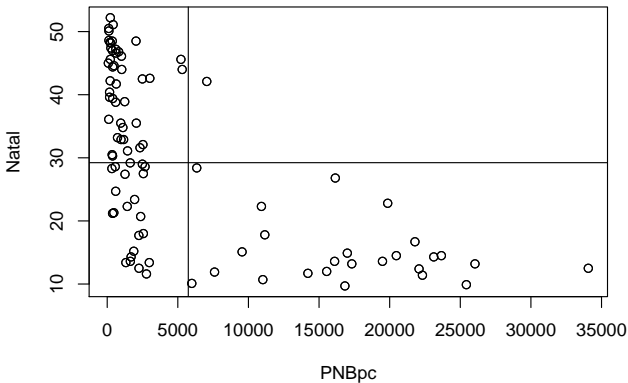
$$r_{xy} = \frac{S_{xy}}{S_x S_y};$$

this measure is the **correlation coefficient**.

An example

- Looking at the following data, what r_{xy} can we expect?

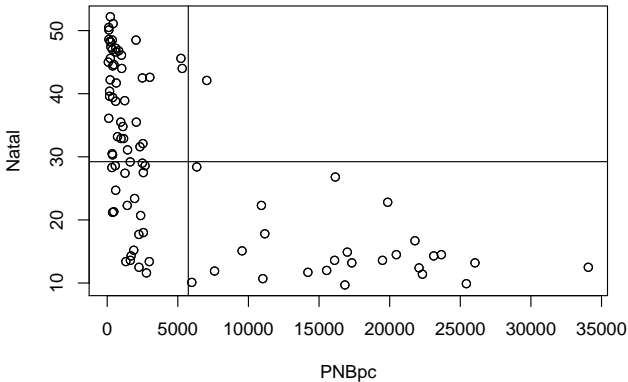
Fertilidad vs. Renta



An example

- Looking at the following data, what r_{xy} can we expect?

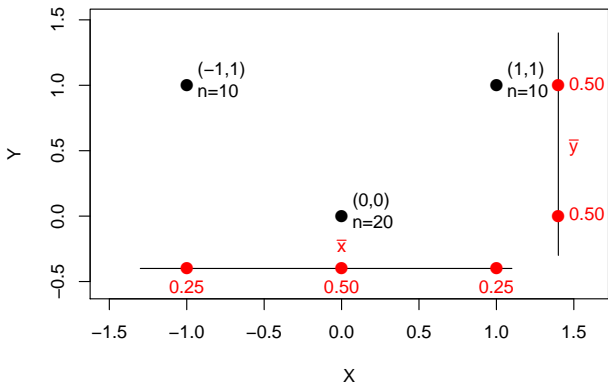
Fertilidad vs. Renta



- Indeed, if we compute r_{xy} we get a value of -0.629 .

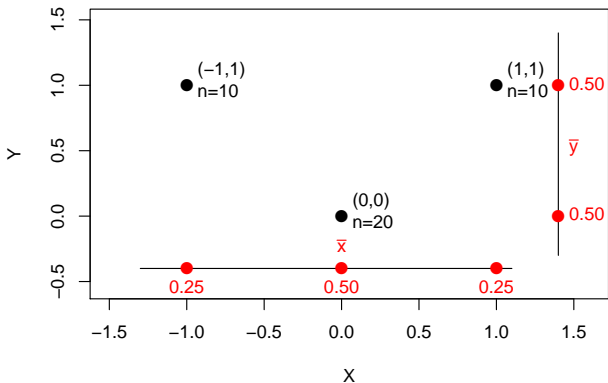
Incorrelation does not imply independence

- Consider the following frequency distribution in R^2 :



Incorrelation does not imply independence

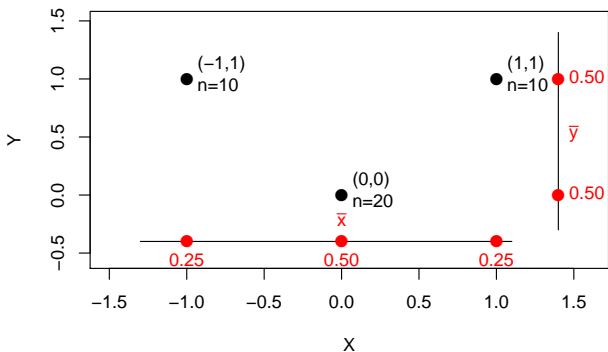
- Consider the following frequency distribution in R^2 :



- Clearly, $1/N \sum_i (X_i - \bar{x})(Y_i - \bar{y}) = 0$.

