Lisbon School of Economics & Management
Universidade de Lisboa

Data Analysis for Economics and Business

Lectures 1 to 3:
Introduction
Main features of empirical analysis of economic and business information

Carlos J.S. Lourenço
carloslourenco@iseg.ulisboa.pt

All course slides adapted from ISEG's Prof. Zbigniew Mogila & Patricia C. Melo

Structure of lecture

- Course objectives, programme & assessment
- Learning resources
- Some examples to get you started
- Main steps in doing empirical research
- Correlation vs causation in social sciences
- Basic notions
- Types of variables

Learning outcomes

- Explain some of the main concepts used in descriptive statistics

- Calculate absolute and relative frequency tables for different types of variables

Module objectives & handbook

1. To ***help*** you develop good quantitative analysis ***skills*** of economic and business data

2. To ***help*** you develop basic data analysis programming and spreadsheet (Excel) ***skills***

3. To ***help*** you develop good communication ***skills*** when reporting quantitative analyses

Lisbon School of Economics & Management
Universidade de Lisboa

A few basic concepts to be aware of

What do economic indicators talk to us about? (1)

Nominal versus real changes

Year 2020

2 apples for 2$ each

INFLATION

Year 2021

2 apples for 3$ each

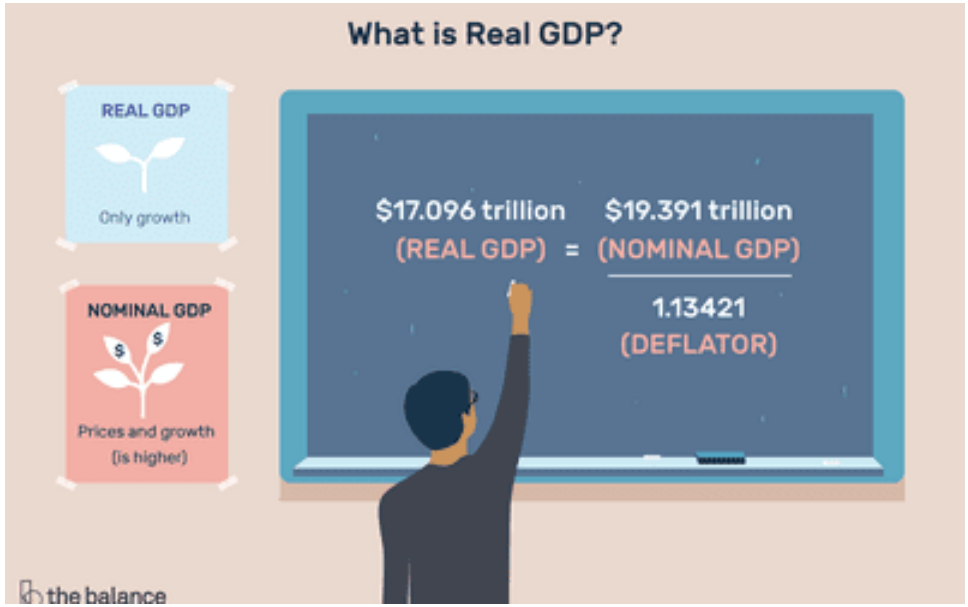
Total nominal value: 4$ (2 apples x 2$)

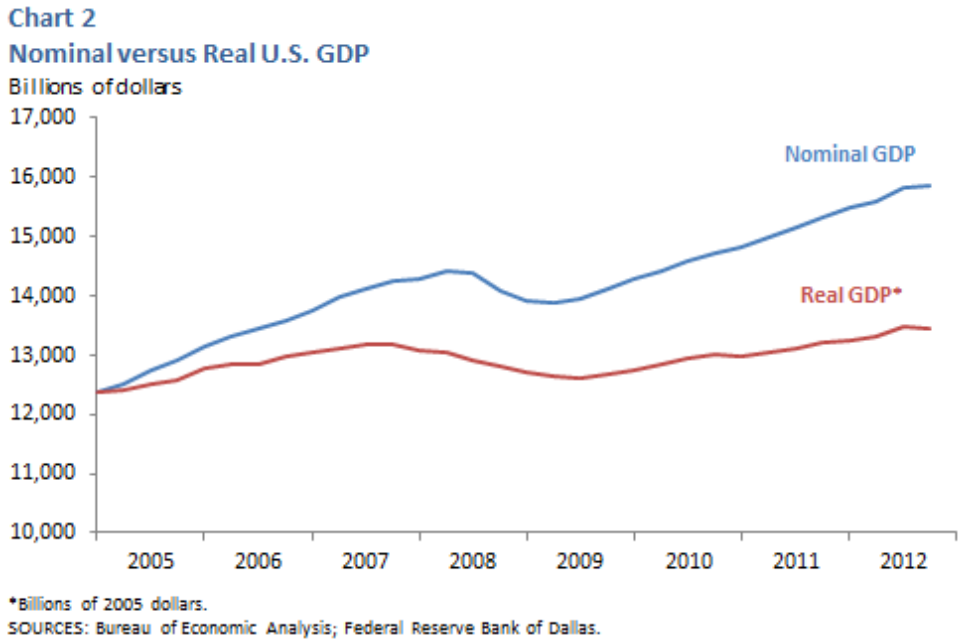
Total nominal value: 6$ (2 apples x 3$)

Did our wealth change in real terms ?

What do economic indicators talk to us about? (2)

Nominal versus real changes



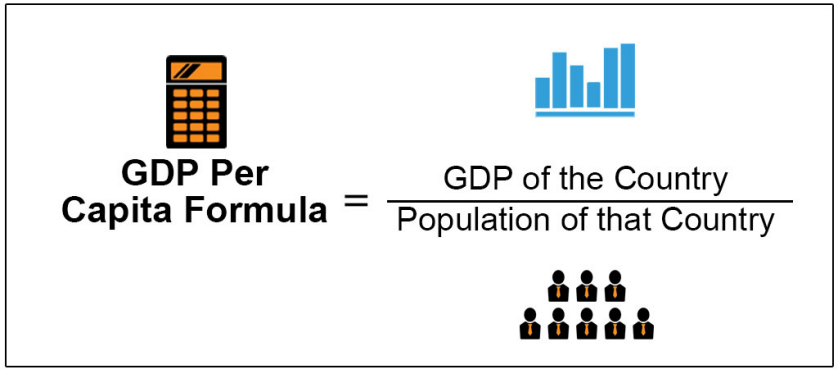


What do economic indicators talk to us about? (3)

Total value versus per capita value

Total GDP in 2018

1. United States ($21.4 bln)
2. China ($15.5 bln)
3. Japan: ($5.4 bln)
4. Germany: ($4.4 bln)
5. India: ($3.2 bln)
6. France: ($3.1 bln)
7. United Kingdom: ($3.0 bln)
8. Italy: ($2.3 bln)
9. Brazil: ($2.2 bln)
10. Canada: ($1.9 bln)



GDP per capita in 2018

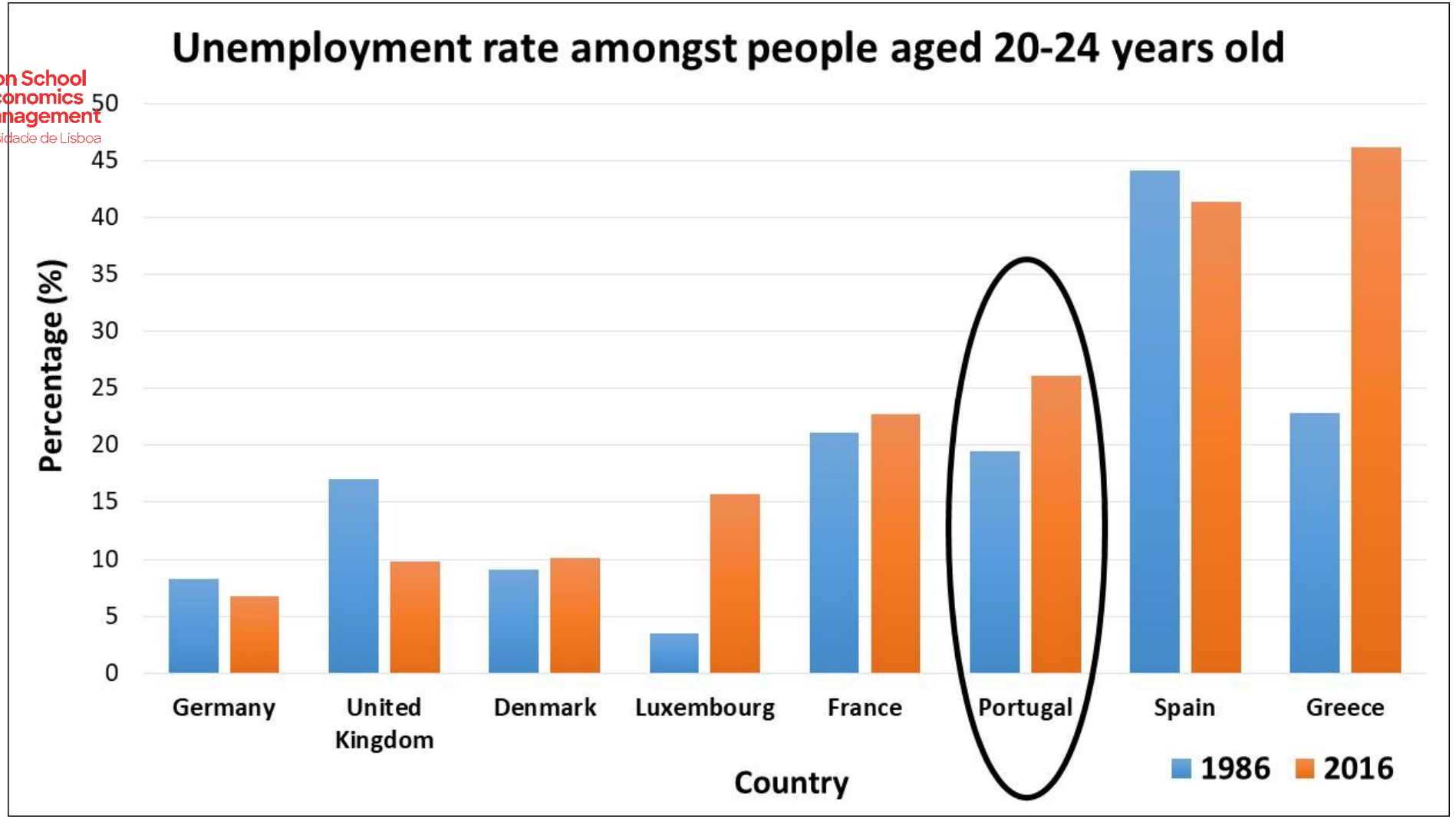
1. Luxembourg ($125.8 thous.)
2. Bermuda ($98 thous.)
3. Iceland ($93.3 thous.)
4. Switzerland ($90.7 thous.)
5. Macau ($89.9 thous.)
6. Norway ($85.6 thous.)
7. Ireland ($83.8 thous.)
8. Qatar ($68.5 thous.)
9. Denmark ($66.9 thous.)
10. United States ($65.1 thous.)

…

84. China ($10.8 thous.)

Lisbon School of Economics & Management
Universidade de Lisboa

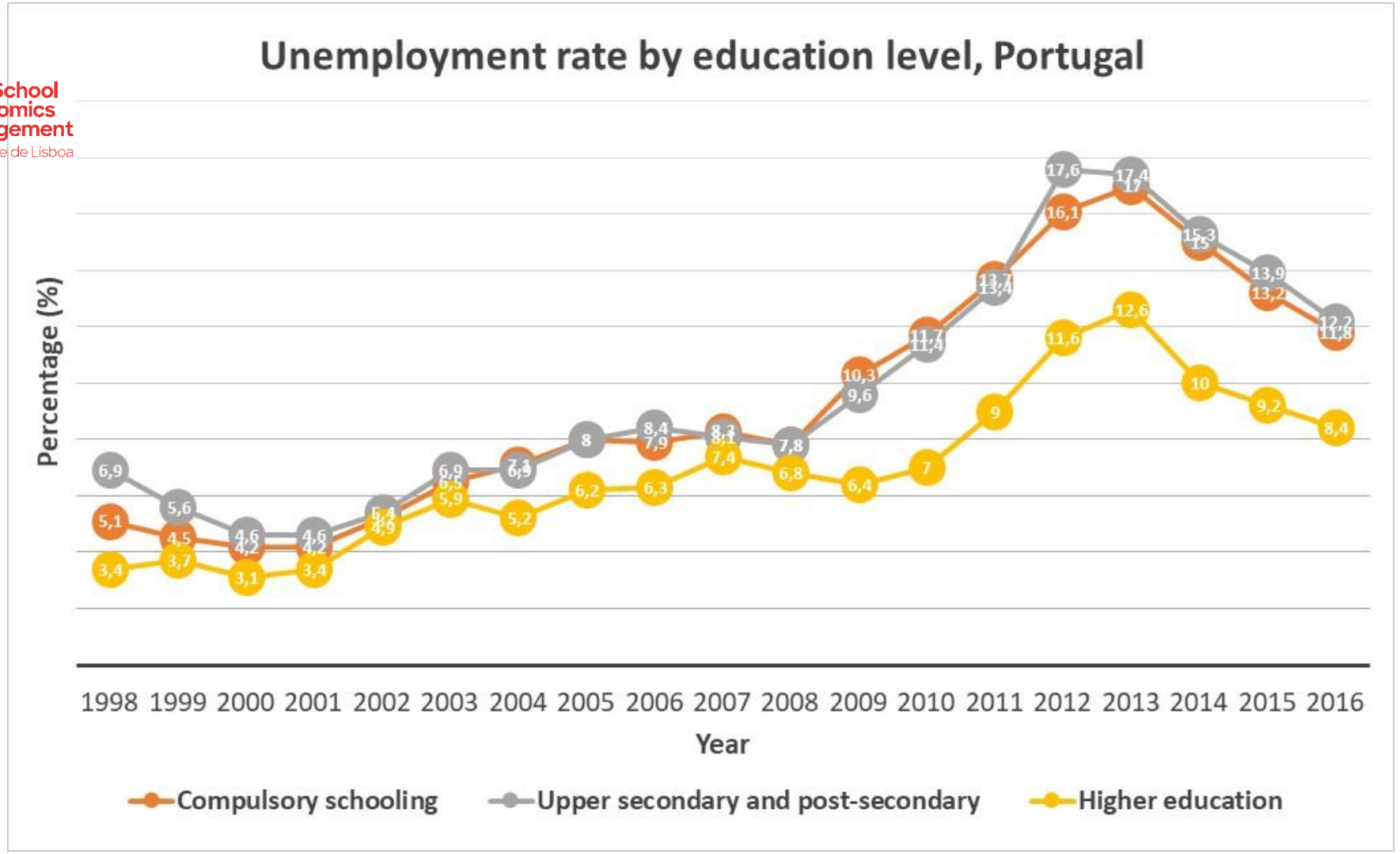
Let's consider some simple examples to get you warmed up



Source: Pordata.

- How many times was a young adult in Portugal more likely to be unemployed compared to Germany?
- Does this mean you should move to Germany?

Lisbon School of Economics & Management
Universidade de Lisboa

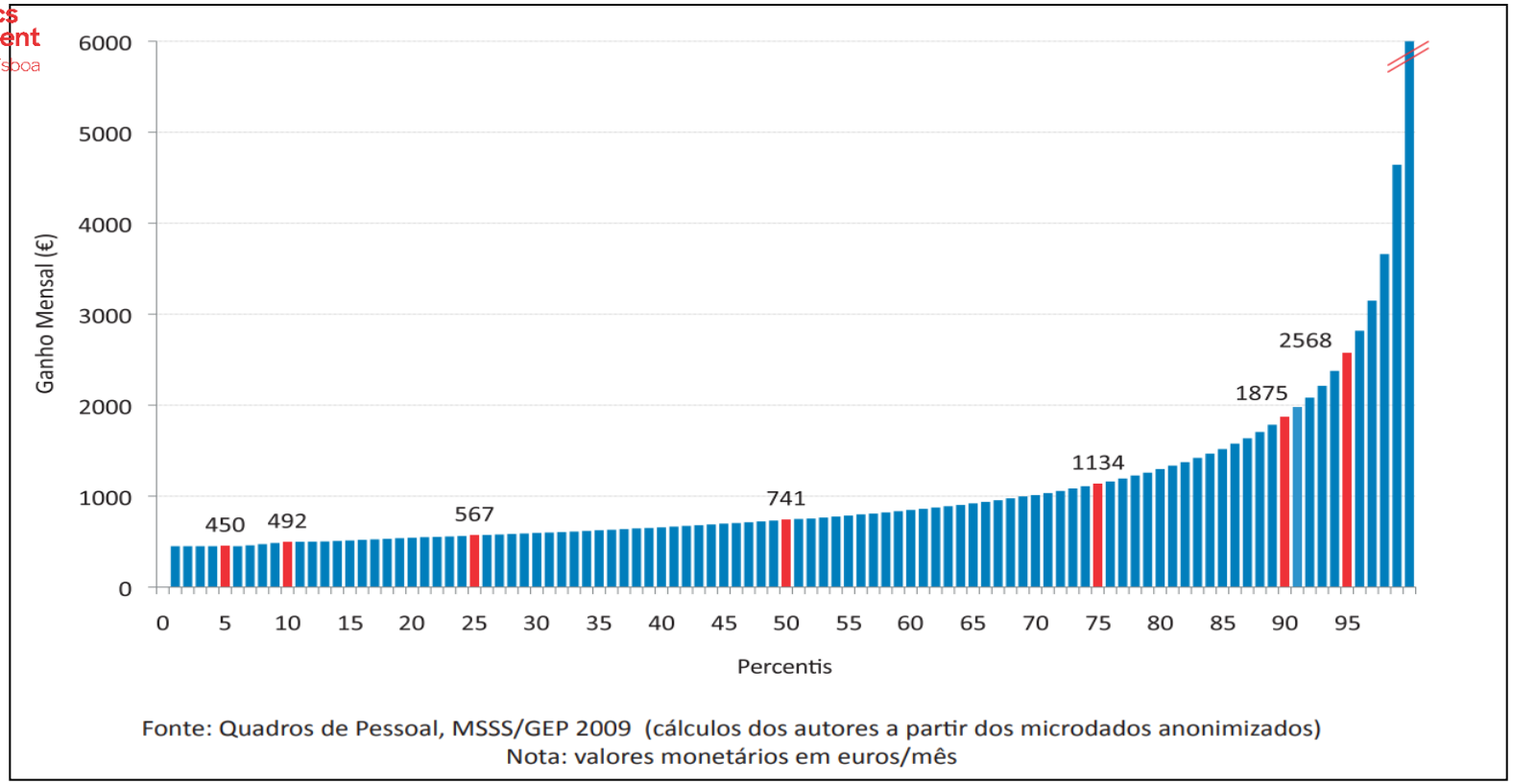


Source: Pordata.

- What is the relationship between education level and unemployment rate?
- Has the relation between education and unemployment increased or reduced over time? Why?

Lisbon School of Economics & Management
Universidade de Lisboa

Monthly wage by percentiles, Portugal, 2009



Source: Farinha, C., Figueiras, R. and Junqueira, V. (2012) Desigualdade Económica em Portugal. Fundação Francisco Manuel dos Santos.

- What was the median wage in Portugal in 2009?
- How many times the percentile 95 wage higher compared to the percentile 5 wage?
- What was the % of people earning 4500 Euros or more per month?

Distribution of individual income in two countries (cont.)