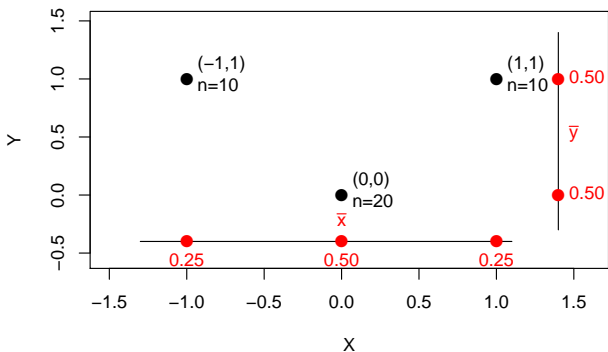
Incorrelation does not imply independence

- $0 = f_{0,1} \neq f_{0.} \times f_{.1} = 0.50 \times 0.50$



Incorrelation does not imply independence

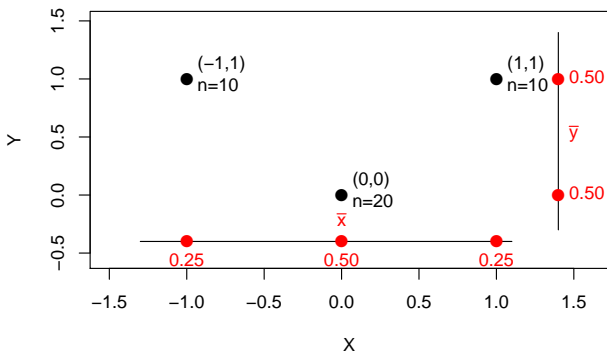
- $0 = f_{0,1} \neq f_{0.} \times f_{.1} = 0.50 \times 0.50$



- However, there is clearly not independence.

Incorrelation does not imply independence

- $0 = f_{0,1} \neq f_{0.} \times f_{.1} = 0.50 \times 0.50$



- However, there is clearly not independence.
- So there may be incorrelation and still the variables not be

Independence does imply incorrelation

- The converse is true, however.

$$\begin{aligned}S_{xy} &= \frac{1}{N} \sum_{i=1}^N (X_i - \bar{x})(Y_i - \bar{y}) \\&= \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^J n_{ij} (X_i - \bar{x})(Y_j - \bar{y}) \\&= \sum_{i=1}^I \sum_{j=1}^J f_{ij} (X_i - \bar{x})(Y_j - \bar{y}) \\&= \sum_{i=1}^I \sum_{j=1}^J f_{i.} f_{.j} (X_i - \bar{x})(Y_j - \bar{y}) \\&= \sum_{i=1}^I f_{i.} (X_i - \bar{x}) \times \sum_{j=1}^J f_{.j} (Y_j - \bar{y})\end{aligned}$$

Independence does imply incorrelation

- However, both of the sums in the last term are zero:

$$\begin{aligned}\sum_{i=1}^I f_{i.}(X_i - \bar{x}) &= \sum_{i=1}^I f_{i.}X_i - \sum_{i=1}^I f_{i.}\bar{x} \\ &= \bar{x} - \bar{x} \\ &= 0\end{aligned}$$

Independence does imply incorrelation

- However, both of the sums in the last term are zero:

$$\begin{aligned}\sum_{i=1}^I f_{i.}(X_i - \bar{x}) &= \sum_{i=1}^I f_{i.}X_i - \sum_{i=1}^I f_{i.}\bar{x} \\ &= \bar{x} - \bar{x} \\ &= 0\end{aligned}$$

- Same for $\sum_{j=1}^J f_{.j}(Y_j - \bar{y})$.

Independence does imply incorrelation

- However, both of the sums in the last term are zero:

$$\begin{aligned}\sum_{i=1}^I f_{i.}(X_i - \bar{x}) &= \sum_{i=1}^I f_{i.}X_i - \sum_{i=1}^I f_{i.}\bar{x} \\ &= \bar{x} - \bar{x} \\ &= 0\end{aligned}$$

- Same for $\sum_{j=1}^J f_{.j}(Y_j - \bar{y})$.
- Therefore,

$$r_{xy} = \sum_{i=1}^I f_{i.}(X_i - \bar{x}) \times \sum_{j=1}^J f_{.j}(Y_j - \bar{y}) = 0$$

Independence and incorrelation

- The example shown suggests that whenever the X - Y relationship is far from monotone, it will be hard to pick up for r_{xy} .

Independence and incorrelation

- The example shown suggests that whenever the X - Y relationship is far from monotone, it will be hard to pick up for r_{xy} .
- On the other hand, nearly linear relationships between X and Y will always give r_{xy} very large in absolute value (close to 1 or -1).

Independence and incorrelation

- The example shown suggests that whenever the X - Y relationship is far from monotone, it will be hard to pick up for r_{xy} .
- On the other hand, nearly linear relationships between X and Y will always give r_{xy} very large in absolute value (close to 1 or -1).
- Thus, r_{xy} is said to measure *linear association* between X and Y .

Effect of linear transformations (I)

- We have seen previously

$$S_{aX+b}^2 = a^2 S_X^2$$

$$S_{aX+b} = |a| S_X$$

Effect of linear transformations (I)

- We have seen previously

$$S_{aX+b}^2 = a^2 S_X^2$$

$$S_{aX+b} = |a| S_X$$

- If we take linear transformations

$$U = aX + b$$

$$V = cY + d,$$

How are $\text{Cov}(X, Y)$ and $\text{Cov}(U, V)$ related? What about r_{xy} and r_{uv} ?

Effect of linear transformations (I)

- We have seen previously

$$S_{aX+b}^2 = a^2 S_X^2$$

$$S_{aX+b} = |a| S_X$$

- If we take linear transformations

$$U = aX + b$$

$$V = cY + d,$$

How are $\text{Cov}(X, Y)$ and $\text{Cov}(U, V)$ related? What about r_{XY} and r_{UV} ?

- We will see that, provided $a \neq 0$ and $c \neq 0$, $r_{UV} = \pm r_{XY}$.

Effect of linear transformations (II)

- $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$

Effect of linear transformations (II)

- $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$
- This specializes to $\text{Var}(aX + b) = a^2\text{Var}(X)$ since $\text{Cov}(X, X) = \text{Var}(X)$.

Effect of linear transformations (II)

- $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$
- This specializes to $\text{Var}(aX + b) = a^2\text{Var}(X)$ since $\text{Cov}(X, X) = \text{Var}(X)$.
- Then,

$$\begin{aligned}r_{uv} &= \frac{\text{Cov}(U, V)}{S_U S_V} \\ &= \frac{ac\text{Cov}(X, Y)}{|a|S_X|c|S_Y} \\ &= \frac{ac}{|a||c|} r_{xy}\end{aligned}$$

Variance of linear combinations of variables

- It is easily shown that

$$S_{aX+bY}^2 = a^2S_X^2 + b^2S_Y^2 + 2abS_{XY}$$

Variance of linear combinations of variables

- It is easily shown that

$$S_{aX+bY}^2 = a^2 S_X^2 + b^2 S_Y^2 + 2ab S_{XY}$$

- For the case of a simple sum,

$$S_{X+Y}^2 = S_X^2 + S_Y^2 + 2S_{XY}$$

Variance of linear combinations of variables

- It is easily shown that

$$S_{aX+bY}^2 = a^2 S_X^2 + b^2 S_Y^2 + 2ab S_{XY}$$

- For the case of a simple sum,

$$S_{X+Y}^2 = S_X^2 + S_Y^2 + 2S_{XY}$$

- Intuition: when X and Y are negatively correlated, their fluctuations tend to compensate each other.

Simple indices (I)

- We have a series of values of a variable, ordered in time.

Simple indices (I)

- We have a series of values of a variable, ordered in time.
- Examples: yearly rain, stock prices, exchange rates...

Simple indices (I)

- We have a series of values of a variable, ordered in time.
- Examples: yearly rain, stock prices, exchange rates...
- Looking directly at the values, ordinarily in different units, makes the comparison hard.

Simple indices (I)

- We have a series of values of a variable, ordered in time.
- Examples: yearly rain, stock prices, exchange rates. . .
- Looking directly at the values, ordinarily in different units, makes the comparison hard.
- We can simplify things expressing all values in terms of a common origin.

Simple indices (II)

- How is this done?

Simple indices (II)

- How is this done?
- We “define” the value at a base period to be 100, and express all other values in proportion.

$$I_{t,0} = \frac{x_t}{x_0} \times 100$$

Simple indices (II)

- How is this done?
- We “define” the value at a base period to be 100, and express all other values in proportion.

$$I_{t,0} = \frac{x_t}{x_0} \times 100$$

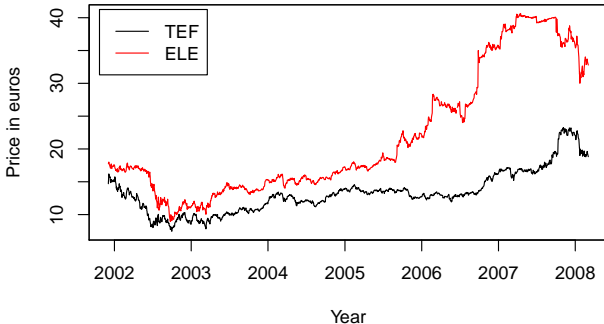
- If we want to express the index in proportion of the base period rather than in percentage, we have:

$$i_{t,0} = \frac{x_t}{x_0}$$

Simple indices (III)

- Which stock performed best from 2002 to 2004?

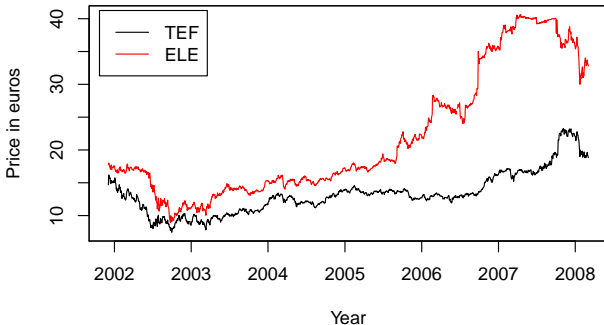
Prices of TEF and ELE



Simple indices (III)

- Which stock performed best from 2002 to 2004?

Prices of TEF and ELE

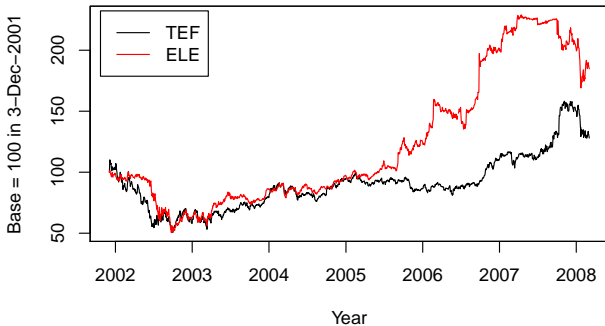


- Hard to say, different starting prices, similar profile.