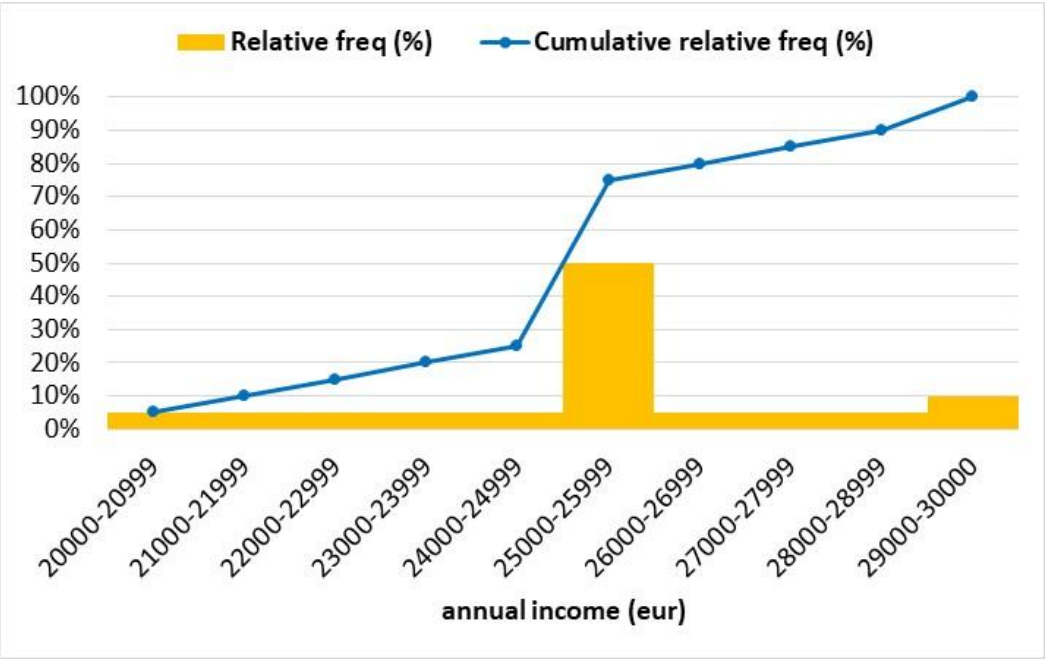
Suppose you also know information about the top and bottom 25% of the income distribution in country A and country B:

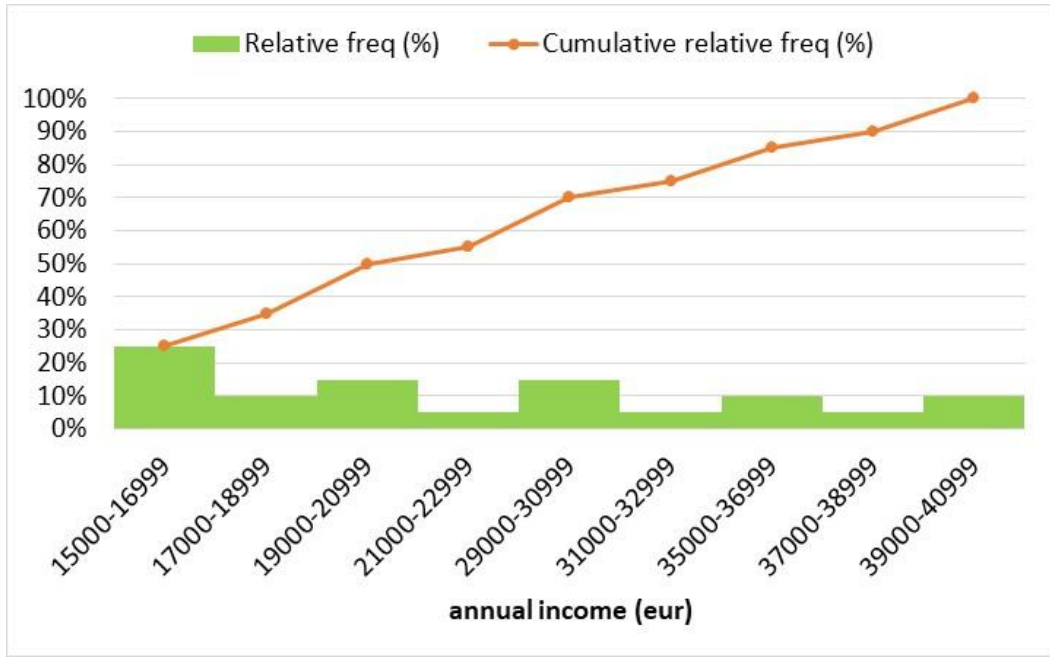
Euros per year	Country A	Country B
Average annual income	25,000.00	25,000.00
Median annual income	21,000.00	25,000.00
Lower quartile (bottom 25%) income	16,500.00	24,750.00
Upper quartile (top 25%) income	32,750.00	25,250.00

- Which country would you like to live in? Why?
- Did you change your answer, if so, why?

Distribution of individual income in two countries (cont.)

- A or B? Match A & B with an income distribution
- Why?





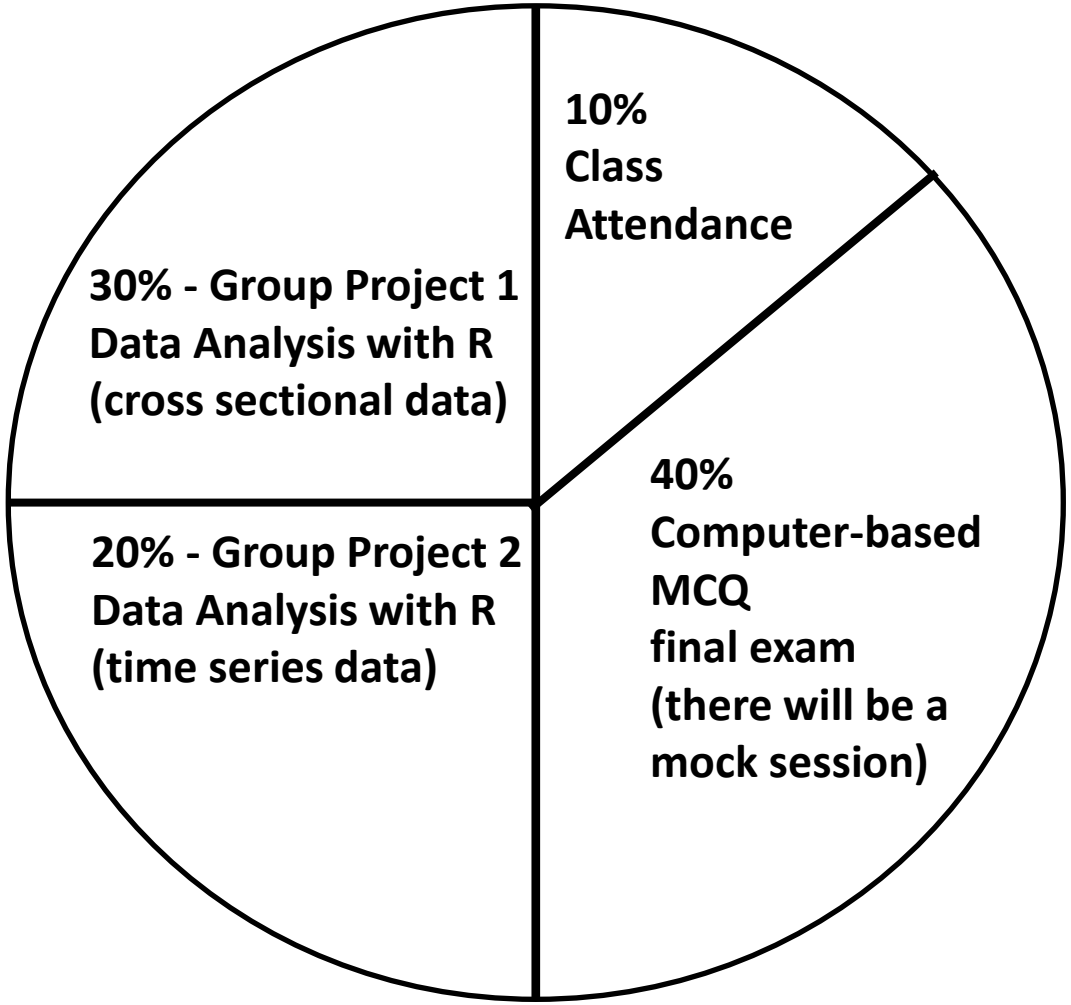
Lisbon School of Economics & Management
Universidade de Lisboa

Overview of module programme

1. Main concepts
2. Data reduction: descriptive statistics (location, dispersion, concentration)
3. Relationships between variables: association and simple regression
5. Variations over time, index numbers, and the deflator

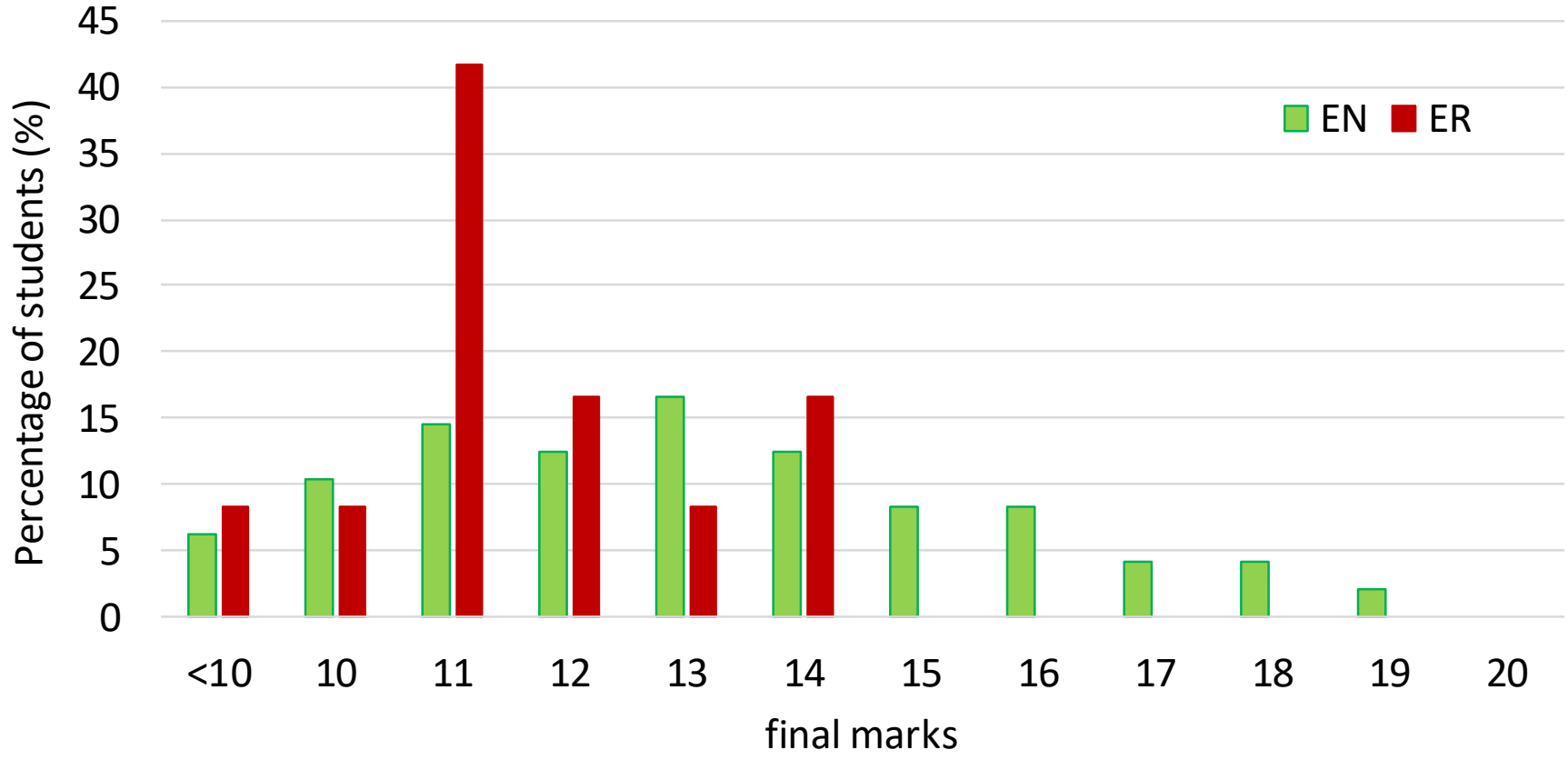
Lisbon School of Economics & Management
Universidade de Lisboa

Course assessment



Lisbon School of Economics & Management
Universidade de Lisboa

Course final marks, Management degree, 2017/18



Economic and Business Information Analysis

What do we (typically) focus on?

- **Decisions and performance of economic agents** – firms, households, individuals, governments, etc. - at the **micro, meso and macro level**

- Examples:
  - Firms: location, sales, investment, recruitment, mergers, etc.
  - Households and individuals: earnings, income, unemployment, education level, poverty, etc.
  - Regions, countries: unemployment, education, average earnings, cost of living, poverty, etc.

Economic and Business Information Analysis

Why is it important? **Why** does it matter?

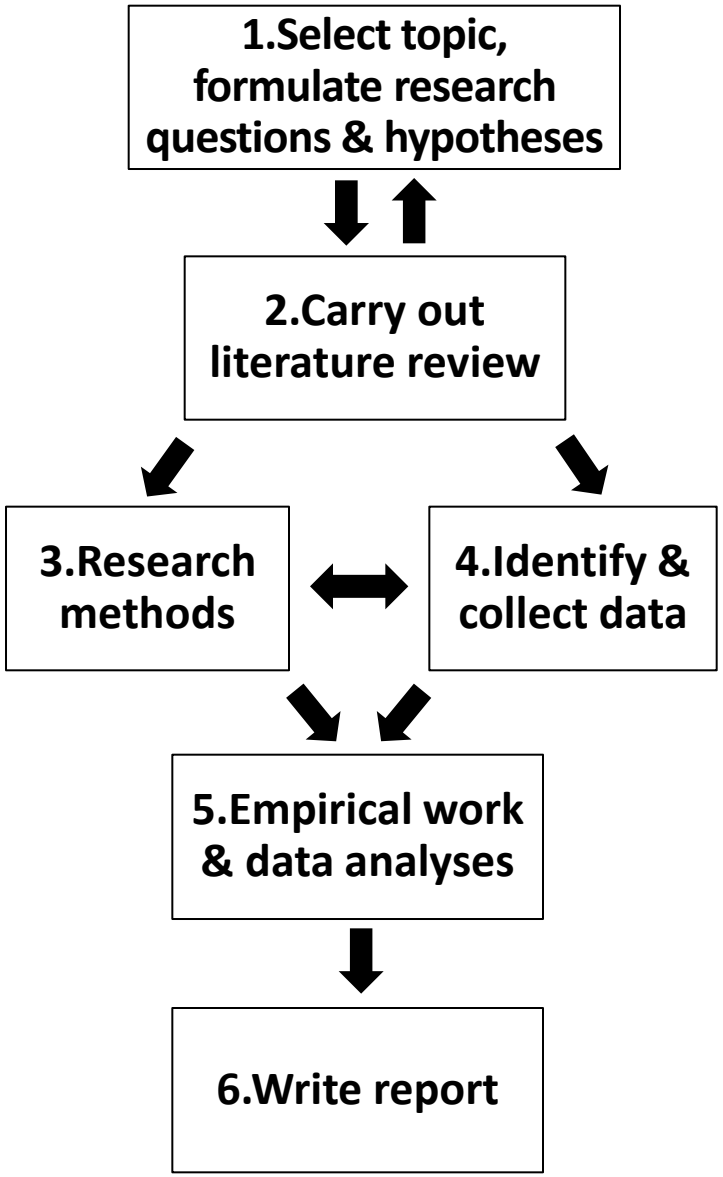
- To help us understand how economic and business phenomena evolve over time and between groups (people, firms, households, regions, countries, etc.),
- To help us understand the relationships between economic variables and how they can be influenced by alternative policy interventions and business decisions,
- To help us distinguish between correlation and causation effects,
- To support evidence-based decisions by policymakers in central and local government, business managers, households, etc.

Economic and Business Information Analysis

How do we do it?

- Using **quantitative data analysis techniques and methods**, ranging from basic descriptive statistics to sophisticated statistical models, to empirically investigate and test hypotheses and theories

Main steps in empirical research



Main steps in empirical research

3. Select appropriate research methods

- from simple to very complex modelling techniques:
    - **Descriptive statistics** making comparisons between and within groups, measuring change over time and across space, etc.
    - **Graphical display** of variables and relationships to be tested
    - **Correlation analysis and regression models**

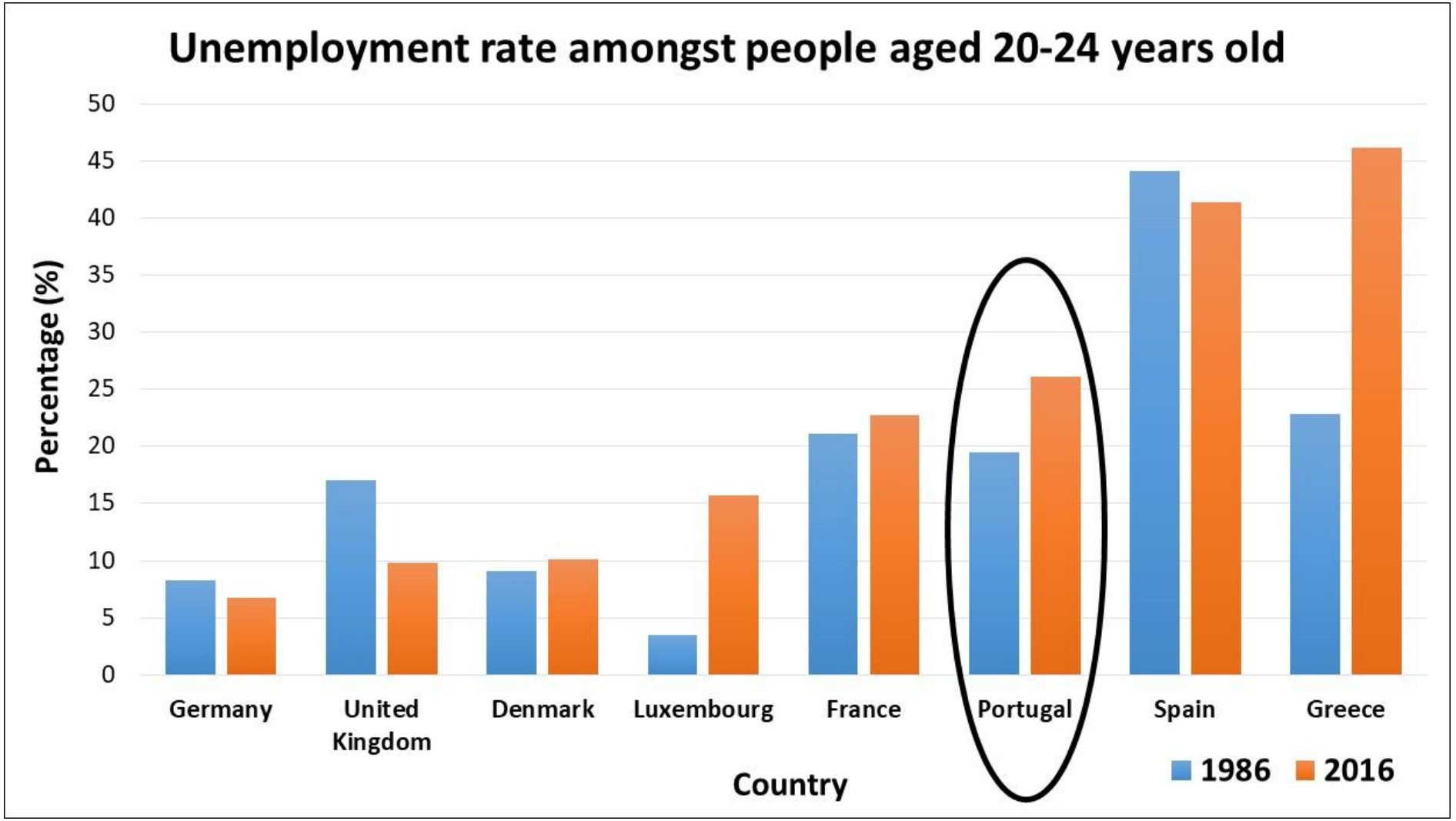
Main steps in empirical research

4. Identifiy, collect, prepare the data for empirical work

- Two main approaches to collecting data:
  - **Primary data**: you collect your own data through questionnaires, interviews, lab and/or field experiments, etc.
  - **Secondary data**: you collect data from others, e.g.: National Statistics Office, Central Banks, EUROSTAT, etc.

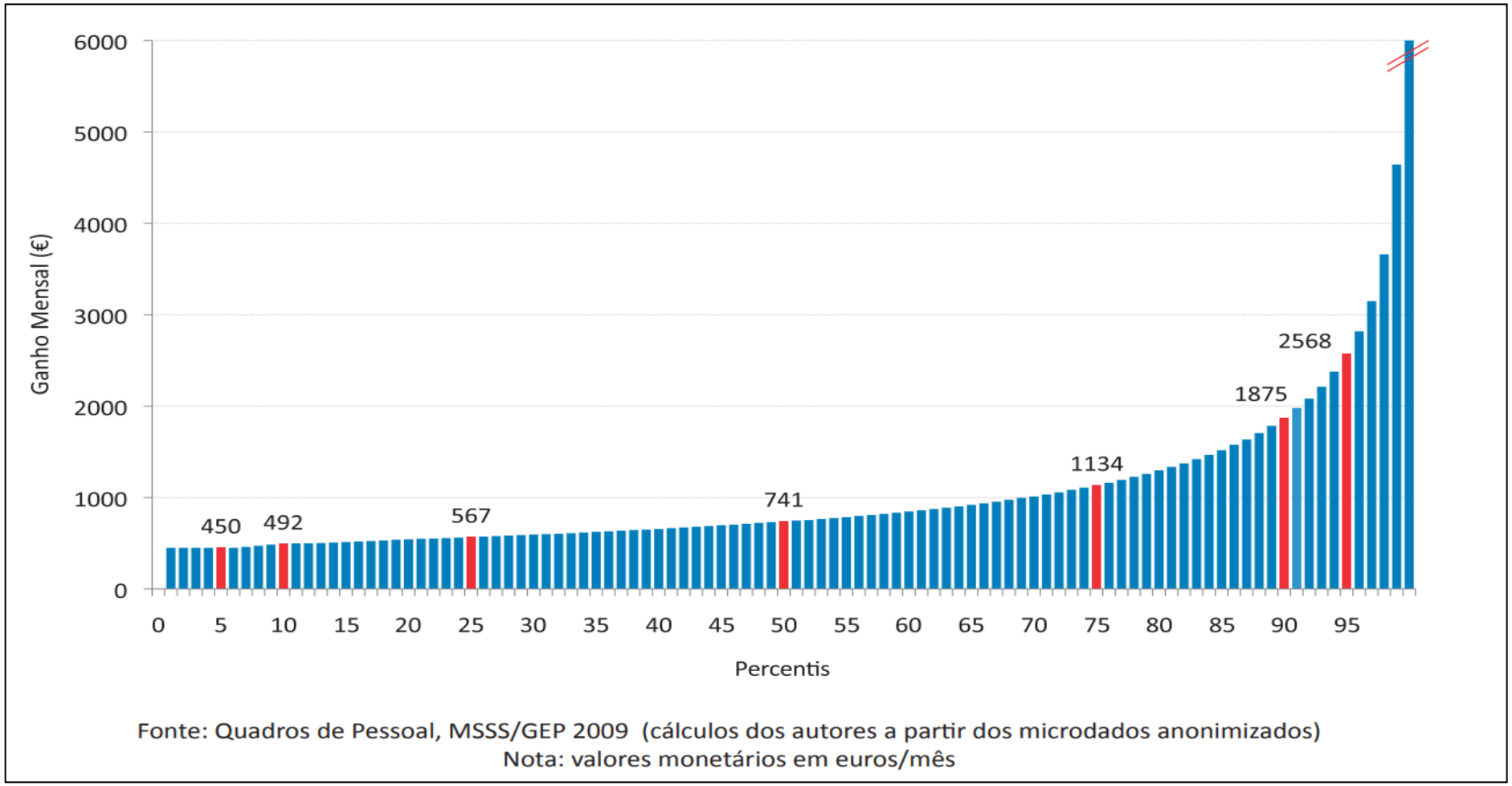
- Types of data for empirical work:
  - **Time series (TS) data**: Multiple time periods for one unit (e.g. Portugal in 1986-2016)
  - **Cross-sectional (CS) data**: Multiple units observed only once (e.g. EU countries in 2016)
  - **Panel data or longitudinal (PD) data**: Multiple units observed multiple times (e.g. EU countries in 1986-2016)

Types of data – TS, CS, PD??



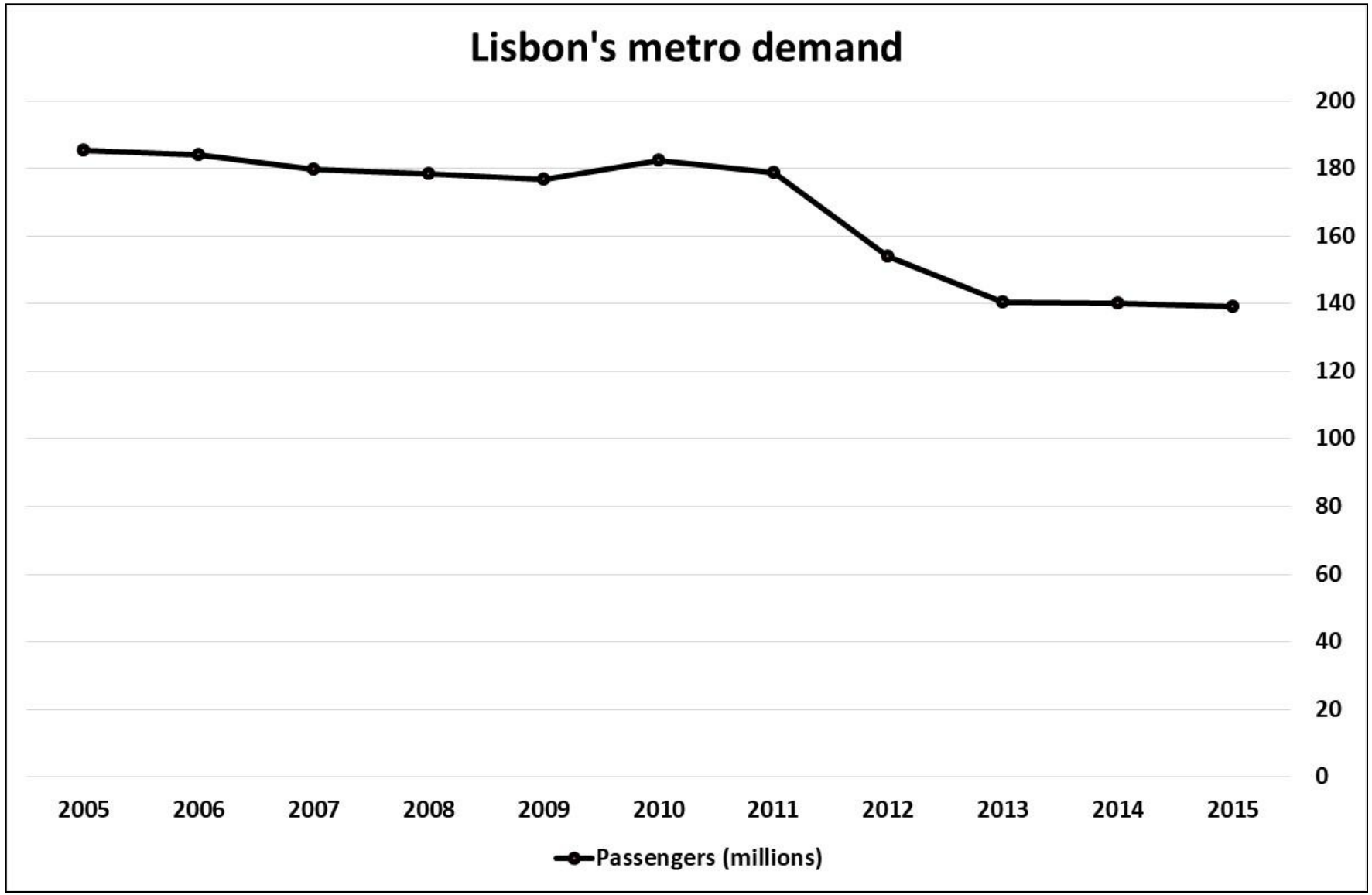
Source: Pordata.

Types of data – TS, CS, PD??



Source: Farinha, C., Figueiras, R. and Junqueira, V. (2012) Desigualdade Económica em Portugal. Fundação Francisco Manuel dos Santos.

Types of data – TS, CS, PD??

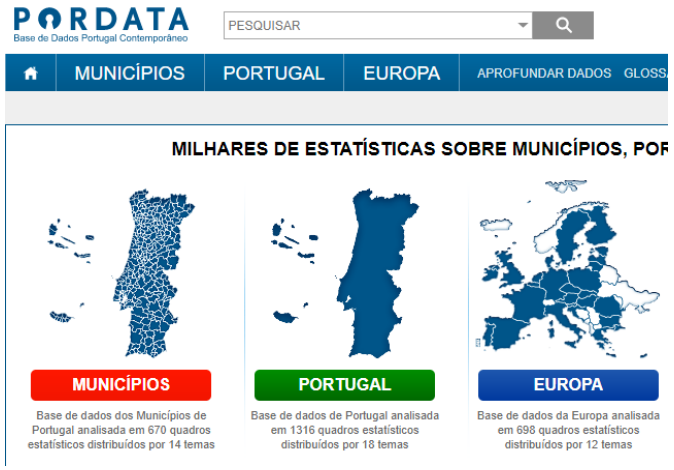


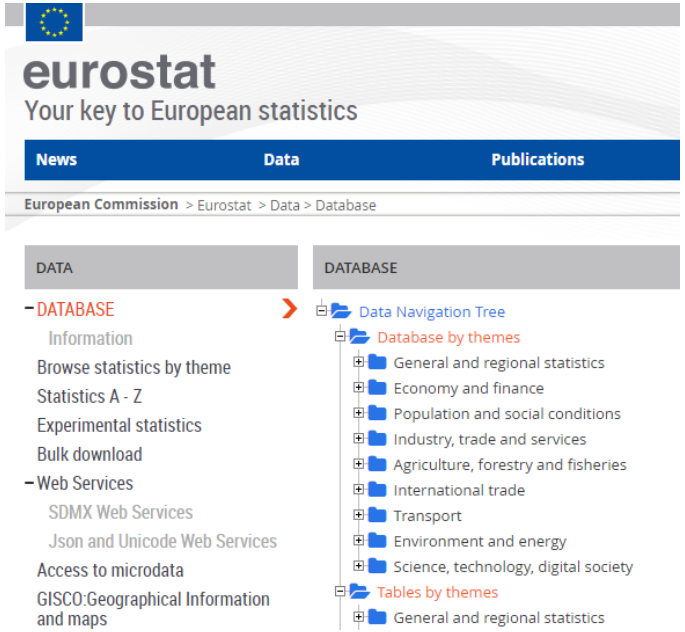
Source: Metropolitano de Lisboa.

Main steps in empirical research

4. Identifiy, collect and prepare the data for empirical work

- Some main sources of secondary data for Portugal and Europe:
  - INE: https://www.ine.pt
  - Banco de Portugal: https://www.bportugal.pt
  - Pordata: https://www.pordata.pt/
  - Eurostat: http://ec.europa.eu/eurostat/data/database
  - European Commission (e.g. AMECO): http://ec.europa.eu/economy_finance/ameco
  - OECD: http://stats.oecd.org
  - etc





Main steps in empirical research

5. Carry out empirical work

- After preparing the data collected and describing the research methods, apply empirical methods and techniques to the data:
  - **Preliminary data analysis** using descriptive statistics and graphical analysis
  - **More structured empirical analysis** using correlation analysis and regression techniques to test the hypothesis

- **Document all the steps of the empirical work** – other people should be able to replicate your analysis and get the same results

- **Be careful: correlation is not causation!**

Correlation versus Causation

- **Correlation:**
  - When two or more variables move together or follow the same trend in a systematic way
  - Correlation can result from the existence of a common unmeasured or unobserved cause, or by chance (spurious relation)

- **Causation:**
  - When changes in a given variable (*cause*) lead to a systematic change in another variable (*effect*)
  - Establishing causation in social sciences is difficult because we generally have observational (rather than experimental) data

- **It is causation, and not correlation, that should guide decision making in public and private sectors**

Main steps in empirical research

6. Writing the project report

- Before you start ask yourself:
  - **Who is the audience/reader** - policy? academics? professional?
  - **What is the purpose** of the report? - communicate knowledge? Raise awareness? Different purposes call for different styles

- Define the **structure** (i.e. sections of the repots)

- Define the **presentation style** (e.g. language, citation)

- Define the **formatting rules** (font size, margins, etc.)

Standard report structure (see/use R Markdown with template)

Citing and using other people's work

- Be careful with **PLAGIARISM**: when using and refering to other people's work, you must explicitly acknowledge the source

- Example 1 - refering to others:

  Gómez-Ibáñez and Meyer (1993) analyze transportation concessions in many industrialized countries and find that renegotiations are also common.

- Example 2 - quotation:

  Cowell (2000 pp 23) argues: "It is almost essential to attempt to 'account for' the level of, or trend in, inequality by components of the population."

References

- There are different styles of presenting references and they can differ between journals and editors. **Chose one style and stick to it - be coherent**.
- Different types of sources (books, journal articles) are referenced differently

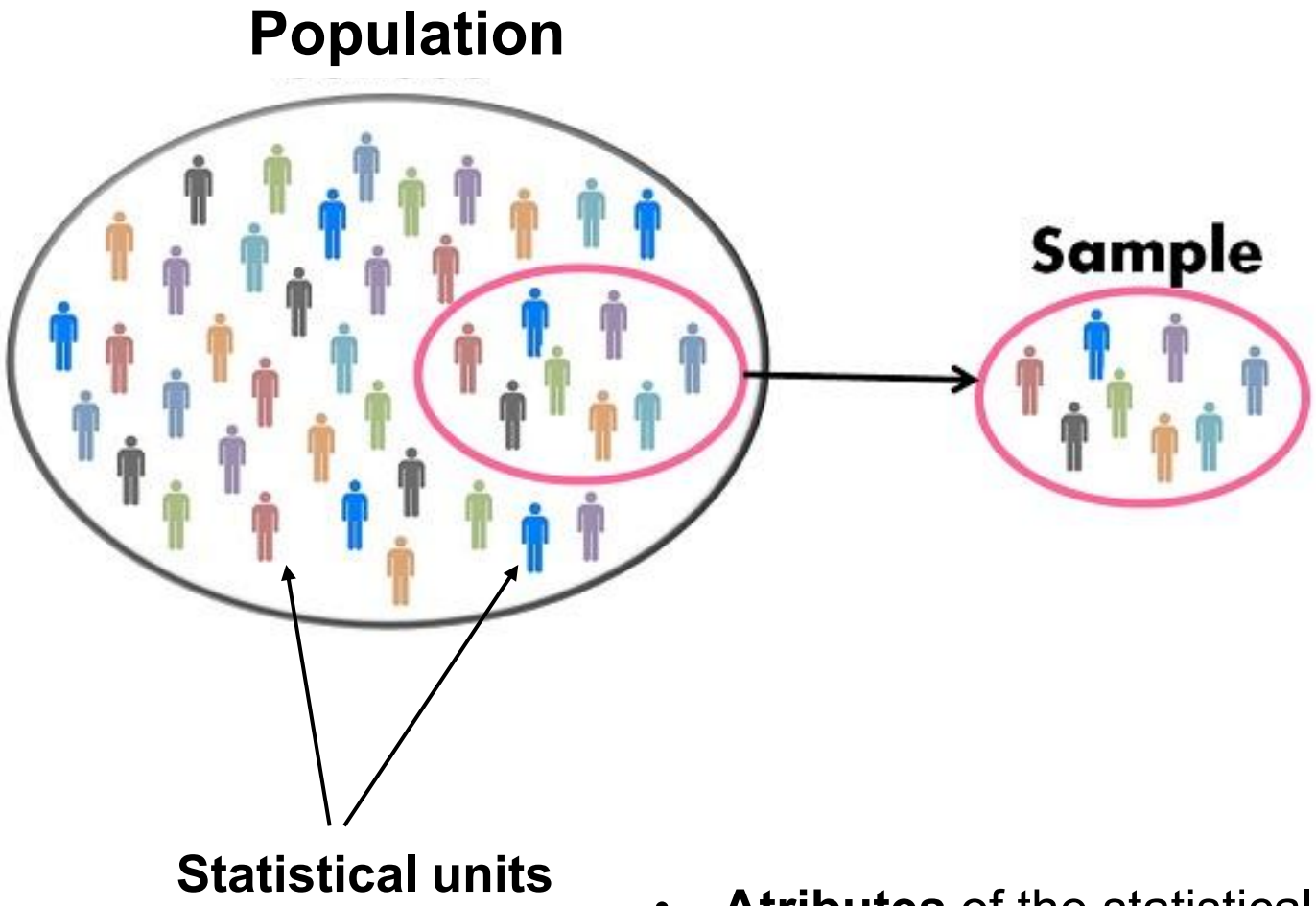
Some of the most commonly used styles:

- **APA**. APA is an author/date based style.
- **Harvard, Chicago** are similar to APA and are widely used in economics.

Some examples:

Type of source	In-text citation	Reference list
Journal article	"Melo and Ramli (2014) estimated an elasticity of …" They obtained a very small elasticity (e.g. Melo and Ramli)	Melo, P.C. & Ramli, A.R. 2014. Estimating Fuel Demand Elasticities to Evaluate CO_2 Emissions: Panel Data Evidence for the Lisbon Metropolitan Area. *Transportation Research Part A: Policy and Practice* Vol.67, pp. 30-46
Book section	"Melo and Graham (2012) estimated a regression model…"	Melo, P.C., Graham, D.J. & Noland, R.B. 2012. A Meta-Analysis of Estimates of Urban Agglomeration Economies, In: Mulley, Corinne (ed) *Urban Form and Transport Accessibility*, Volume 3 of Classics in Transport and Environmental Valuation Series, Cheltenham: Edward Elgar Publishing Ltd, Chapter 24, pp. 449-459

Basic notions (I)



- **Atributes** of the statistical units = Variables

e.g. gender, age, income, occupation

Basic notions (II)

- **Population** – the full set of statistical units (or universe) – e.g. individuals, firms, etc.
- **Sample** - sub-set of units drawn from the entire population or universe
- **Statistical unit** – each individual member of the population
- **Variables**: the features or attributes of the statistical units we want to study – e.g. age, income, etc.

Basic notions (III)

- Examples of **statistical units**: firms, farms, households, children, products, etc.

- Examples of **attributes of statistical units**: sales, age, gender, hair colour, eye colour, income, employment status, etc.

- The attributes of the statistical units - i.e. “variables” – can be **qualitative** or **quantitative**

Types of variables (I)

- **Qualitative** – also called categorical variables because they express classes or categories of non-numerical attributes
  - **Nominal variable** – expresses non-ordered categories which allow us to identify, classify, distinguish (e.g. gender, ethnicity, hair colour)
  - **Ordinal variable** – expresses categories that can be ordered (e.g. educational level; “low” to “high”; “fully disagree” to “fully agree”)

- **Quantitative** – numerical attributes, represent intensities
  - **Discrete variable** – takes on a finite (countable) number of values (e.g. age, number of firms, number of accidents, etc.)
  - **Continuous variable** – can take on any infinite value inside an interval (e.g. income, rent, profit, revenue, distance)

Types of variables (II)

- **NB! We can represent qualitative variables numerically:**
  - For **qualitative nominal variables**: the numbers given to each class do not have any numerical or quantitative meaning.
    - ✓ Yes/No: yes = 0, no = 1
    - ✓ Gender: male = 0, female =1
    - ✓ Employment status: employed = 1, unemployed = 0
  - For **qualitative ordinal variables**: the order of the numbers must respect the order of the categories.
    - ✓ Rating or Likert-type scales: e.g. five-level scale
      1: Strongly disagree
      2: Disagree
      3: Neither agree nor disagree
      4: Agree
      5: Strongly agree
    - ✓ Educational level: None =1, Primary = 2, Secondary=3, Tertiary=4

Types of variables (III) – Examples

- Q1: Do you enjoy food in ISEG's canteen?