Simple indices (III)

- Now it is easier!

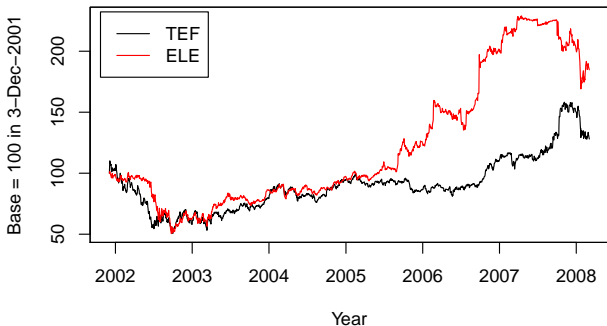
Prices of TEF and ELE



Simple indices (III)

- Now it is easier!

Prices of TEF and ELE



- The starting points have been lined up and made equal to 100. At each and every point we can compare the indices.

Chaining indices

- Very easy to chain indices:

$$i_{t,0} = \frac{x_t}{x_0} = \frac{x_t}{x_s} \times \frac{x_s}{x_0} = i_{t,s} \times i_{s,0} \quad (1)$$

Chaining indices

- Very easy to chain indices:

$$i_{t,0} = \frac{x_t}{x_0} = \frac{x_t}{x_s} \times \frac{x_s}{x_0} = i_{t,s} \times i_{s,0} \quad (1)$$

- The index base 0 at time t is the index base s at time t multiplied by the index base 0 at time s .

Chaining indices

- Very easy to chain indices:

$$i_{t,0} = \frac{x_t}{x_0} = \frac{x_t}{x_s} \times \frac{x_s}{x_0} = i_{t,s} \times i_{s,0} \quad (1)$$

- The index base 0 at time t is the index base s at time t multiplied by the index base 0 at time s .
- If indices are expressed in percentages, we have to divide by 100 each, then multiply by 100:

$$I_{t,0} = \frac{I_{t,s}}{100} \times \frac{I_{s,0}}{100} \times 100$$

Rate of change

- The **rate of change** from $t - 1$ to t is computed as follows:

$$\frac{x_t}{x_{t-1}} = 1 + \alpha_t \implies \alpha_t = \frac{x_t}{x_{t-1}} - 1$$

Rate of change

- The **rate of change** from $t - 1$ to t is computed as follows:

$$\frac{x_t}{x_{t-1}} = 1 + \alpha_t \implies \alpha_t = \frac{x_t}{x_{t-1}} - 1$$

- We can replace absolute values (prices, quantities, whatever) with indices:

$$\frac{i_{t,0}}{i_{t-1,0}} = 1 + \alpha_t \implies \alpha_t = \frac{i_{t,0}}{i_{t-1,0}} - 1$$

Average rate of change

- From that formula,

$$\alpha = \sqrt[k]{\frac{I_{t+k,0}}{I_{t,0}}} - 1$$

Average rate of change

- From that formula,

$$\alpha = \sqrt[k]{\frac{I_{t+k,0}}{I_{t,0}}} - 1$$

- It can be computed even if the period-to-period rates are unknown (we only need the values of the first and last period).

Average rate of change

- From that formula,

$$\alpha = \sqrt[k]{\frac{I_{t+k,0}}{I_{t,0}}} - 1$$

- It can be computed even if the period-to-period rates are unknown (we only need the values of the first and last period).
- If the period-to-period rates are known, it is the case that:

$$(1 + \alpha)^k = (1 + \alpha_{t+k}) \times \dots \times (1 + \alpha_{t+1})$$

Complex indices (I)

- So far, all our indices have been a simple change of units: the base period was defined as 100, and everything else changed accordingly.

Complex indices (I)

- So far, all our indices have been a simple change of units: the base period was defined as 100, and everything else changed accordingly.
- No we want to summarize in a *single* index information concerning *more than one* item or simple index.

Complex indices (I)

- So far, all our indices have been a simple change of units: the base period was defined as 100, and everything else changed accordingly.
- No we want to summarize in a *single* index information concerning *more than one* item or simple index.
- Examples: IPC (Consumer Price Index), IPI (Industrial Production Index), IBEX35 (a stock market price index), etc.

Complex indices (I)

- So far, all our indices have been a simple change of units: the base period was defined as 100, and everything else changed accordingly.
- No we want to summarize in a *single* index information concerning *more than one* item or simple index.
- Examples: IPC (Consumer Price Index), IPI (Industrial Production Index), IBEX35 (a stock market price index), etc.
- How do we go about?

Complex indices (II)

- Consider the stock exchange example, and assume only ELE and TEF were traded.

Complex indices (II)

- Consider the stock exchange example, and assume only ELE and TEF were traded.
- We might construct a stock market index as follows:

$$\tilde{I}_{t,0} = \frac{I_{t,0}^{(TEF)} + I_{t,0}^{(ELE)}}{2}$$

Complex indices (II)

- Consider the stock exchange example, and assume only ELE and TEF were traded.
- We might construct a stock market index as follows:

$$\tilde{I}_{t,0} = \frac{I_{t,0}^{(TEF)} + I_{t,0}^{(ELE)}}{2}$$

- If we have indices for more than two stocks, we might generalize that to:

$$\tilde{I}_{t,0} = \frac{\sum_{i=1}^n I_{t,0}^{(i)}}{n}$$

Complex indices (II)

- Consider the stock exchange example, and assume only ELE and TEF were traded.
- We might construct a stock market index as follows:

$$\tilde{I}_{t,0} = \frac{I_{t,0}^{(TEF)} + I_{t,0}^{(ELE)}}{2}$$

- If we have indices for more than two stocks, we might generalize that to:

$$\tilde{I}_{t,0} = \frac{\sum_{i=1}^n I_{t,0}^{(i)}}{n}$$

- This is an unweighted index: TEF would have the same weight as the smallest company among the n considered.

Complex indices (III)

- Notice: the **simple average of indices**

$$\tilde{l}_{t,0} = \frac{\sum_{i=1}^n l_{t,0}^{(i)}}{n}$$

is equivalent to

$$\tilde{l}_{t,0} = \frac{100 \times \sum_{i=1}^n \frac{x_t^{(i)}}{x_0^{(i)}}}{n}$$

Complex indices (III)

- Notice: the **simple average of indices**

$$\tilde{l}_{t,0} = \frac{\sum_{i=1}^n l_{t,0}^{(i)}}{n}$$

is equivalent to

$$\tilde{l}_{t,0} = \frac{100 \times \sum_{i=1}^n \frac{x_t^{(i)}}{x_0^{(i)}}}{n}$$

- Sometimes, the **unweighted basket index** is used:

$$\tilde{G}_{t,0} = 100 \times \frac{\sum_{i=1}^n x_t^{(i)}}{\sum_{i=1}^n x_0^{(i)}};$$

This last formula requires all x 's in same units, the previous one did not.

Complex indices (IV)

- The simple indices and the “unweighted basket index” are related:

$$\tilde{G}_{t,0} = 100 \times \frac{\sum_{i=1}^n x_t^{(i)}}{\sum_{i=1}^n x_0^{(i)}} \quad (2)$$

$$= 100 \times \frac{\sum_{i=1}^n \frac{x_t^{(i)}}{x_0^{(i)}} x_0^{(i)}}{\sum_{i=1}^n x_0^{(i)}} \quad (3)$$

$$= \frac{\sum_{i=1}^n x_0^{(i)} I_{t,0}^{(i)}}{\sum_{i=1}^n x_0^{(i)}} \quad (4)$$

$$(5)$$

Complex indices (IV)

- The simple indices and the “unweighted basket index” are related:

$$\tilde{G}_{t,0} = 100 \times \frac{\sum_{i=1}^n x_t^{(i)}}{\sum_{i=1}^n x_0^{(i)}} \quad (2)$$

$$= 100 \times \frac{\sum_{i=1}^n \frac{x_t^{(i)}}{x_0^{(i)}} x_0^{(i)}}{\sum_{i=1}^n x_0^{(i)}} \quad (3)$$

$$= \frac{\sum_{i=1}^n x_0^{(i)} I_{t,0}^{(i)}}{\sum_{i=1}^n x_0^{(i)}} \quad (4)$$

$$(5)$$

- $\tilde{G}_{t,0}$ is a weighted average of the simple indices $I_{t,0}^{(i)}$.

Weighted indices (I)

- We would rather fix our weights in our own, meaningful way.

Weighted indices (I)

- We would rather fix our weights in our own, meaningful way.
- Consumer Price Index: $w_i =$ fraction of income spent on item i .

Weighted indices (I)

- We would rather fix our weights in our own, meaningful way.
- Consumer Price Index: $w_i =$ fraction of income spent on item i .
- Stock Exchange Index: $w_i =$ market capitalization of stock i .

Weighted indices (I)

- We would rather fix our weights in our own, meaningful way.
- Consumer Price Index: $w_i =$ fraction of income spent on item i .
- Stock Exchange Index: $w_i =$ market capitalization of stock i .
- Oil Price Index: $w_i =$ fraction of oil production of each class (Brent, Texas Light, etc.)

Weighted indices (I)

- We would rather fix our weights in our own, meaningful way.
- Consumer Price Index: $w_i =$ fraction of income spent on item i .
- Stock Exchange Index: $w_i =$ market capitalization of stock i .
- Oil Price Index: $w_i =$ fraction of oil production of each class (Brent, Texas Light, etc.)
- Industrial Production Prices: $w_i =$ fraction of total output of sector i .