Yes	No

- Q2: How much do you agree with the following statement: "The food in ISEG's canteen is great!"

Fully disagree	Disagree	No opinion	Agree	Fully agree

- Q3: How many people do you share your flat with?
- Q4: What is your monthly rent in Euros?
- Q5: Select the interval that best describes your room rent per month (in euros)?
  - ✓ Less than 150
  - ✓ (150-250(
  - ✓ (250-350(
  - ✓ 350 or more

Simple frequency tables (I)

- Suppose the values observed for a given variable X for statistical units i (i=1,…, n) are represented by:

$$x_1, x_2, x_3, \ldots, x_n$$

- **Frequency distribution:** set of observed values and their respective frequencies

- **Absolute frequency:** F_j is the number of times (i.e. count) each value of the variable X is observed.

- **Relative frequency:** f_j is the proportion of times each value of variable *X* is observed. N is the total number of observations.

$$f_j = \frac{F_j}{N}$$

- NB! *i* denotes the statistical units, while *j* denotes the set of observed values, which may or may not be the same number (if there are repeated values)

Example: Age of DAEB students (hypothetical)

- **Sample**: 50 students
- **Statistical unit**: student
- **Variable of interest**: age
- **Type of variable**: quantitative discrete

Table of simple frequencies

Age	Nr. students (Fj)	Share students (fj)
18	25	0.50
19	10	0.20
20	8	0.16
21	5	0.10
22	2	0.04
Total	**50**	**1.00**

Student No	Age	Student No	Age
1	18	26	20
2	19	27	20
3	18	28	20
4	19	29	18
5	18	30	18
6	19	31	18
7	18	32	18
8	19	33	19
9	18	34	19
10	18	35	18
11	18	36	18
12	20	37	18
13	18	38	18
14	18	39	19
15	19	40	18
16	19	41	18
17	22	42	18
18	22	43	18
19	21	44	21
20	18	45	20
21	18	46	20
22	18	47	20
23	18	48	21
24	19	49	20
25	21	50	21

Cummulative frequency tables (I)

- **Cumulative absolute frequency:** number of observations with values lower or equal than X_j

$$cum\ F_j = \sum_j F_j = N$$

- **Cumulative relative frequency:** proportion of observations with values lower or equal to X_j

$$cum\ f_j = \sum_j f_j = \sum_j \frac{F_j}{N} = 1$$

- Cumulative frequencies make sense for qualitative ordinal and quantitative variables, but not for qualitative nominal variables because they cannot be ordered.

Example: Age of DAEB students (hypothetical)

Table of simple and cumulative frequencies

Age	Nr. students (Fj)	Share students (fj)	cum Fj	cum fj
18	25	0.50	25	0.50
19	10	0.20	35	0.70
20	8	0.16	43	0.86
21	5	0.10	48	0.96
22	2	0.04	50	1.00
Total	**50**	**1.00**	**-**	**-**

$$cum\ F_j = \sum_j F_j = N$$

$$cum\ f_j = \sum_j f_j = \sum_j \frac{F_j}{N} = 1$$

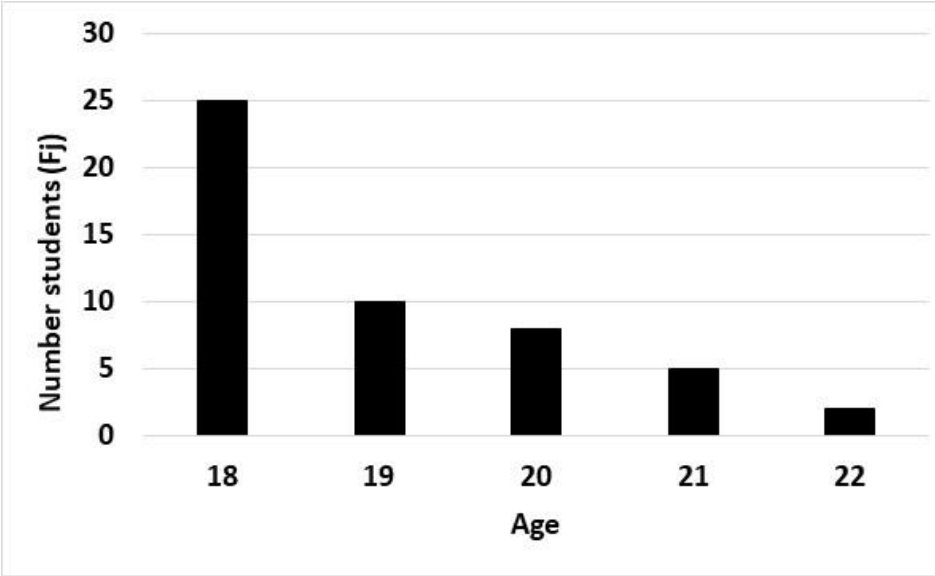
Frequency diagrams (I)

- **Quantitative discrete** variables:

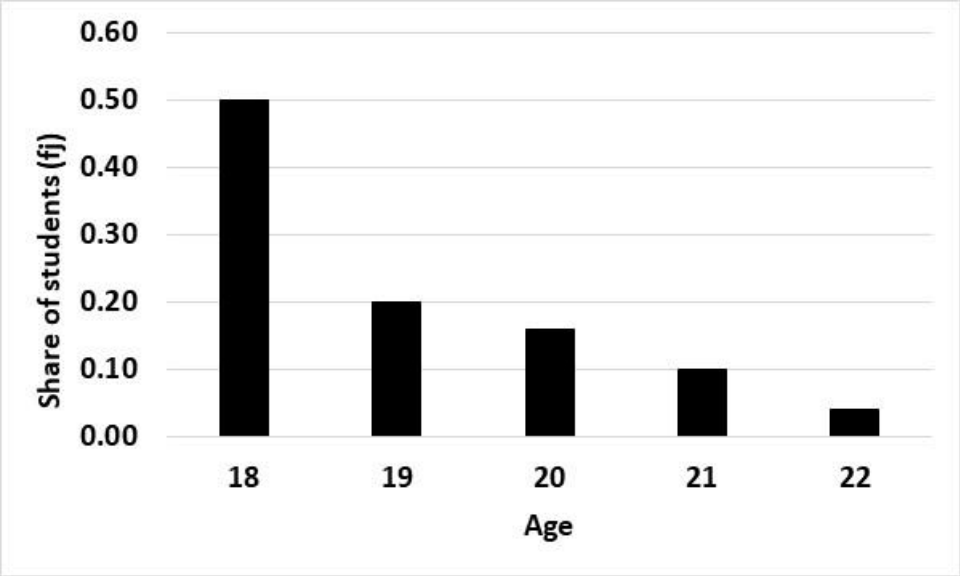
Bar charts: vertical axis shows the absolute or relative frequency for each value of the variable X in horizontal axis.

Example: Age of DAEB students (hypothetical)

Absolute frequency chart



Relative frequency chart



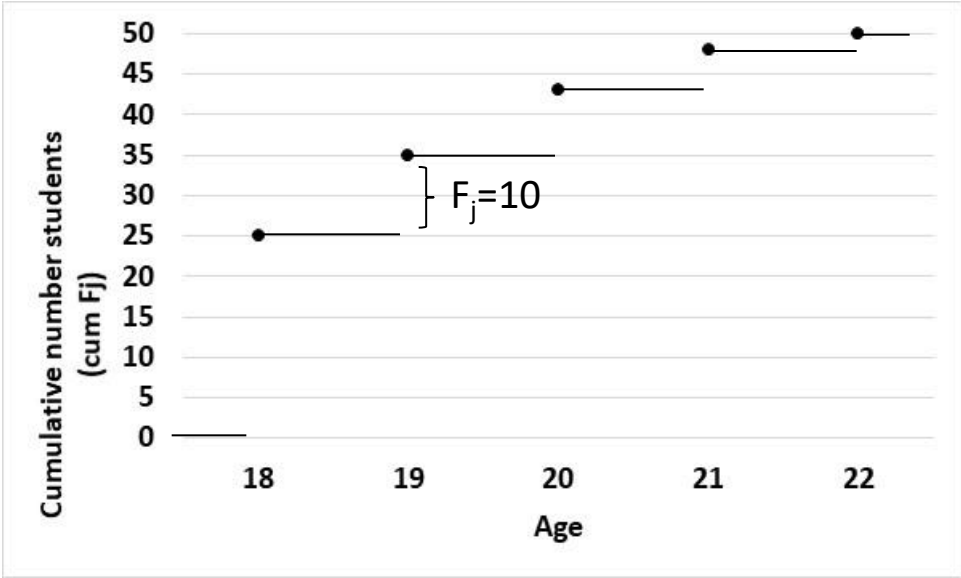
Frequency diagrams (II)

- **Quantitative discrete** variables:

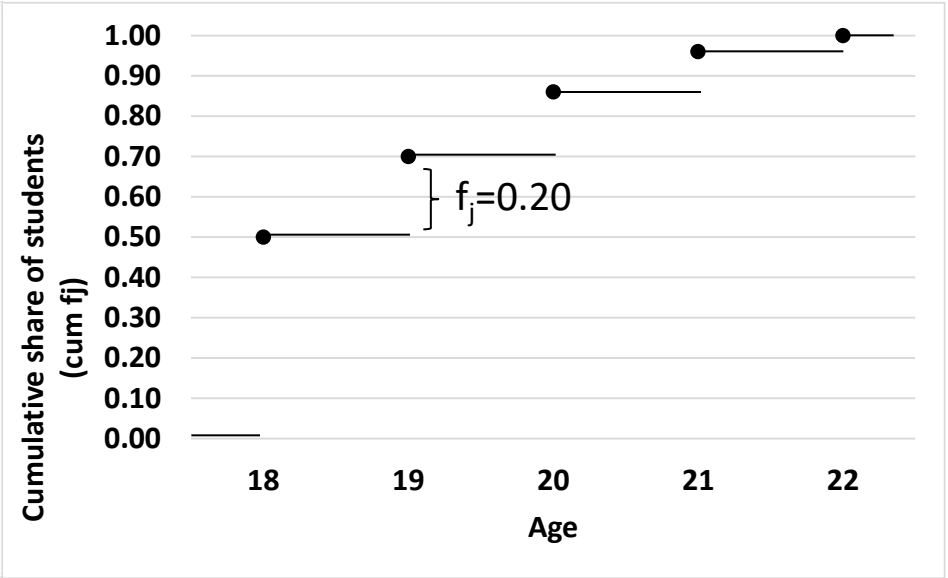
Step charts: vertical axis shows the cumulative absolute or relative frequency for each value of the variable X in horizontal axis.

Example: Age of DAEB students (hypothetical)

Cumulative absolute frequency chart



Cumulative relative frequency chart



Frequency tables for continuous variables (I)

- Continuous variables can take on any infinite value inside an interval (e.g. income, rent, profit, revenue, distance)

- As a consequence, it makes little sense to have simple or cumulative frequencies for individual isolated values

- Therefore, we often find **individual values grouped into *m* groups or classes or intervals l_j**

- The definition of class l_j should be such that:
  1. Classes should be mutually exclusive, i.e. non-intersection between classes
  2. Classes should be exhaustive – cover the entire range of values observed
  3. Number of classes: “not too many, nor too few”, no magic rule
  4. Class width or length a_j should be, when possible and appropriate, of equal size: $a_j = l_j - l_{j-1} = a, \qquad j = 1, 2, \ldots, m$

Frequency tables for continuous variables (II)

- Sometimes it is not appropriate to have equal-length classes, especially for variables with very dissimilar distributions (e.g. income, firm size, farm size)

- As a general rule you should avoid having indefinite / unlimited lower and upper limits in the first and last classes

- Group I_j is open to the left and closed to the right $I_j =]l_{j-1}, l_j]$, except for the first class (closed to the left) and last class (can be open to the right)

- Once you have created the groups or classes I_j, you can compute relative and absolute frequencies as defined earlier

- NB! You can also create groups or classes for discrete variables when there are too many values

Example: Average monthly earnings of individuals aged 25-30 (hypothetical)

- **Sample**: 35000 workers
- **Statistical unit**: worker
- **Variable of interest**: monthly earnings
- **Type of variable**: continuous (grouped)

Table of simple and cumulative frequencies

Income groups	Nr. Individuals (Fj)	Share individuals (fj)	cum Fj	cum fj
[250-450[	500	0.01	500	0.01
[450, 650[	5000	0.14	5500	0.16
[650, 850[	5000	0.14	10500	0.30
[850, 1050[	5000	0.14	15500	0.44
[1050,1250[	8000	0.23	23500	0.67
[1250,1450[	2500	0.07	26000	0.74
[1450,1650[	2500	0.07	28500	0.81
[1650,1850[	2500	0.07	31000	0.89
[1850,2050[	2000	0.06	33000	0.94
[2050,2250[	1250	0.04	34250	0.98
[2250,2450[	500	0.01	34750	0.99
[2450,2650]	250	0.01	35000	1.00
Total	**35000**	**1.00**	-	-

Data Analysis for Economics and Business

Lecture 2: Analysing numerical information

Basic notions; Frequency tables and diagrams
location measures (central tendency)

Structure of lecture

- Frequency tables and diagrams
- Define the different location measures of a distribution

- Calculate the different location measures of a distribution

Learning outcomes

- Draw absolute and relative frequency diagrams for different types of variables

Frequency diagrams

- **Discrete variables**:
  - Bar chart – for simple absolute and relative frequencies
  - Step chart – for cumulative absolute and relative frequencies

- **Continuous (grouped) variables**: First you need to group the values into m classes I_j with class width a_j
  - Histogram & frequency density polygon - for simple absolute and relative frequencies

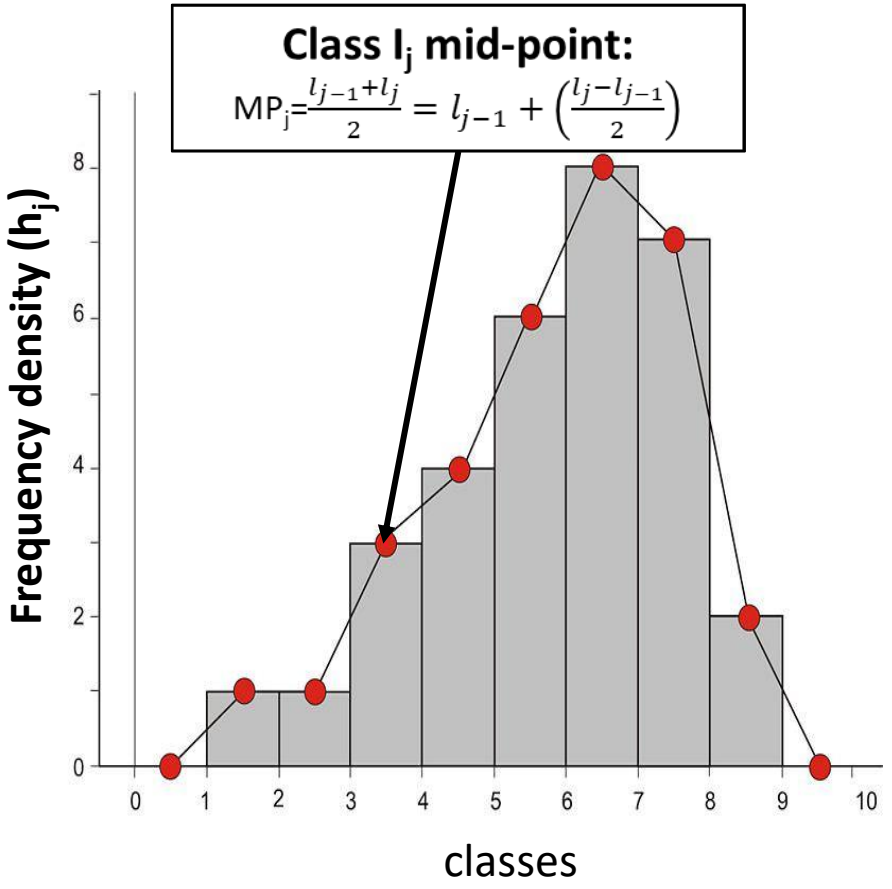
  NB! The height of the rectangles (h_j) differs for equal- vs. unequal-length classes:

  - Cumulative histogram & cumulative frequency polygon (ogive) - for cumulative absolute and relative frequencies

Frequency diagrams for continuous (grouped) variables (I)

Histogram (or area chart) and frequency density polygon:

- The histogram shows the **frequency density (h_j)** for each corresponding class length (a_j) of variable X
- **Frequency density = (frequency / class length)**
- The area of each rectangle (A_j) corresponds to the frequency: $A_j = h_j * a_j = (\frac{f_j}{a_j}) * a_j = f_j$
- Sum of A_j=$\sum_j (h_j * a_j) = \sum_j \left(\frac{f_j}{a_j}\right) * a_j = \sum_j f_j = 1$
- The **frequency density polygon** is obtained by joining the lines connecting the mid-points of each rectangle. It corresponds to the probability density function and its total area is also 1.



Frequency diagrams for continuous (grouped) variables (II)

Histogram (or area chart) and frequency density polygon:

- **If classes have the same length, $a_j=a_k=a$.** Because a_j is a constant we can set the frequency density (h_j) to be equal to the relative frequency f_j (or absolute frequency F_j). **It is a simple scale effect.**

  The area of each rectangle of the histogram is $A_j = (h_j * a) = (f_j * a)$ and $\sum_j \boldsymbol{A_j} = \sum_j (\boldsymbol{f_j} * \boldsymbol{a}) = \boldsymbol{a} \sum_j \boldsymbol{f_j} = \boldsymbol{a}$.