Weighted indices (II)

- Consider the first case: it makes sense to weight each product index by the fraction of income that item represents:

Weighted indices (II)

- Consider the first case: it makes sense to weight each product index by the fraction of income that item represents:
- Fraction of income in base year spent on i :

$$w_i = \frac{p_{i0}q_{i0}}{\sum_{j=1}^n p_{j0}q_{j0}}$$

Weighted indices (II)

- Consider the first case: it makes sense to weight each product index by the fraction of income that item represents:
- Fraction of income in base year spent on i :

$$w_i = \frac{p_{i0}q_{i0}}{\sum_{j=1}^n p_{j0}q_{j0}}$$

- Single item index:

$$I_{t,0}^{(i)} = \frac{p_{it}}{p_{i0}} \times 100$$

Weighted indices (II)

- Consider the first case: it makes sense to weight each product index by the fraction of income that item represents:
- Fraction of income in base year spent on i :

$$w_i = \frac{p_{i0}q_{i0}}{\sum_{j=1}^n p_{j0}q_{j0}}$$

- Single item index:

$$I_{t,0}^{(i)} = \frac{p_{it}}{p_{i0}} \times 100$$

- Weighted index:

$$\sum_{j=1}^n w_j I_{t,0}^{(j)} = \frac{\sum_{i=1}^n p_{i0}q_{i0} \left(\frac{p_{it}}{p_{i0}} \times 100 \right)}{\sum_{j=1}^n p_{j0}q_{j0}} = \frac{\sum_{i=1}^n p_{it}q_{i0}}{\sum_{j=1}^n p_{j0}q_{j0}} \times 100$$

The Laspeyres index

- The index just found is the **Laspeyres index**:

$$\tilde{I}_{t,0} = \frac{\sum_{i=1}^n p_{it} q_{i0}}{\sum_{j=1}^n p_{j0} q_{j0}} \times 100$$

The Laspeyres index

- The index just found is the **Laspeyres index**:

$$\tilde{I}_{t,0} = \frac{\sum_{i=1}^n p_{it} q_{i0}}{\sum_{j=1}^n p_{j0} q_{j0}} \times 100$$

- It is just the ratio of the values of a basket of n goods in quantities $q_{10}, q_{20}, \dots, q_{n0}$ valued at the prices prevailing at the 0 and t periods. The period 0 is the base period.

The Laspeyres index

- The index just found is the **Laspeyres index**:

$$\tilde{I}_{t,0} = \frac{\sum_{i=1}^n p_{it} q_{i0}}{\sum_{j=1}^n p_{j0} q_{j0}} \times 100$$

- It is just the ratio of the values of a basket of n goods in quantities $q_{10}, q_{20}, \dots, q_{n0}$ valued at the prices prevailing at the 0 and t periods. The period 0 is the base period.
- The quantities $q_{10}, q_{20}, \dots, q_{n0}$ remain fixed for the duration of the index.

The Laspeyres index

- The index just found is the **Laspeyres index**:

$$\tilde{I}_{t,0} = \frac{\sum_{i=1}^n p_{it} q_{i0}}{\sum_{j=1}^n p_{j0} q_{j0}} \times 100$$

- It is just the ratio of the values of a basket of n goods in quantities $q_{10}, q_{20}, \dots, q_{n0}$ valued at the prices prevailing at the 0 and t periods. The period 0 is the base period.
- The quantities $q_{10}, q_{20}, \dots, q_{n0}$ remain fixed for the duration of the index.

The Paasche index

- Rather than compare values of a constant basket of goods in quantities q_{i0} , it compares a *varying* basket with quantities q_{it} .

$$\tilde{I}_{t,0} = \frac{\sum_{i=1}^n p_{it} q_{it}}{\sum_{j=1}^n p_{j0} q_{jt}} \times 100$$

The Paasche index

- Rather than compare values of a constant basket of goods in quantities q_{i0} , it compares a *varying* basket with quantities q_{it} .

$$\tilde{I}_{t,0} = \frac{\sum_{i=1}^n p_{it} q_{it}}{\sum_{j=1}^n p_{j0} q_{jt}} \times 100$$

- More cumbersome; it requires all quantities q_{it} for all t at which we want to compute the index (Laspeyres used always q_{i0}).

The Paasche index

- Rather than compare values of a constant basket of goods in quantities q_{i0} , it compares a *varying* basket with quantities q_{it} .

$$\tilde{I}_{t,0} = \frac{\sum_{i=1}^n p_{it} q_{it}}{\sum_{j=1}^n p_{j0} q_{jt}} \times 100$$

- More cumbersome; it requires all quantities q_{it} for all t at which we want to compute the index (Laspeyres used always q_{i0}).
- The Laspeyres index tends to overstate inflation, the Paasche index to understate it.

Chaining of Laspeyres indices (I)

- An index such as $i_{t,0} = \tilde{I}_{t,0}/100$ answers the question: “How much have prices changed from 0 to t ?”

Chaining of Laspeyres indices (I)

- An index such as $i_{t,0} = \tilde{I}_{t,0}/100$ answers the question: “How much have prices changed from 0 to t ?”
- We could choose any other pair of moments. Then, $i_{t,t-1}$ for instance answers the question: “How much have prices changed from $t - 1$ to t ?”