- **If classes have varying lengths ($a_j \neq a_k$)** the frequency density (h_j) must be used (as explained in previous slide)

Frequency diagrams - continuous (grouped) variables (III)

Histogram: equal-length groups

Average monthly earnings of individuals aged 25-30 (hypothetical)

Income groups	Nr. Individuals (Fj)	Share individuals (fj)	cum Fj	cum fj	aj	hj=Fj	hj=fj
[250-450[	500	0.01	500	0.01	200	500	0.014
[450, 650[	5000	0.14	5500	0.16	200	5000	0.143
[650, 850[	5000	0.14	10500	0.30	200	5000	0.143
[850, 1050[	5000	0.14	15500	0.44	200	5000	0.143
[1050,1250[	8000	0.23	23500	0.67	200	8000	0.229
[1250,1450[	2500	0.07	26000	0.74	200	2500	0.071
[1450,1650[	2500	0.07	28500	0.81	200	2500	0.071
[1650,1850[	2500	0.07	31000	0.89	200	2500	0.071
[1850,2050[	2000	0.06	33000	0.94	200	2000	0.057
[2050,2250[	1250	0.04	34250	0.98	200	1250	0.036
[2250,2450[	500	0.01	34750	0.99	200	500	0.014
[2450,2650]	250	0.01	35000	1.00	200	250	0.007
Total	**35000**	**1.00**	-	-	-	-	-

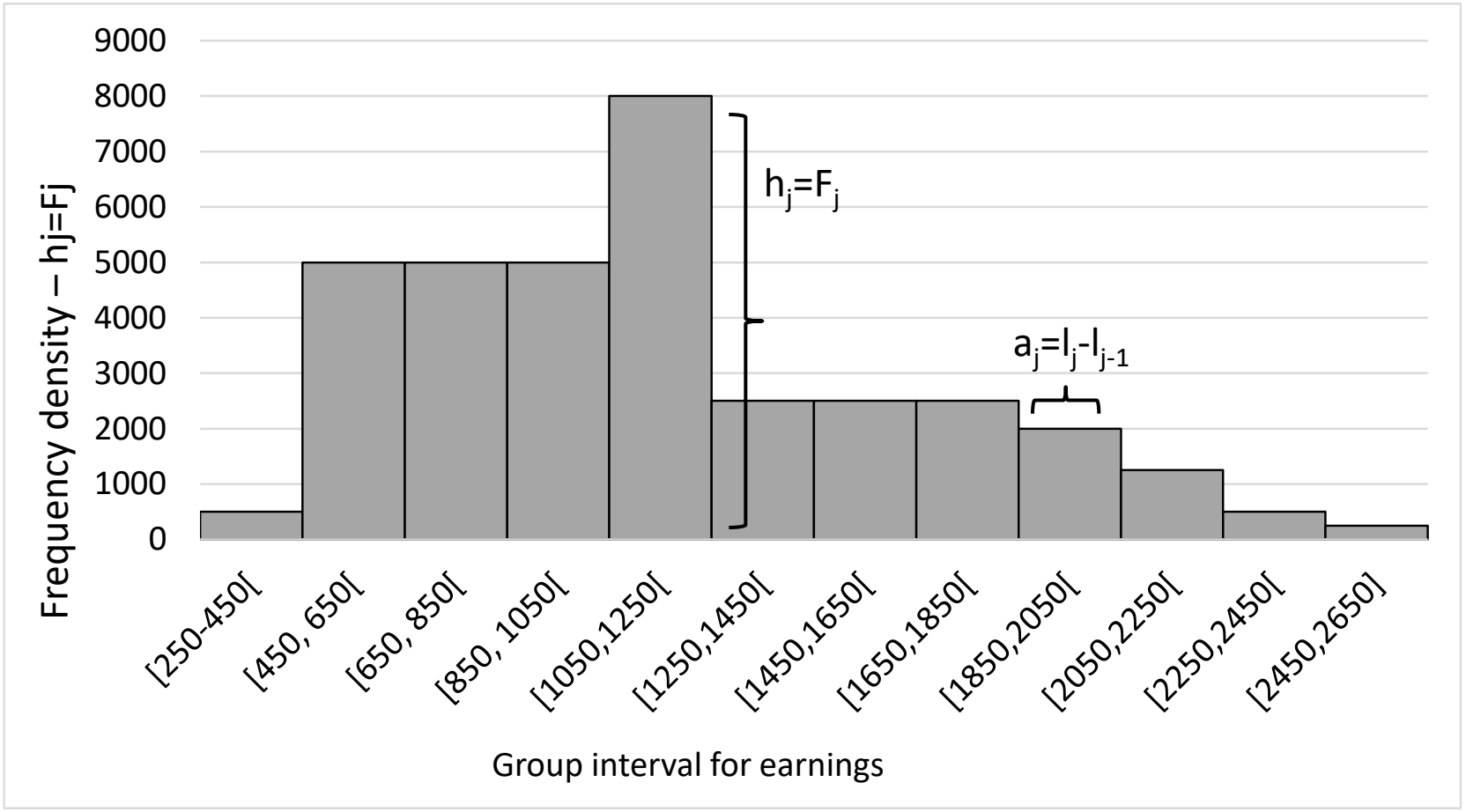
Group lengths

Frequency density (= simple frequency)

Frequency diagrams - continuous (grouped) variables (IV)

Histogram: equal-length groups

Average monthly earnings of individuals aged 25-30 (hypothetical)



Frequency diagrams - continuous (grouped) variables (V)

Histogram: unequal-length groups

Average monthly earnings of individuals aged 25-30 (hypothetical)

Income groups	Nr. Individuals (Fj)	Share individuals (fj)	cum Fj	cum fj	aj	hj=Fj/aj	hj=fj/aj
[250-450[	500	0.01	500	0.01	200	2.5	0.000071
[450, 850[	10000	0.29	10500	0.30	400	25	0.000714
[850, 1850[	20500	0.59	31000	0.89	1000	20.5	0.000586
[1850,2250[	3250	0.09	34250	0.98	400	8.125	0.000232
[2250,2450[	500	0.01	34750	0.99	200	2.5	0.000071
[2450,2650]	250	0.01	35000	1.00	200	1.25	0.000036
Total	35000	1.00	-	-	-	-	-

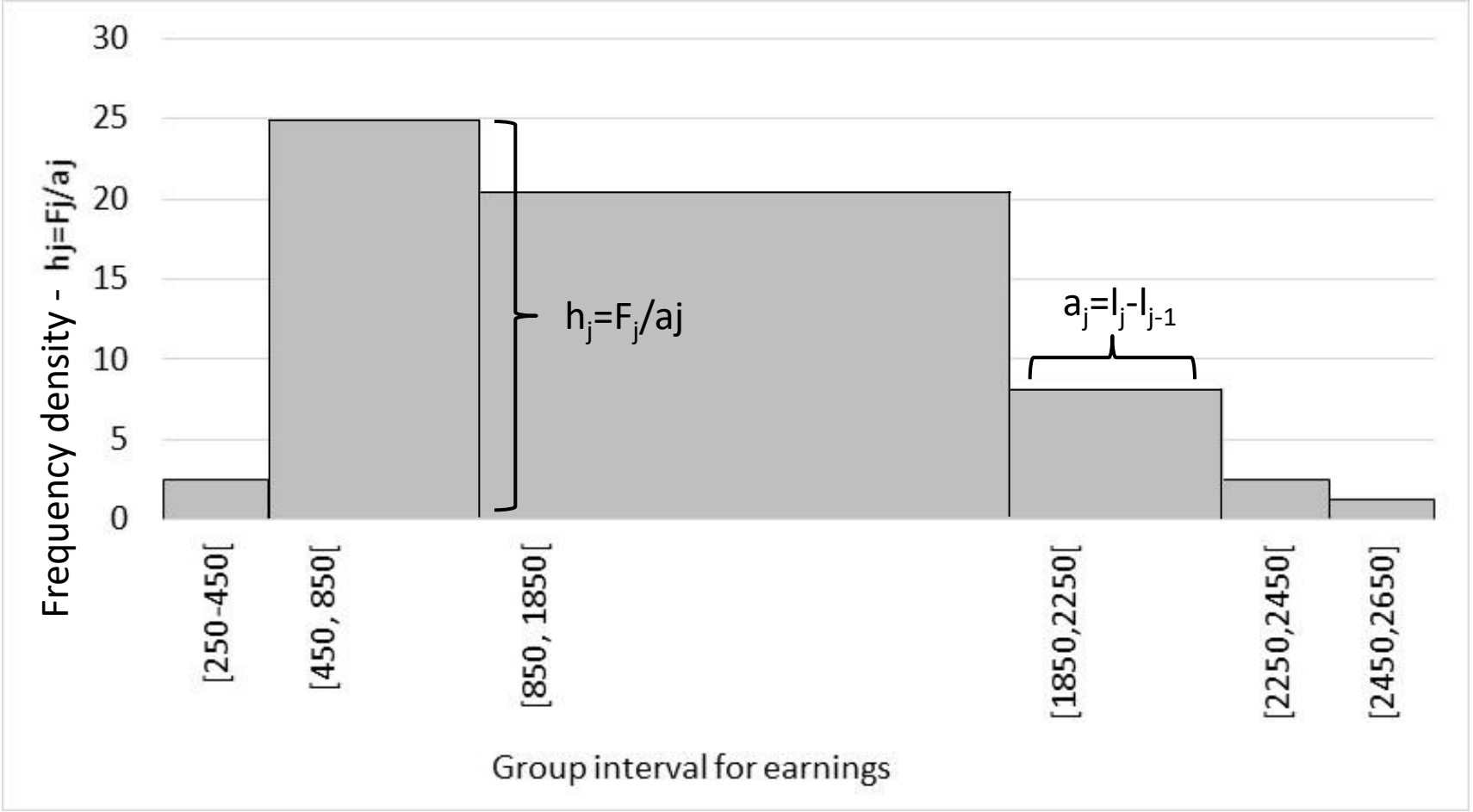
Group lengths

Frequency density = frequency divided by class length

Frequency diagrams - continuous (grouped) variables (VI)

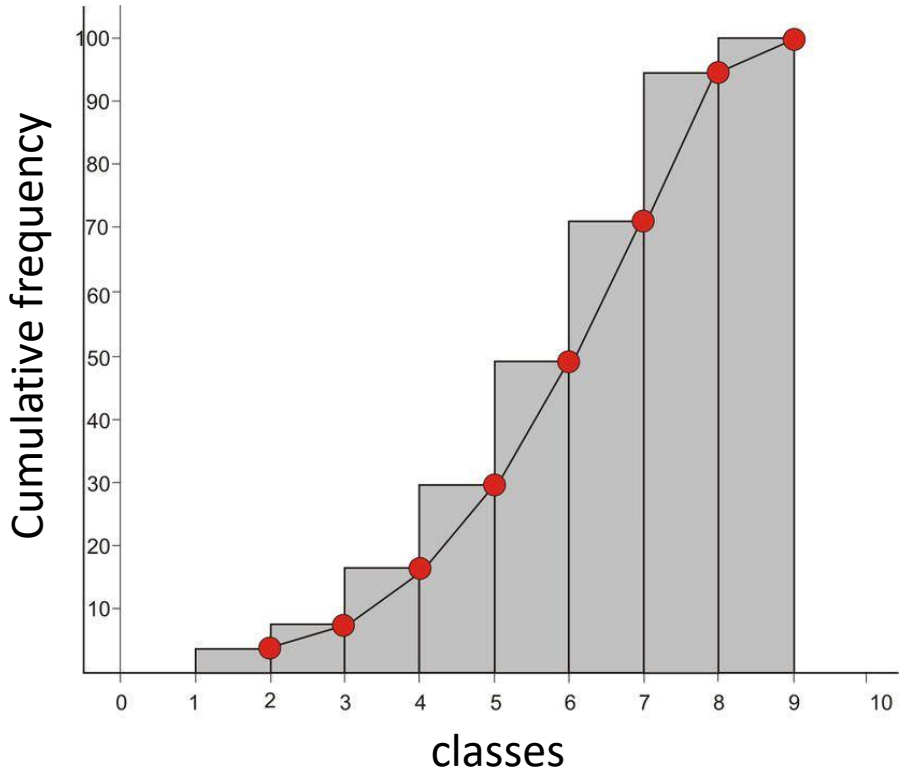
Histogram: unequal-length groups

Average monthly earnings of individuals aged 25-30 (hypothetical)



Frequency diagrams - continuous (grouped) variables (VII)

- Cumulative frequency polygon (also called ogive) is plotted by joining the line segments connecting each class upper limit



Cumulative frequency diagrams

- Differences and similarities between cumulative frequency diagrams for quantitative discrete and continuous variables:

Similarities:

1. $0 \leq \text{cum } f(x) \leq 1$
2. cum f(x) is a non-decreasing function

(same applies to cum F(x))

Differences refer mainly to the discontinuity of cum f(x) for values x_j in the case of discrete variables, between which cum f(x) is constant, giving it the apparent shape of a “step” line

Location Measures

- Location or position measures:
  - **Central tendency measures** – the centre of the distribution
    - Mean
    - Median
    - Mode
  - **Non-central tendency measures** – a given part of the distribution
    - Quartiles
    - Deciles
    - Percentiles

Location Measures - Central tendency

- The **mean** – gravity centre of the distribution
  - Only for quantitative data
  - Makes use of all available values
  - The value may not be observed
  - Influenced by extreme values (outliers)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{x} = \frac{1}{n}\sum_{j=1}^{m} F_j x_j = \sum_{j=1}^{m} f_j x_j$$

- Calculation of mean for grouped data is based on the weighted average of each class mid-points

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{m} F_j MP_j = \sum_{j=1}^{m} f_j MP_j$$

By assuming that all values within a given class are equal to its mid-point we incur a "*tabulation error*"

Location Measures - Central tendency

Example: Age of students in AIEE (hypothetical)

Age (xj)	Nr. students (Fj)	Share students (fj)	cum Fj	cum fj	fj*xj
18	25	0,50	25	0,50	9,0
19	10	0,20	35	0,70	3,8
20	8	0,16	43	0,86	3,2
21	5	0,10	48	0,96	2,1
22	2	0,04	50	1,00	0,9
Total	**50**	**1,00**	-	-	19,0

$$\bar{x} = \sum_{j=1}^{m} f_j x_j$$

Location Measures - Central tendency

Example: Average monthly earnings of individuals aged 25-30 (hypothetical)

Income groups	Nr. Individuals (Fj)	Share individuals (fj)	cum Fj	cum fj	aj	MPj	fj*MPj
[250-450[	500	0,0143	500	0,0143	200	350	5,0
[450, 650[	5000	0,1429	5500	0,1571	200	550	78,6
[650, 850[	5000	0,1429	10500	0,3000	200	750	107,1
[850, 1050[	5000	0,1429	15500	0,4429	200	950	135,7
[1050,1250[	8000	0,2286	23500	0,6714	200	1150	262,9
[1250,1450[	2500	0,0714	26000	0,7429	200	1350	96,4
[1450,1650[	2500	0,0714	28500	0,8143	200	1550	110,7
[1650,1850[	2500	0,0714	31000	0,8857	200	1750	125,0
[1850,2050[	2000	0,0571	33000	0,9429	200	1950	111,4
[2050,2250[	1250	0,0357	34250	0,9786	200	2150	76,8
[2250,2450[	500	0,0143	34750	0,9929	200	2350	33,6
[2450,2650]	250	0,0071	35000	1,0000	200	2550	18,2
Total	**35000**	**1,00**	-	-	-	-	**1161,4**

$$\bar{x} = \sum_{j=1}^{m} f_j MP_j$$

Location Measures - Central tendency

Example: Average monthly earnings of individuals aged 25-30 (hypothetical) – unequal class width

Income groups	Nr. Individuals (Fj)	Share individuals (fj)	cum Fj	cum fj	aj	MPj	fj*MPj
[250-450[	500	0,014	500	0,014	200	350	5,0
[450, 850[	10000	0,286	10500	0,300	400	650	185,7
[850, 1850[	20500	0,586	31000	0,886	1000	1350	790,7
[1850,2250[	3250	0,093	34250	0,979	400	2050	190,4
[2250,2450[	500	0,014	34750	0,993	200	2350	33,6
[2450,2650]	250	0,007	35000	1,000	200	2550	18,2
Total	35000	1,00	-	-	-	-	**1223,6**

$$\bar{x} = \sum_{j=1}^{m} f_j MP_j$$

• The grouping of the continuous variable into fewer (greater width) classes incurs tabulation error by producing a different mean value for the monthly earnings

Location Measures - Central tendency

- **Properties of the mean**:
  - Adding / subtracting a non-zero constant C to the distribution leads to an increase / reduction in the mean equal to constant C

$$\boldsymbol{m(x+c) = m(x) + m(C) = m(x) + C}$$

  - Multiplying / diving the distribution by a non-zero constant C leads to the multiplication / division of the mean by constant C

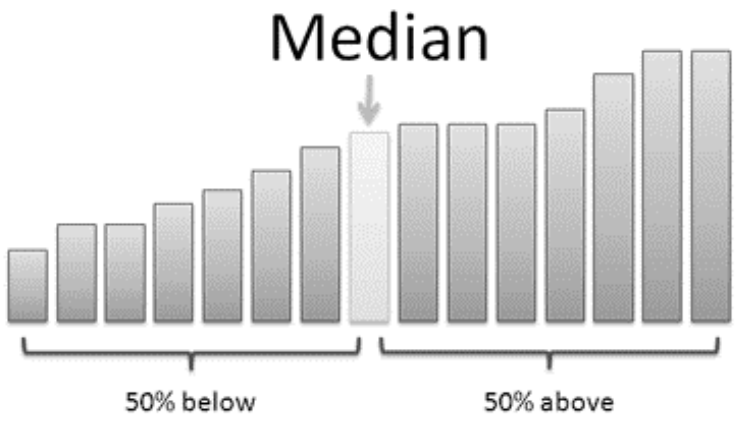
$$\boldsymbol{m(x * c) = c * m(x)}$$

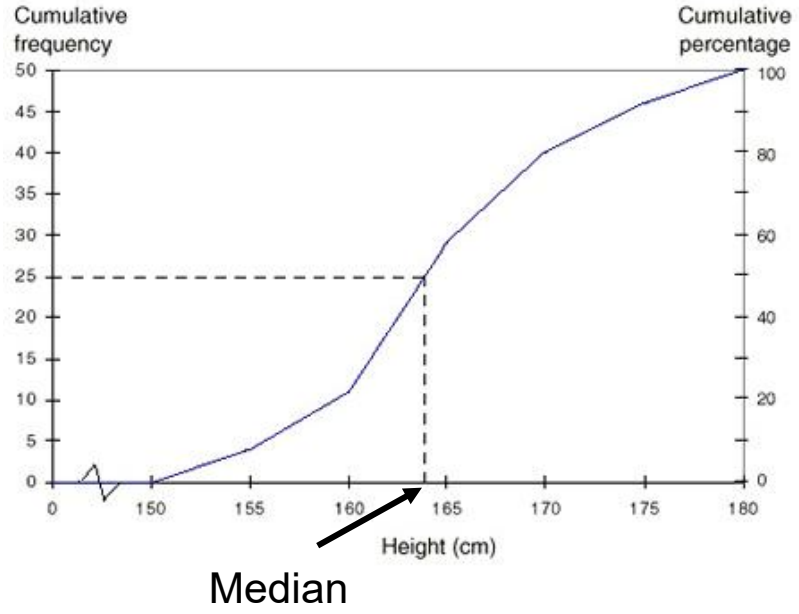
  - The global mean of the distribution is equal the mean of the group means

$$\bar{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2} + \cdots + n_k\overline{x_k}}{n_1 + n_2 + \cdots + n_k} = \frac{1}{n}\sum_{j=1}^{m} n_j\overline{x_j}$$

Location Measures - Central tendency

- **Median** - positional centre of the distribution – **central observation** (i.e. half or **50% of the observations above and below**)
  - Quantitative or ordinal qualitative data
  - Not influenced by extreme values (outliers)
  - More difficult to compute than the mean
  - Makes use of only a few observations





Location Measures - Central tendency

- **Median** for discrete data:
  - Uneven number of observations: central observation

$$x_{Me} = x_{\frac{n+1}{2}}$$

  - Even number of observations: mean of central observations

$$x_{Me} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

Example: Age of students in AIEE (hypothetical)

Age (xj)	Nr. students (Fj)	Share students (fj)	cum Fj	cum fj	fj*xj
18	25	0,50	25	0,50	9,0
19	10	0,20	35	0,70	3,8
20	8	0,16	43	0,86	3,2
21	5	0,10	48	0,96	2,1
22	2	0,04	50	1,00	0,9
Total	**50**	**1,00**	-	-	19,0

N=50 is an even number, so: $x_{Me} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_{25} + x_{25+1}}{2} = \frac{18+19}{2} = 18.5$

Location Measures - Central tendency

- **Median** for continuous grouped data:
  1. Calculate the cumulative frequency distribution
  2. Find the class that contains the median, i.e. cum $F_j \geq N/2$ or cum $f_j \geq 0.5$
  3. Find the median by using the formula:

$$x_{Me} = l_{j-1}(Me) + \frac{0.5 \ - cum\ f(Me-1)}{f(Me)} a(Me)$$

where:

- l_{j-1} (Me): lower limit of the median class
- cum f(Me-1): cumulative frequency of the class before the median class
- f(Me): relative frequency of the median class
- a(Me): width of the median class

Location Measures - Central tendency

Example: Average monthly earnings of individuals aged 25-30 (hypothetical)

Income groups	Nr. Individuals (Fj)	Share individuals (fj)	cum Fj	cum fj
[250-450[	500	0,0143	500	0,0143
[450, 650[	5000	0,1429	5500	0,1571
[650, 850[	5000	0,1429	10500	0,3000
[850, 1050[	5000	0,1429	15500	0,4429
[1050,1250[	8000	0,2286	23500	0,6714
[1250,1450[	2500	0,0714	26000	0,7429
[1450,1650[	2500	0,0714	28500	0,8143
[1650,1850[	2500	0,0714	31000	0,8857
[1850,2050[	2000	0,0571	33000	0,9429
[2050,2250[	1250	0,0357	34250	0,9786
[2250,2450[	500	0,0143	34750	0,9929
[2450,2650]	250	0,0071	35000	1,0000
Total	**35000**	**1,00**	-	-

cum f(Me-1)

cum f(Me)≥0.50

$l_{j=Me}$ is the class that contains the median, i.e. median class

$$x_{Me} = l_{j-1}(Me) + \frac{0.5 - cum\, f(Me-1)}{f(Me)} a(Me) = 1050 + \frac{0.5 - 0.443}{0.229} * 200 = 1102$$