Chaining of Laspeyres indices (I)

- An index such as $i_{t,0} = \tilde{I}_{t,0}/100$ answers the question: “How much have prices changed from 0 to t ?”
- We could choose any other pair of moments. Then, $i_{t,t-1}$ for instance answers the question: “How much have prices changed from $t - 1$ to t ?”
- Hence, it makes sense to define a **chained Laspeyres index** as follows:

$$i_{t,0} = i_{t,t-1} \times i_{t-1,t-2} \times \dots \times i_{1,0}$$

Chaining of Laspeyres indices (II)

- Notice! This allows the quantities to change from period to period (not necessarily each and every period).

Chaining of Laspeyres indices (II)

- Notice! This allows the quantities to change from period to period (not necessarily each and every period).
- The Spanish Consumer Price Index (IPC) base 2006 is computed as a chained Laspeyres index.

Chaining of Laspeyres indices (II)

- Notice! This allows the quantities to change from period to period (not necessarily each and every period).
- The Spanish Consumer Price Index (IPC) base 2006 is computed as a chained Laspeyres index.
- Each year, the quantities are recalculated, and the indices computed with reference to the previous December.

Chaining of Laspeyres indices (III)

- For the j month of year $t + 1$ the index is computed as:

$$\frac{\sum p_{i,(j,t+1)} q_{i,(12,t)}}{\sum p_{i,(12,t)} q_{i,(12,t)}} \times \frac{\sum p_{i,(12,t)} q_{i,(12,t-1)}}{\sum p_{i,(12,t-1)} q_{i,(12,t-1)}} \times \dots$$

Chaining of Laspeyres indices (III)

- For the j month of year $t + 1$ the index is computed as:

$$\frac{\sum p_{i,(j,t+1)} q_{i,(12,t)}}{\sum p_{i,(12,t)} q_{i,(12,t)}} \times \frac{\sum p_{i,(12,t)} q_{i,(12,t-1)}}{\sum p_{i,(12,t-1)} q_{i,(12,t-1)}} \times \dots$$

- The basket of good changes each year, but the index behaves like a Laspeyres index *within* the year.

Chaining of Laspeyres indices (III)

- For the j month of year $t + 1$ the index is computed as:

$$\frac{\sum p_{i,(j,t+1)} q_{i,(12,t)}}{\sum p_{i,(12,t)} q_{i,(12,t)}} \times \frac{\sum p_{i,(12,t)} q_{i,(12,t-1)}}{\sum p_{i,(12,t-1)} q_{i,(12,t-1)}} \times \dots$$

- The basket of good changes each year, but the index behaves like a Laspeyres index *within* the year.
- Each year, the quantities are recalculated, and the indices computed with reference to the previous December.

Nominal and real terms: deflated series

- **Nominal** or monetary values are expressed plainly in units of currency. euros, US\$ and the like.

Nominal and real terms: deflated series

- **Nominal** or monetary values are expressed plainly in units of currency. euros, US\$ and the like.
- Currency units do not have a constant value over time, which makes comparisons difficult.

Nominal and real terms: deflated series

- **Nominal** or monetary values are expressed plainly in units of currency. euros, US\$ and the like.
- Currency units do not have a constant value over time, which makes comparisons difficult.
- **Real** or deflated values are nominal values after taking into account changes in value of the currency: they are expressed in “real” currency units, referred to a time period: “In pesetas of 1980”, “in 1945 US\$”, etc.

Example: real and monetary wages

- Consider

Period	Salary	Price index	Real salary
1980	1354	100	1354
1985	1678	123	1364
1990	2450	212	1155
1995	2630	223	1179
2000	2810	232	1211

Example: real and monetary wages

- Consider

Period	Salary	Price index	Real salary
1980	1354	100	1354
1985	1678	123	1364
1990	2450	212	1155
1995	2630	223	1179
2000	2810	232	1211

- Real salary = $\frac{\text{Salary}}{\text{Price index}} \times 100$

Example: real terms of trade

- Ratio of exports price index to imports price index.

$$TOT = \frac{\frac{\sum_{i=1}^n p_{it}^X q_{i0}^X}{\sum_{j=1}^n p_{j0}^X q_{j0}^X}}{\frac{\sum_{i=1}^n p_{it}^I q_{i0}^I}{\sum_{j=1}^n p_{j0}^I q_{j0}^I}}$$

Example: real terms of trade

- Ratio of exports price index to imports price index.

$$TOT = \frac{\frac{\sum_{i=1}^n p_{it}^X q_{j0}^X}{\sum_{j=1}^n p_{j0}^X q_{j0}^X}}{\frac{\sum_{i=1}^n p_{it}^I q_{i0}^I}{\sum_{j=1}^n p_{j0}^I q_{j0}^I}}$$

- An increase in the TOT means that the country's export prices are faring better than its import prices —hence with the same exports the country can finance more imports than in the base year.

Example: real terms of trade

- Ratio of exports price index to imports price index.

$$TOT = \frac{\frac{\sum_{i=1}^n p_{it}^X q_{j0}^X}{\sum_{j=1}^n p_{j0}^X q_{j0}^X}}{\frac{\sum_{i=1}^n p_{it}^I q_{i0}^I}{\sum_{j=1}^n p_{j0}^I q_{j0}^I}}$$

- An increase in the TOT means that the country's export prices are faring better than its import prices —hence with the same exports the country can finance more imports than in the base year.
- Usually Laspeyres indices for numerator and denominator.