Location measures - Central tendency

- **Mode** – the most frequent value, i.e. **value with the highest frequency**
  - Quantitative, qualitative ordinal and qualitative nominal data
  - Not affected by outliers
  - May assume more than one value or be undefined

- Mode for grouped data - it belongs to the **class with the highest frequency, i.e. the modal class**. To obtain the mode apply the formula:

$$x_{Mo} = l_{j-1}(Mo) + \frac{f(Mo+1)}{f(Mo-1) + f(Mo+1)}\, a(Mo)$$

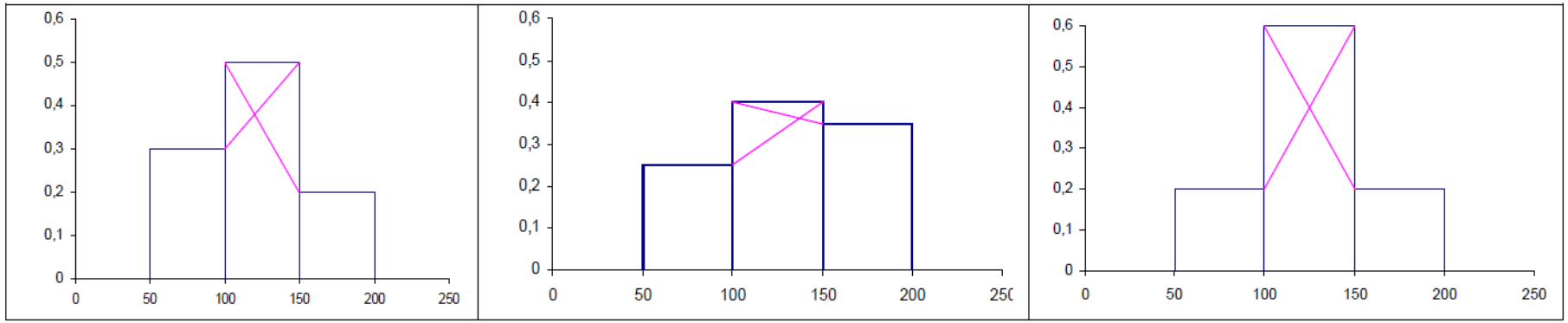
Location measures - Central tendency

- Identifying the mode from histograms and relative frequencies

Table of relative frequencies (%) for variables A, B, and C

Classe	A	B	C
0-50	0	0	0
50-100	30	25	20
100-150	50	40	60
150-200	20	35	20
200-250	0	0	0

Histograms for variables A, B, and C



- The **mode is closer to the adjacent class with larger relative frequency**

Location measures - Central tendency

Example: Age of students in AIEE (hypothetical)

Age (xj)	Nr. students (Fj)	Share students (fj)	cum Fj	cum fj
18	25	0,50	25	0,50
19	10	0,20	35	0,70
20	8	0,16	43	0,86
21	5	0,10	48	0,96
22	2	0,04	50	1,00
Total	**50**	**1,00**	-	-

- The mode is 18 because it is the value with the highest absolute / relative frequency

Location measures - Central tendency

Example: Average monthly earnings of individuals aged 25-30 (hypothetical)

Income groups	Nr. Individuals (Fj)	Share individuals (fj)	cum Fj	cum fj	
[250-450[	500	0,014	500	0,014	← f(Mo-1)
[450, 850[	10000	0,286	10500	0,300	← f(Mo)
[850, 1850[	20500	0,586	31000	0,886	← f(Mo+1)
[1850,2250[	3250	0,093	34250	0,979	
[2250,2450[	500	0,014	34750	0,993	
[2450,2650]	250	0,007	35000	1,000	
Total	35000	1,00	-	-	

$l_{j=Mo}$ is the class that contains the mode, i.e. modal class

$$x_{Mo} = l_{j-1}(Mo) + \frac{f(Mo+1)}{f(Mo-1)+f(Mo+1)}\,a(Mo)$$

Location measures - Central tendency

Summary table:

	Mean	Median	Mode
Characterization	Gravity centre	Central position value	Most frequent
Application	Quantitative data	Quantitative or ordinal qualitative data	All
Characteristics	All available values Easy to compute Influenced by extreme values Need to define class limits	Determined by the order and not by the value of the observations Hard to compute Not influenced by extreme values No need to define class limits	May exist more than one or not be defined Not influenced by extreme values No need to define class limits

Data Analysis for Economics and Business

Lectures 8 and 9: Analyzing numerical information

Data reduction: Location measures (non-central tendency);
Asymmetry and skewness profile of distributions
Measures of dispersion

Academic Year 2023/24

Structure of lecture

- Location measures – Non-central tendency
- Asymmetry and skewness profile of numerical distributions
- Measures of dispersion
  - Measures of absolute dispersion
  - Measures of relative dispersion

Learning outcomes

- Define the different location measures of a distribution
- Calculate the different location measures of a distribution
- Classify distributions according to asymmetry/skewness profile
- Define and explain the different measures of dispersion
- Calculate the different measures of dispersion

Location Measures

- Location or position measures:
  - **Central tendency measures** – the centre of the distribution
    - Mean
    - Median
    - Mode
  - **Non-central tendency measures** – a given part of the distribution
    - Quartiles
    - Deciles
    - Percentiles

Non-central tendency location measures

- **Quantiles**: separate the distribution into a set of equal size partitions
  - **Quartiles** – 4 equal parts – Q_1, Q_2, Q_3

$$Q_1 = Q_{0,25}: contains\ 25\%\ of\ the\ observations$$
$$Q_2 = Q_{0,50}: contains\ 50\%\ of\ the\ observations\ (= median)$$
$$Q_3 = Q_{0,75}: contains\ 75\%\ of\ the\ observations$$

  - **Deciles** – 10 equal parts – D_1, D_2,..., D_9

$$D_5 = Q_2: contains\ 50\%\ of\ the\ observations\ (= median)$$
$$D_9: contains\ 90\%\ of\ the\ observations$$

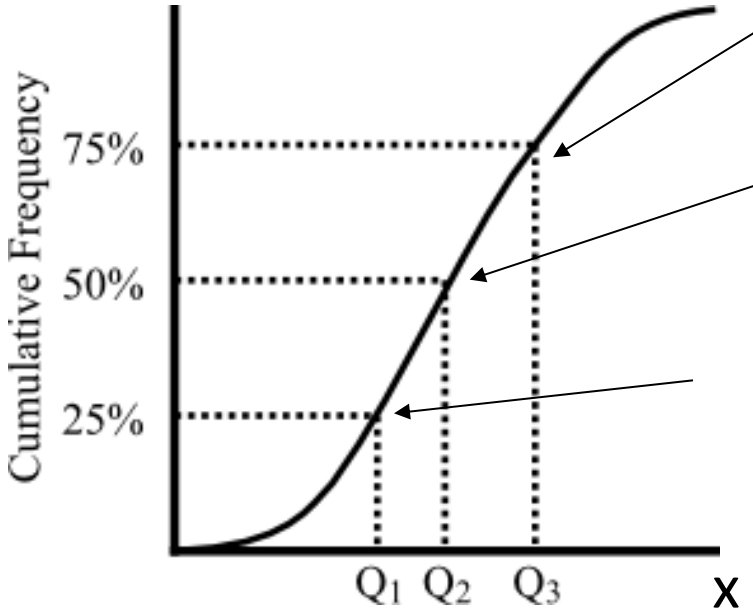
  - **Percentiles** – 100 equal parts – P_1, P_2,...,P_{99}

$$P_{25} = Q_1: contains\ 25\%\ of\ the\ observations$$
$$P_{75} = Q_3: contains\ 75\%\ of\ the\ observations$$

Quantiles are computed in the same way as the median both for discrete or grouped data

Non-central tendency location measures

Identifying quartiles from cumulative frequencies:



Q_3=$Q_{0,75}$:contains 75% of observations
x: cum f(x)=75%

Q_2 = $Q_{0,50}$:contains 50% of observations and is equal to the median
x: cum f(x)=50%

Q_1 = $Q_{0,25}$:contains 25% of observations
x: cum f(x)=25%

- Similar for deciles (10%, …, 90%) and percentiles (1%, …, 99%)

Non-central tendency location measures

- Quartiles: separate the distribution into 4 equal parts – Q_1, Q_2, Q_3

  - Q_1 contains 25% of observations

$$x_{Q_1} = l_{j-1}(Q_1) + \frac{0.25 - cum\, f(Q_1 - 1)}{f(Q_1)} a(Q_1)$$

  - Q_2 contains 50% of observations = median

$$x_{Q_2} = x_{Me} = l_{j-1}(Me) + \frac{0.50 - cum\, f(Me - 1)}{f(Me)} a(Me)$$

  - Q_3 contains 75% of observations

$$x_{Q_3} = l_{j-1}(Q_3) + \frac{0.75 - cum\, f(Q_3 - 1)}{f(Q_3)} a(Q_3)$$

Non-central tendency location measures

Example: Average monthly earnings of individuals aged 25-30 (hypothetical)

Income groups	Nr. Individuals (Fj)	Share individuals (fj)	cum Fj	cum fj	aj
[250-450[	500	0,01	500	0,01	200
[450, 650[	5000	0,14	5500	0,16	200
[650, 850[	5000	0,14	10500	0,30	200
[850, 1050[	5000	0,14	15500	0,44	200
[1050,1250[	8000	0,23	23500	0,67	200
[1250,1450[	2500	0,07	26000	0,74	200
[1450,1650[	2500	0,07	28500	0,81	200
[1650,1850[	2500	0,07	31000	0,89	200
[1850,2050[	2000	0,06	33000	0,94	200
[2050,2250[	1250	0,04	34250	0,98	200
[2250,2450[	500	0,01	34750	0,99	200
[2450,2650]	250	0,01	35000	1,00	200
Total	**35000**	**1,00**	-	-	-

cum $f(Q_1-1)$

cum $f(Q_1)\geq 0.25$

Class that contains Q_1

$f(Q_1)$

$$x_{Q_1} = l_{j-1}(Q_1) + \frac{0.25 - cum\, f(Q_1-1)}{f(Q_1)} a(Q_1) = 650 + \frac{0.25-0.16}{0.14} * 200 = 778.57$$

Non-central tendency location measures

Example: Average monthly earnings of individuals aged 25-30 (hypothetical)

Income groups	Nr. Individuals (Fj)	Share individuals (fj)	cum Fj	cum fj	aj
[250-450[	500	0,01	500	0,01	200
[450, 650[	5000	0,14	5500	0,16	200
[650, 850[	5000	0,14	10500	0,30	200
[850, 1050[	5000	0,14	15500	0,44	200
[1050,1250[	8000	0,23	23500	0,67	200
[1250,1450[	2500	0,07	26000	0,74	200
[1450,1650[	2500	0,07	28500	0,81	200
[1650,1850[	2500	0,07	31000	0,89	200
[1850,2050[	2000	0,06	33000	0,94	200
[2050,2250[	1250	0,04	34250	0,98	200
[2250,2450[	500	0,01	34750	0,99	200
[2450,2650]	250	0,01	35000	1,00	200
Total	**35000**	**1,00**	-	-	-

cum $f(Q_3-1)$

cum $f(Q_3)\geq 0.75$

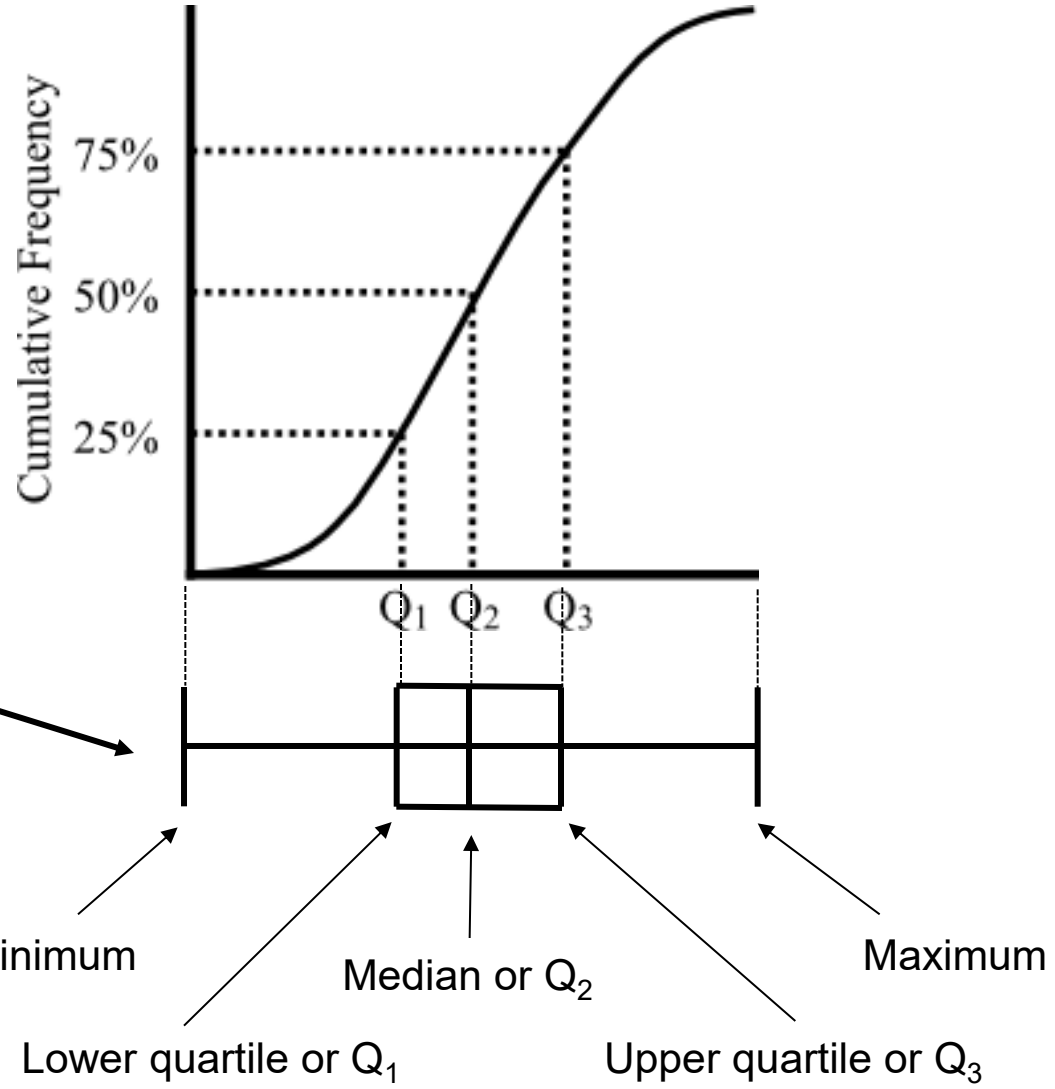
Class that contains Q_3

$f(Q_3)$

$$x_{Q_3} = l_{j-1}(Q_3) + \frac{0.75 - cum\, f(Q_3-1)}{f(Q_3)} a(Q_3) = 1450 + \frac{0.75-0.74}{0.07} * 200 = 1478.57$$