Example: implicit deflator of GDP

- Weighted price index of goods produced by an economy.

Example: implicit deflator of GDP

- Weighted price index of goods produced by an economy.
- Used to deflate nominal magnitudes in the National Accounts.

Example: implicit deflator of GDP

- Weighted price index of goods produced by an economy.
- Used to deflate nominal magnitudes in the National Accounts.
- Can sometimes be quite different from consumer price indices (imported inflation, etc.)

Chaining indices (I)

- Sometimes, there is not an index covering the whole time span we are interested in.

Chaining indices (I)

- Sometimes, there is not an index covering the whole time span we are interested in.
- Then, we may resort to chaining two indices.

Chaining indices (I)

- Sometimes, there is not an index covering the whole time span we are interested in.
- Then, we may resort to chaining two indices.
- A simple proportional rule is all that is required, and at least a period of overlap.

Chaining indices (I)

- Sometimes, there is not an index covering the whole time span we are interested in.
- Then, we may resort to chaining two indices.
- A simple proportional rule is all that is required, and at least a period of overlap.
- Notice the homogeneity of the index is lost.

Chaining indices (II)

- Let's look at the table next:

Year	Price Index (base 1980 = 100)	Price Index (base 1990 = 100)
1980	100	—
1985	132	—
1990	144	100
1995	163	114
2000	—	120

Chaining indices (II)

- Let's look at the table next:

Year	Price Index (base 1980 = 100)	Price Index (base 1990 = 100)
1980	100	—
1985	132	—
1990	144	100
1995	163	114
2000	—	120

- Suppose we are required to calculate the equivalent in year 2000 of 350€ of year 1980.

Chaining indices (III)

Year	Price Index (base 1980 = 100)	Price Index (base 1990 = 100)
1980	100	—
1985	132	—
1990	144	100
1995	163	114
2000	—	120

- 350€ of year 1980 are $\frac{350}{100} \times 144 = 504€$ in 1990.

Chaining indices (III)

Year	Price Index (base 1980 = 100)	Price Index (base 1990 = 100)
1980	100	--
1985	132	--
1990	144	100
1995	163	114
2000	--	120

- 350€ of year 1980 are $\frac{350}{100} \times 144 = 504\text{€}$ in 1990.
- 504€ of 1990 are $\frac{504}{100} \times 120 = 604.80\text{€}$ of year 2000.

Fisher's "Ideal Index"

- It is the harmonic mean of the Paasche and Laspeyres indices:

$$I_{\text{Fisher}} = \sqrt{I_{\text{Laspeyres}} I_{\text{Paasche}}}$$

Fisher's "Ideal Index"

- It is the harmonic mean of the Paasche and Laspeyres indices:

$$I_{\text{Fisher}} = \sqrt{I_{\text{Laspeyres}} I_{\text{Paasche}}}$$

- Why? It is the only one to verify a number of properties, desirable in principle (circularity, time reversal, etc.)

Fisher's "Ideal Index"

- It is the harmonic mean of the Paasche and Laspeyres indices:

$$I_{\text{Fisher}} = \sqrt{I_{\text{Laspeyres}} I_{\text{Paasche}}}$$

- Why? It is the only one to verify a number of properties, desirable in principle (circularity, time reversal, etc.)
- Lacks the easy interpretation of the Laspeyres index, which is in general preferred.





Fisher's "Ideal Index"

- It is the harmonic mean of the Paasche and Laspeyres indices:




$$I_{\text{Fisher}} = \sqrt{I_{\text{Laspeyres}} I_{\text{Paasche}}}$$

- Why? It is the only one to verify a number of properties, desirable in principle (circularity, time reversal, etc.)
- Lacks the easy interpretation of the Laspeyres index, which is in general preferred.
- Details in Vogt, A. - Barta, J. *The Making of Tests for Index Numbers*, Physica-Verlag, 1997.

Textbooks and monographs I

-  M.J. Bárcena, K. Fernández, E. Ferreira, and A. Garín.
Elementos de Probabilidad y Estadística Descriptiva.
Serv. Editorial UPV/EHU, Bilbao, 2003.
-  P. Dalgaard.
Introductory statistics with R.
Statistics and Computing. Springer-Verlag, 2002.
Signatura: 519.682 DAL.
-  A.M. Montiel, F. Rius, and F.J. Barón.
Elementos básicos de Estadística Económica y Empresarial.
Prentice-Hall, 1997.
-  D. Peña and J. Romo.
Introducción a la Estadística para las Ciencias Sociales.
McGraw-Hill, 1997.

Textbooks and monographs II

-  F.J. Martín Pliego and L. Ruiz Maya.
Problemas de Probabilidad.
Editorial AC, Madrid, 2002.
-  F.J. Martín Pliego and L. Ruiz Maya.
Estadística I: Probabilidad.
Editorial AC, Madrid, 2 edition, 2004.
-  V. Tomeo and I. Uña.
Lecciones de Estadística Descriptiva.
Thomson, 2003.