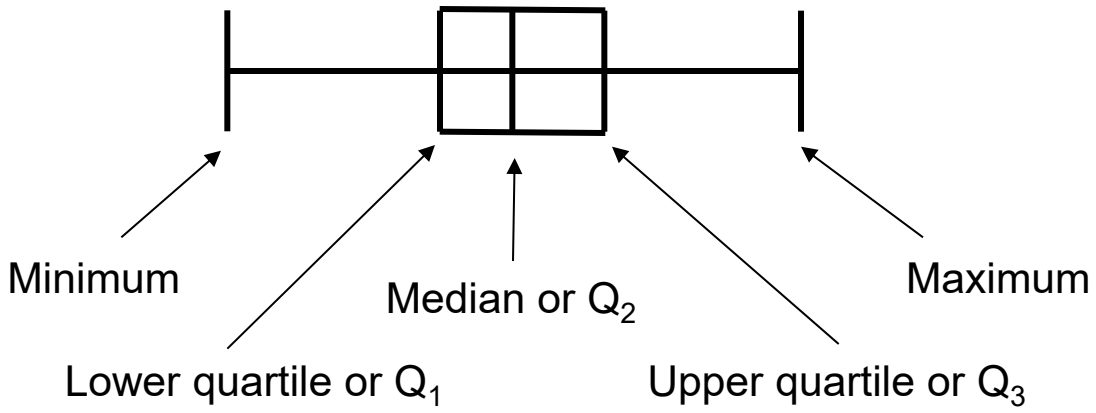
Location measures

Box-Whisker Diagram



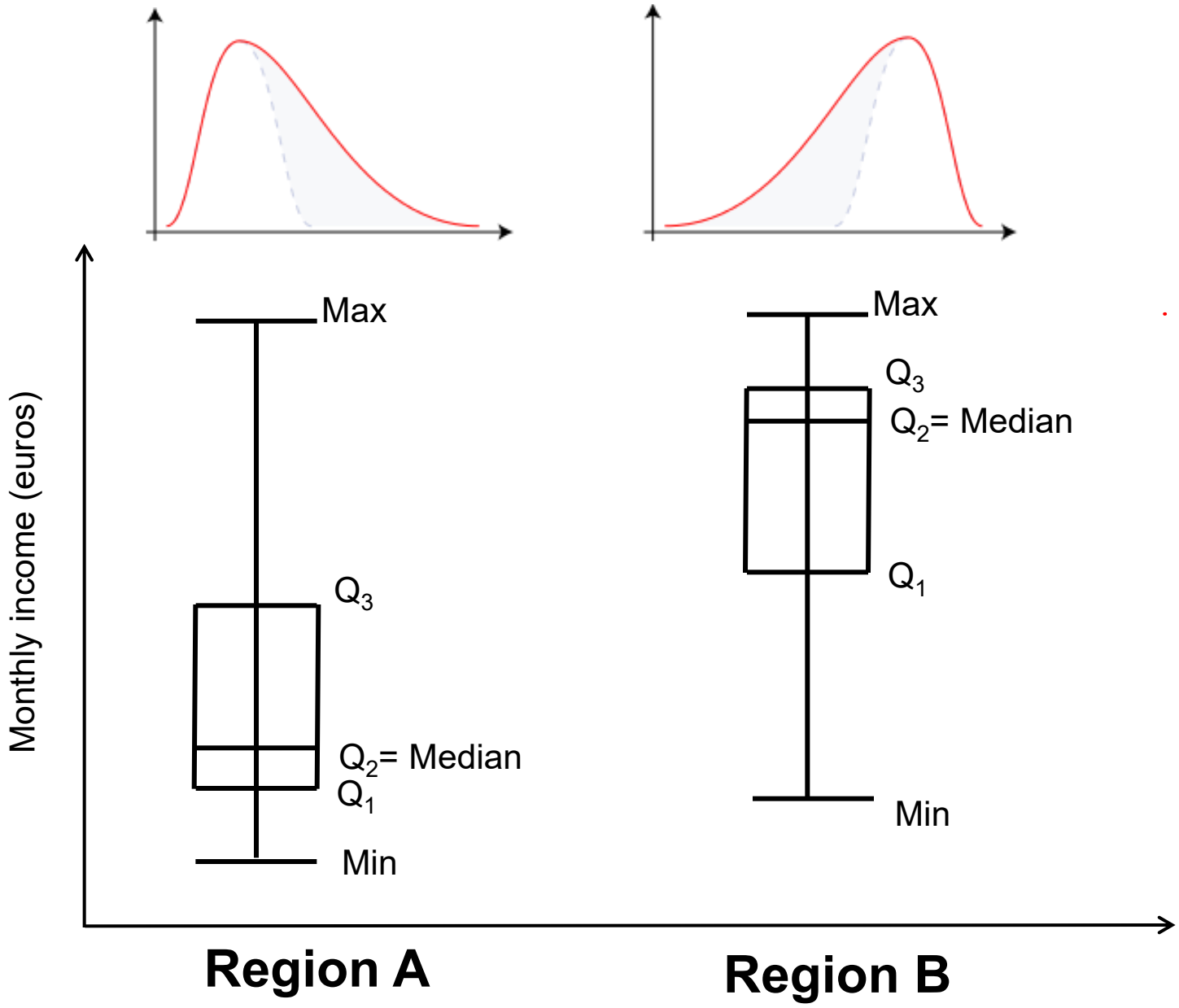
Box-Whisker Diagram

- Assessing location, dispersion and asymmetry using position measures

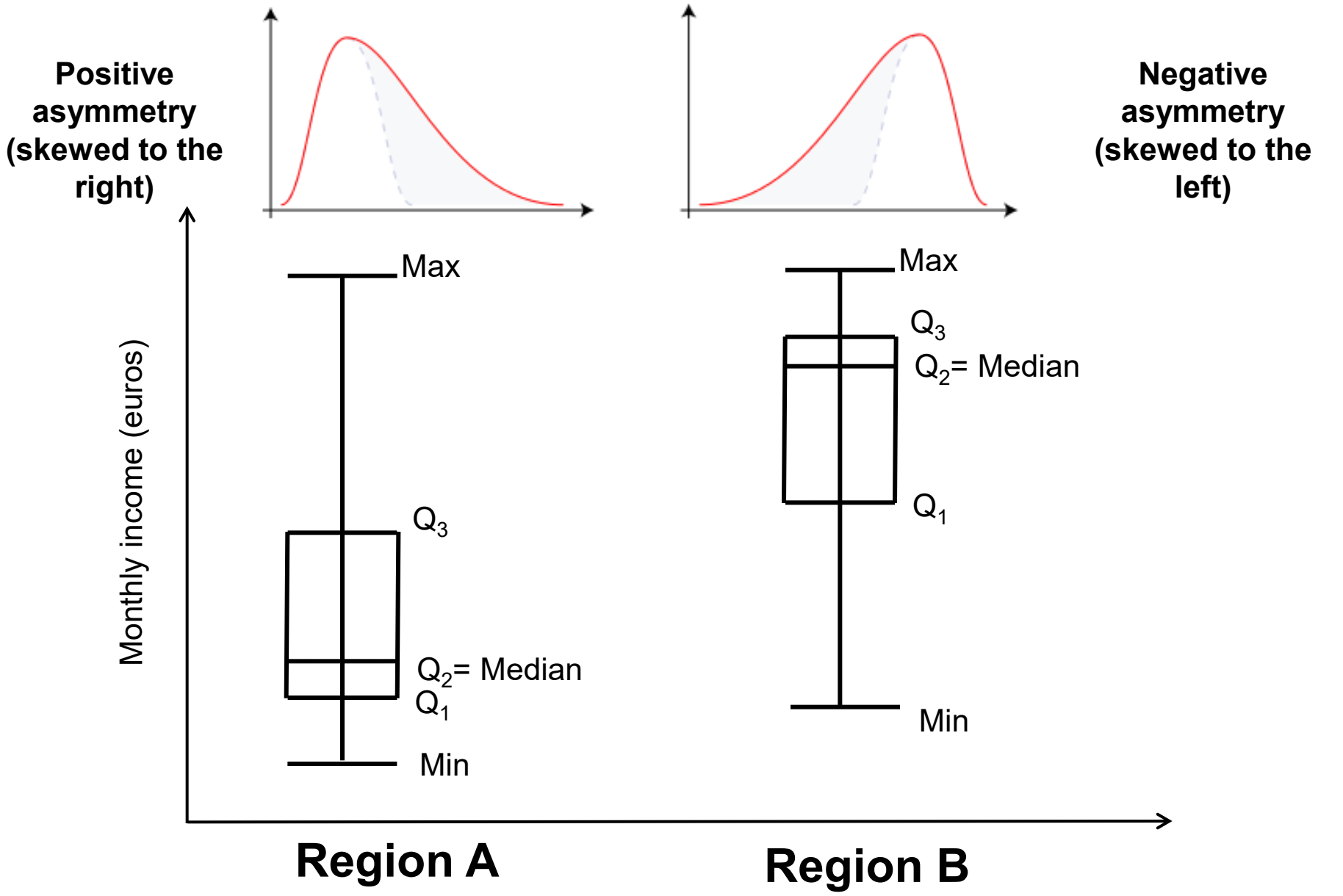


- **Classifying asymmetry profiles based on box-whisker diagrams**:
  - ❖ Positive asymmetry (skewed to the right): Median closer to Q1 and Q1 closer to Min
  - ❖ Negative asymmetry (skewed to the left): Median closer to Q3 and Q3 closer to Max
  - ❖ Symmetric: Median divides range and IQR in two equal parts; Q1, Q2, and Q3 divide range in four equal parts

Asymmetry or Skewness profiles

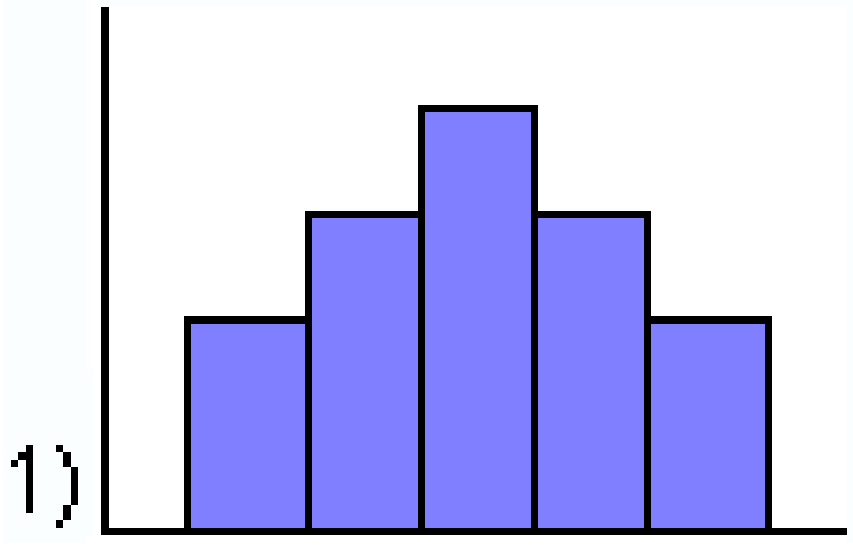


Asymmetry or Skewness profiles



Asymmetry or Skewness profiles

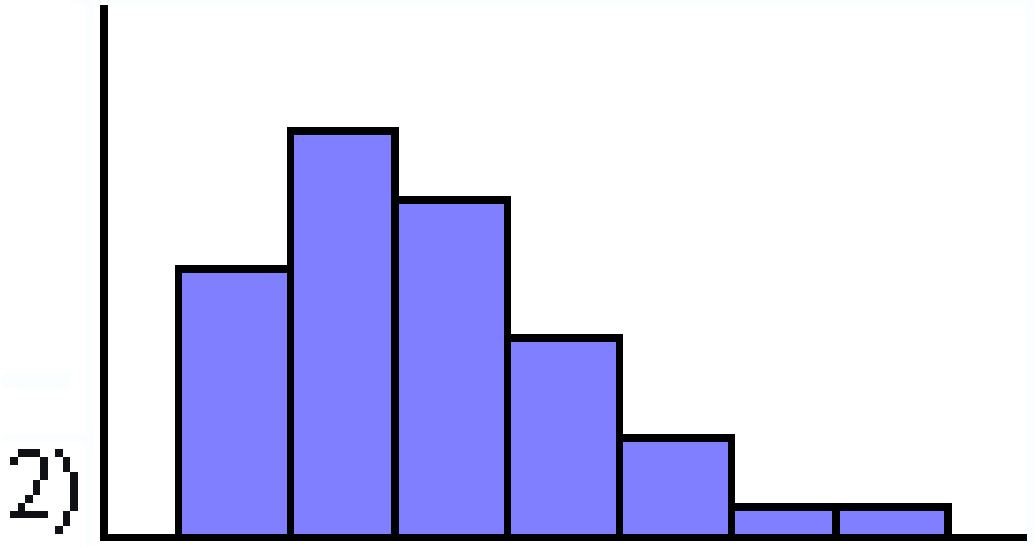
- **Identifying asymmetry or skewness profiles by comparing the relative position of the mean, median, and mode**:



- Which statistic do you think is greater/lower - mean, median or mode?
- Positive asymmetry, negative asymmetry, or symmetric distribution?

Asymmetry or Skewness profiles

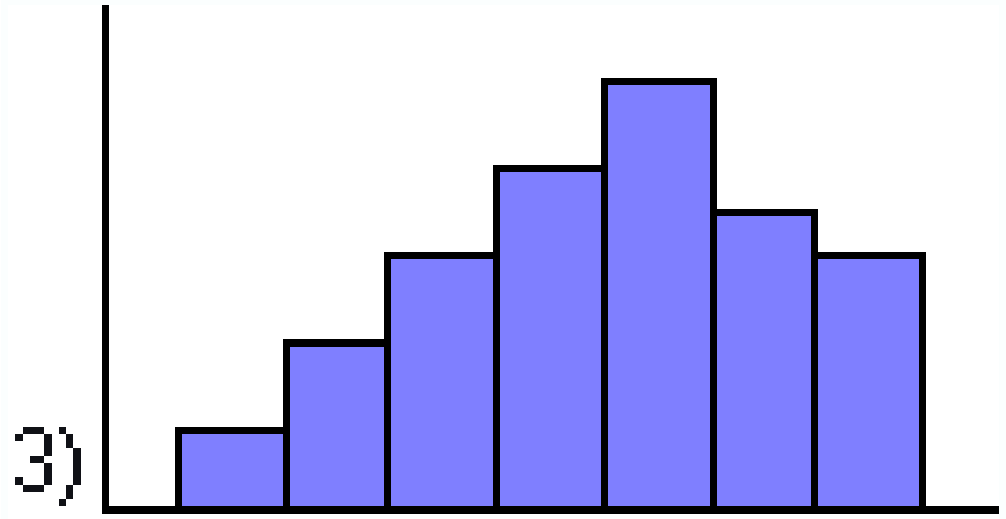
- **Identifying asymmetry or skewness profiles by comparing the relative position of the mean, median, and mode**:



- Which statistic do you think is greater/lower - mean, median or mode?
- Positive asymmetry, negative asymmetry, or symmetric distribution?

Asymmetry or Skewness profiles

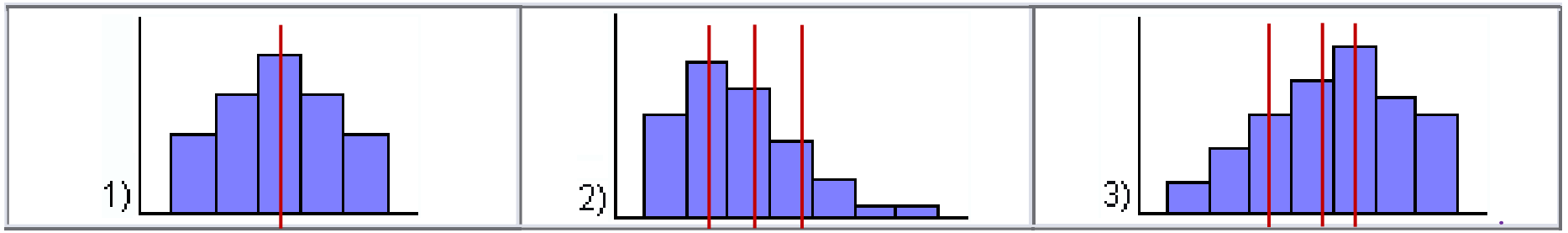
- **Identifying asymmetry or skewness profiles by comparing the relative position of the mean, median, and mode**:



- Which statistic do you think is greater/lower - mean, median or mode?
- Positive asymmetry, negative asymmetry, or symmetric distribution?

Asymmetry or Skewness profiles

- **Identifying asymmetry or skewness profiles by comparing the relative position of the mean, median, and mode**:
  - mean=median=mode – symmetrical distribution
  - mean>median>mode – positive asymmetry – skewed to the right
  - mode>median>mean – negative asymmetry – skewed to the left



Mean=Median=Mode
Symmetric distribution

Mean>Median>Mode
Positive asymmetry
(skewed to the right)

Mean<Median<Mode
Negative asymmetry
(skewed to the left)

Measures of dispersion

- The location measures (central tendency + non-central tendency) summarize the distribution of variables, but provide limited information about the degree of spread or dispersion in the data and cannot describe well how different – how heterogeneous – the data are.

- There are two types of measures of dispersion:
  - Measures of absolute dispersion: expressed in the same units as the data
  - Measures of relative dispersion: unit free or dimensionless measures

- Generally, a value of zero indicates there is no dispersion (i.e. all observations have the same value) and higher values indicate greater dispersion in the data

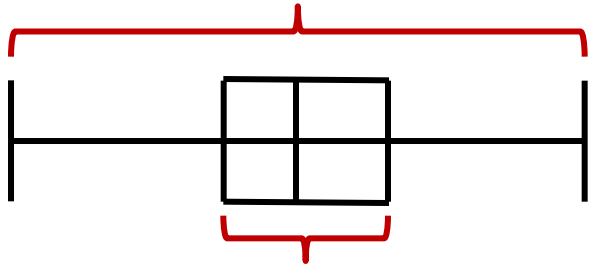
Measures of dispersion

- **Measures of absolute dispersion** – expressed in the same unit as the variable being studied, and thus not well suited to compare spread amongst different distributions
    1. Range
    2. Interquartile range
    3. Mean deviation
    4. Standard deviation and Variance

- **Measures of relative dispersion** – these are unit free, or dimensionless, measures and thus are suitable to compare spread of different distributions
    1. Interquartile relative range
    2. Coefficient of variation

Measures of dispersion

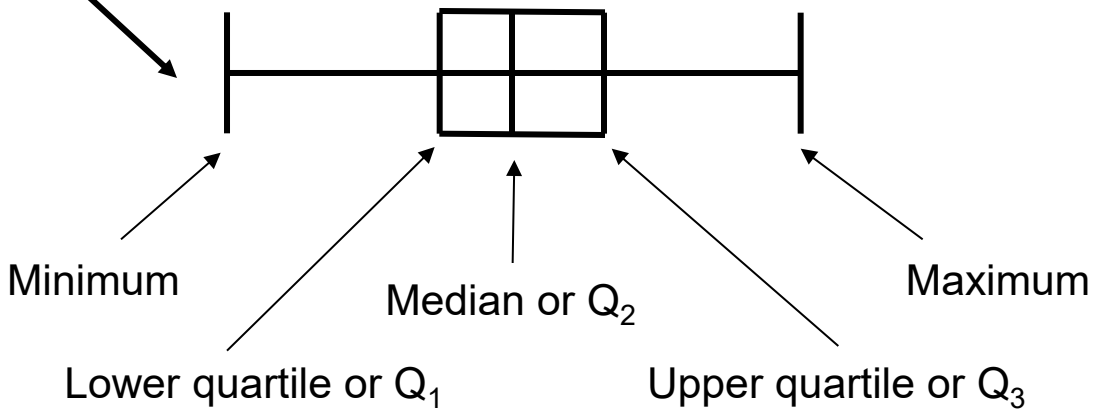
Range R = Max - Min



Interquartile range

IQR = Q_3-Q_1

Box-Whisker Diagram



Measures of dispersion

- Based on position measures:

  - **Range:** $\boldsymbol{R = Max(x) - Min(x) = X_{max} - X_{min}}$
  - Easy to compute
  - Extreme sensitivity to extreme values
  - Does not consider the values in between
  - Sensitivity to population or sample size

  - **Interquartile range:** $\boldsymbol{IRQ = Q_3 - Q_1}$
  - Lower sensitivity to extreme values
  - Does not consider all values

Measures of dispersion

- Measures of dispersion considering all the observations, ie: measure the spread amongst all the observations

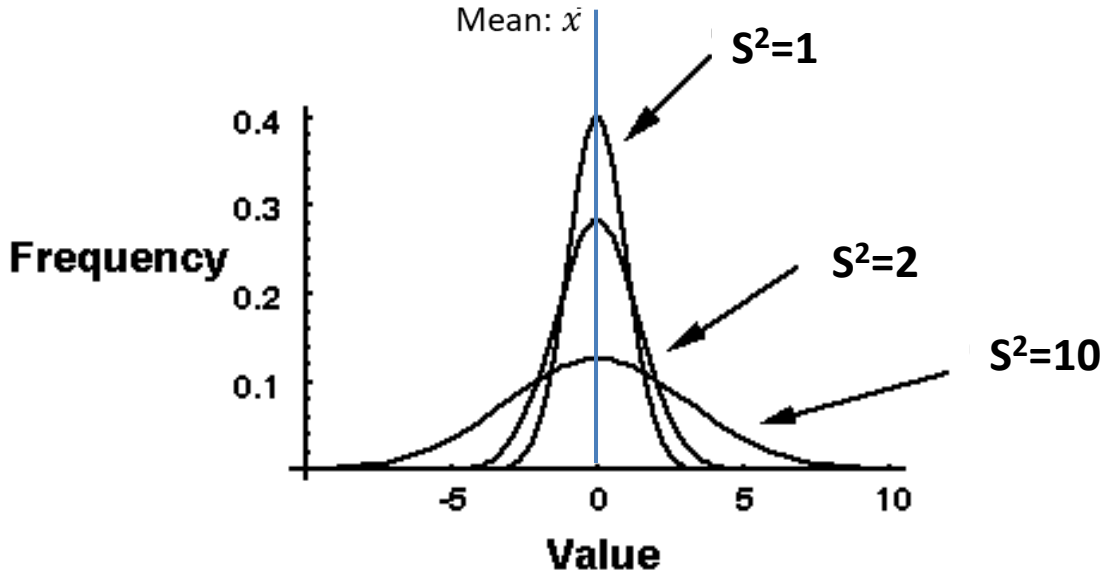
  - **Mean deviation** (not commonly used)
    - Ungrouped data: $MD_x = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$
    - Grouped data: $MD_x = \frac{\sum_{j=1}^{m} n_j|MP_j - \bar{x}|}{n} = \sum_{j=1}^{m} f_j\left|MP_j - \bar{x}\right|$

  - **Standard deviation** (very commonly used)
    - Ungrouped data: $S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$
    - Grouped data: $S_x = \sqrt{\frac{\sum_{j=1}^{m} n_j\left(MP_j - \bar{x}\right)^2}{n}} = \sqrt{\sum_{j=1}^{m} f_j\left(MP_j - \bar{x}\right)^2}$

Measures of dispersion

- Measures of dispersion considering all the observations, ie: measure the spread amongst all the observations

  - **Variance** - the square of the standard deviation: S_x^2. Very popular.
    - Ungrouped data: $S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$
    - Grouped data: $S_x^2 = \frac{\sum_{j=1}^{m} n_j \left(MP_j - \bar{x}\right)^2}{n} = \sum_{j=1}^{m} f_j \left(MP_j - \bar{x}\right)^2$



Main properties of variance and standard deviation

- The variance and standard deviation of a constant C is zero: $S^2(C) = 0$
- Adding (subtracting) a constant C to each observation of a given distribution does not change the variance nor the standard deviation:

$$S^2(\mathrm{x} + \mathrm{C}) = S^2(\mathrm{x}) \qquad S^2(\mathrm{x} - \mathrm{C}) = S^2(\mathrm{x})$$

$$\mathrm{S}(\mathrm{x} + \mathrm{C}) = S(\mathrm{x}) \qquad \mathrm{S}(\mathrm{x} - \mathrm{C}) = S(\mathrm{x})$$

- Multiplying (dividing) the observations of a given distribution by a non-zero constant C results in the multiplication (division) of:
  - the variance by the square of that constant:

$$S^2(\mathrm{xC}) = C^2 * S^2(\mathrm{x}) \qquad S^2(\mathrm{x/C}) = S^2(\mathrm{x})/C^2$$

  - the standard deviation by that constant:

$$\mathrm{S}(\mathrm{xC}) = C * S(\mathrm{x}) \qquad \mathrm{S}(\mathrm{x/C}) = S(\mathrm{x})/C$$

Measures of relative dispersion

- Allow making comparisons between distributions - **generally take the ratio between absolute dispersion measure and location measure**

- **Relative interquartile range (RIQR)**:
  - Interquartile range (IQR) divided by the median (Q_2): $RIQR = \frac{IQR}{Q_2} = \frac{Q_3 - Q_1}{Q_2} = \frac{Q_3 - Q_1}{x_{ME}}$
  - Interpretation: e.g. RIQR=0.25 means the interquartile range (i.e. the middle 50% of the distribution) is 25% of the value of the median

- **Coefficient of variation (CV)**:
  - Standard deviation divided by the mean: $CV_x = \frac{S_x}{\bar{x}}$
  - Interpretation: e.g. CV=0.25 means the standard deviation is 25% of the value of the mean

Data Analysis for Economics and Business

Lecture 10: Analyzing numerical information

Measures of concentration

Academic Year 2023/24

Structure of lecture

- Measures of concentration
  - Gini coefficient
  - Lorenz curve

Learning outcomes

- Define and explain key measures of concentration
- Calculate and interpret the Gini Index
- Draw the Lorenz curve

Measures of concentration

- Measure the extent to which a given attribute or variable is more evenly or unevenly distributed across the members of the population or sample

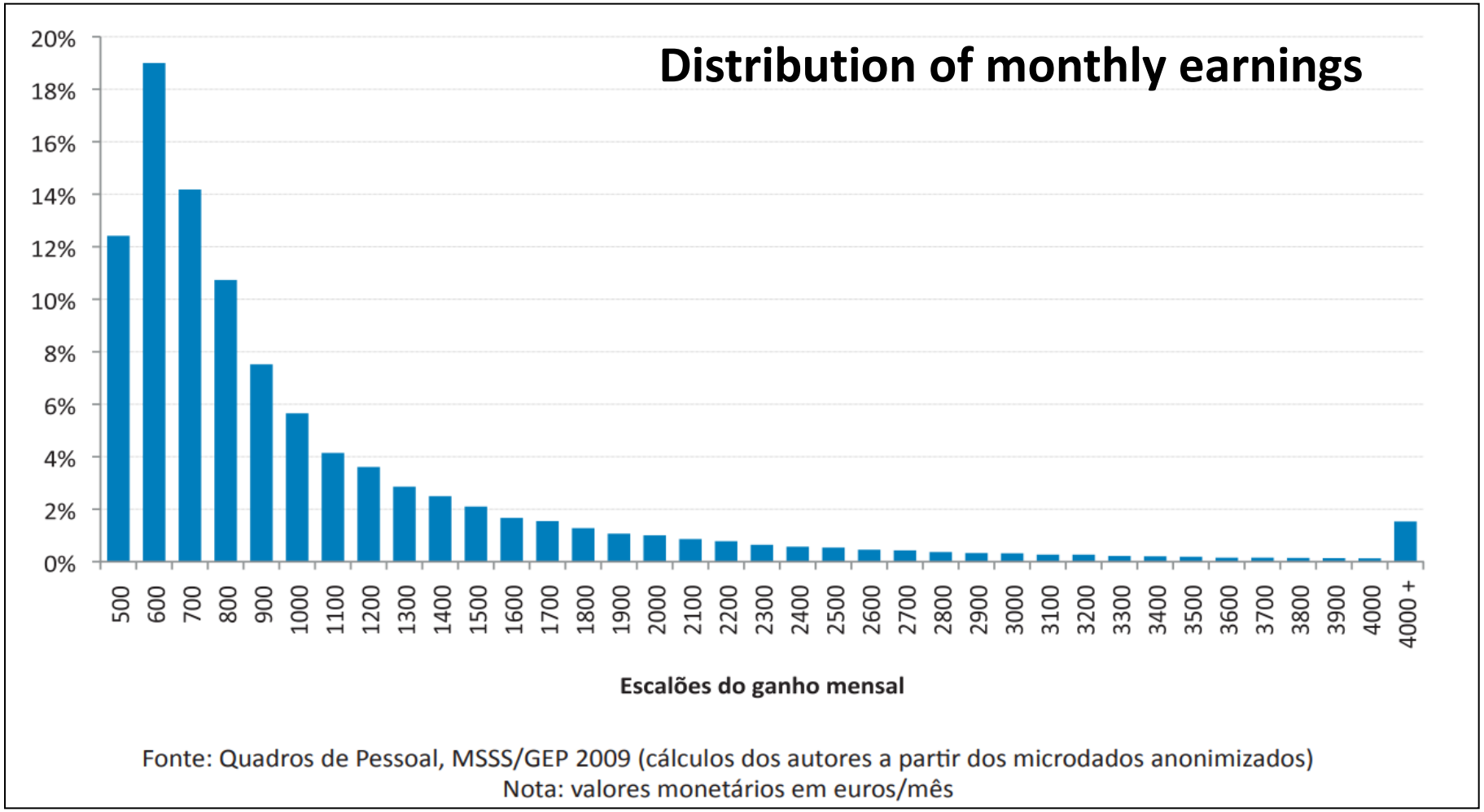
- Two extreme cases:
  - Attribute (e.g. income) is equally distributed amongst members – zero concentration
  - Attribute concentrated in a single member or group of members – maximum concentration

- Generally we have an intermediate situation where there is some degree of "unevenness" in the distribution of the attribute amongst the population or sample

Measures of concentration

What is the level of inequality in the distribution of earnings in Portugal?

Right-skewed (long tail to the right) due to large earning values

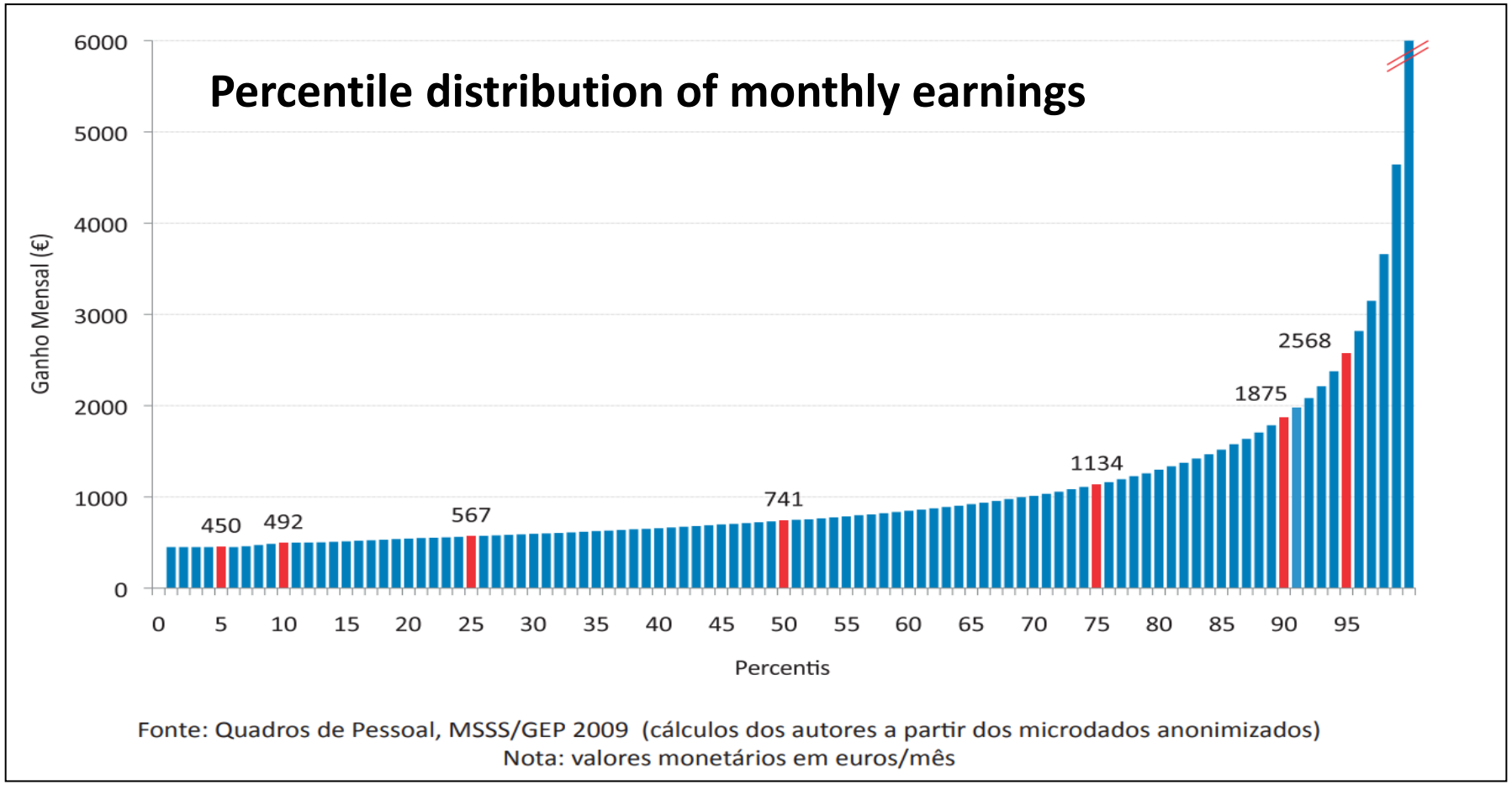


Fonte: Quadros de Pessoal, MSSS/GEP 2009 (cálculos dos autores a partir dos microdados anonimizados)
Nota: valores monetários em euros/mês

Measures of concentration

What is the level of inequality in the distribution of earnings in Portugal?

Looking at cum. rel. frequencies: percentile bars become increasingly taller than previous ones at the top of the distribution (around P75(=Q3) or so)



Source: Farinha, C., Figueiras, R. and Junqueira, V. (2012) Desigualdade Económica em Portugal. Fundação Francisco Manuel dos Santos.

Measures of concentration

- **What is the level of inequality in the distribution of earnings in Portugal? or**
- **How uneven is the distribution of earnings amongst workers in Portugal?**
  - compare:

    the cumulative relative distribution of the variable

    across (i.e., “accumulated” over)

    the (groups of) statistical units

    using a given metric (e.g. Gini Index, P80/P20 ratios)

  - high concentration:
    a small proportion of statistical units has a large proportion of the variable analysed

- Variables must be ranked and additive (ex: sales, income, etc.)

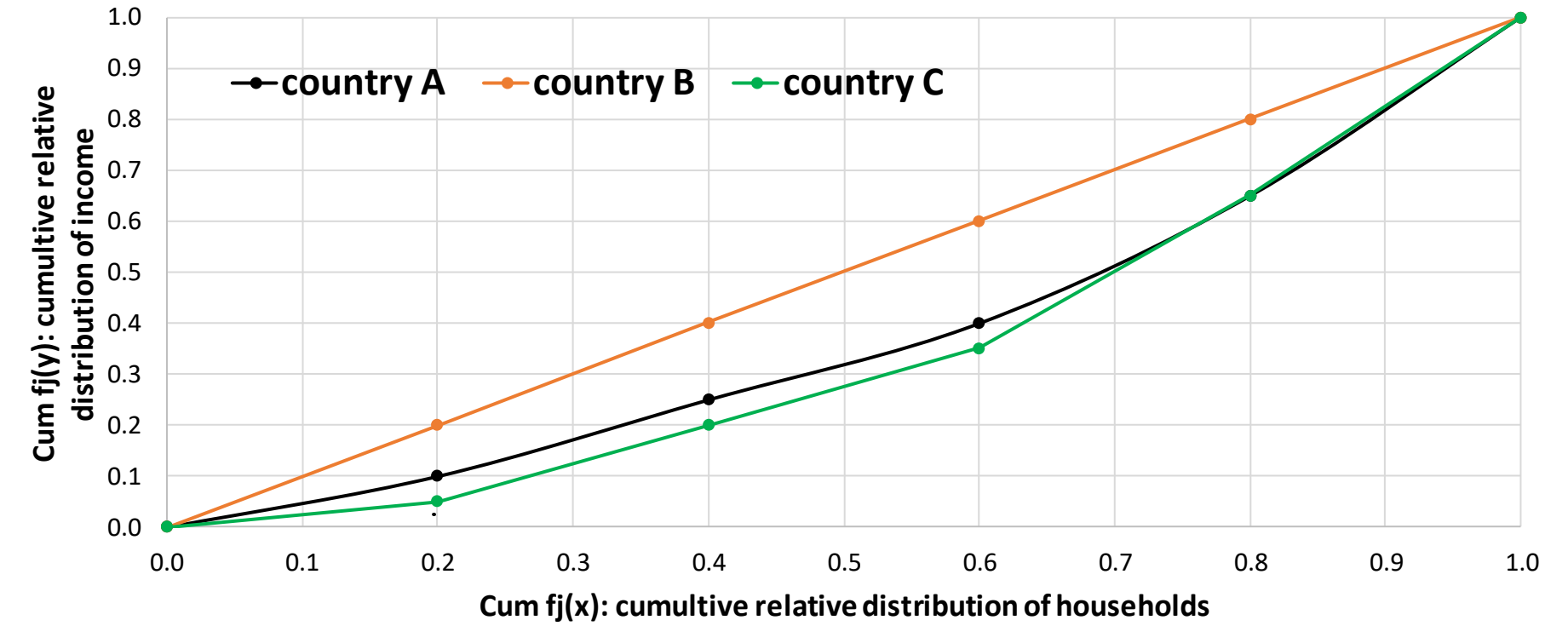
Meaures of concentration

Which country has a greater level of concentration (or inequality) in the distribution of income? (Y = income, X = households)

Quintiles of households (X)	country A	country B	country C
	% income (fj(Y))	% income (fj(Y))	% income (fj(Y))
1 (bottom 20%)	0.10	0.20	0.05
2	0.15	0.20	0.15
3	0.15	0.20	0.15
4	0.25	0.20	0.30
5 (top 20%)	0.35	0.20	0.35

compare the cumulative relative distribution of the variable "accumulated" over the statistical units

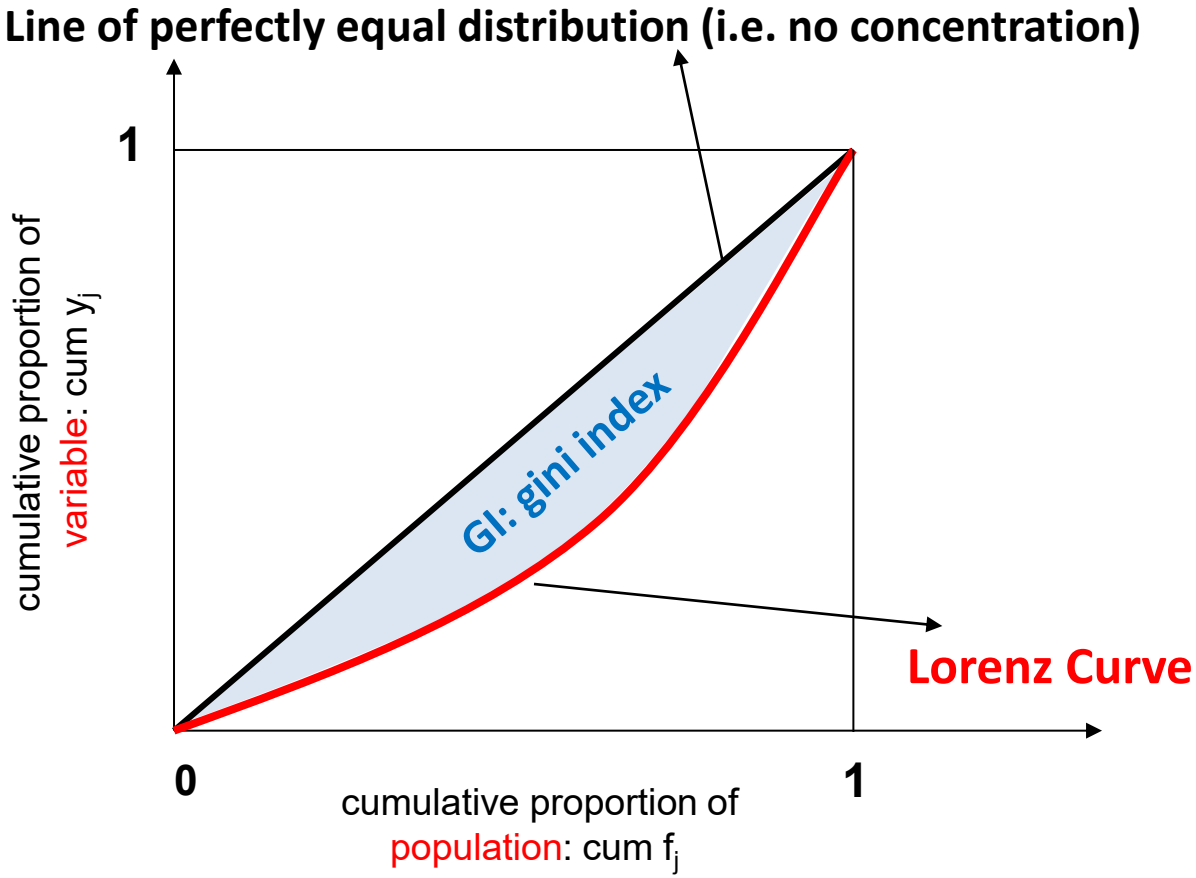
Quintiles of households (X)	country A	country B	country C	country A	country B	country C
	% income (fj(Y))	% income (fj(Y))	% income (fj(Y))	cum fj(y)=qj	cum fj(y)=qj	cum fj(y)=qj
1 (bottom 20%)	0.10	0.20	0.05	0.10	0.20	0.05
2	0.15	0.20	0.15	0.25	0.40	0.20
3	0.15	0.20	0.15	0.40	0.60	0.35
4	0.25	0.20	0.30	0.65	0.80	0.65
5 (top 20%)	0.35	0.20	0.35	~~1.00~~	~~1.00~~	~~1.00~~



note: *cum fj(x) is your typical observed cumulative frequencies;*
to get cum fj(y) one needs to first add all values of the variable analysed

Measures of concentration

- The **Lorenz curve plots Cum $f_j(x)$ against Cum $f_j(y)$**
- **Gini Index (GI)**: measures the area between the two lines; the greater the area the more unequal (i.e. greater concentration) the distribution



Measures of concentration

- **Gini Index**: one popular measure of concentration or inequality

Cumulative relative distribution of the population

Cumulative relative distribution of the variable

$$GI = \frac{\sum_{j=1}^{m-1}\left(cumf_j(x) - cum\ f_j(y)\right)}{\sum_{j=1}^{m-1} cumf_j(x)} = \frac{\sum_{j=1}^{m-1}\left(p_j - q_j\right)}{\sum_{j=1}^{m-1} p_j} = 1 - \frac{\sum_{j=1}^{m-1} q_j}{\sum_{j=1}^{m-1} p_j}$$

Where:

$$q_j = cum\ f_j(y)$$

$$p_j = cum\ f_j(x)$$

note: *the summation signs go from j=1 to m-1, i.e., sum all classes but the last one!*

- 0≤ GI ≤1: the greater the GI the greater the level of concentration
- GI=0: for perfectly even or equal distribution of the attribute
- GI=1: when the distribution of the attribute is fully concentrated in one member

Quintiles of households (X)	country A	country B	country C	country A	country B	country C	pj=cum fj(X)
	cum fj(y)=qj	cum fj(y)=qj	cum fj(y)=qj	pj-qj	pj-qj	pj-qj	
1 (bottom 20%)	0.10	0.20	0.05	0.10	0.00	0.15	0.2
2	0.25	0.40	0.20	0.15	0.00	0.20	0.4
3	0.40	0.60	0.35	0.20	0.00	0.25	0.6
4	0.65	0.80	0.65	0.15	0.00	0.15	0.8
5 (top 20%)	~~1.00~~	~~1.00~~	~~1.00~~	~~0.00~~	~~0.00~~	~~0.00~~	~~1~~
	1.40	**2.00**	**1.25**	**0.60**	**0.00**	**0.75**	**2**

$$GI(country\ A) = \frac{\sum_{j=1}^{m-1}\left(cumf_j(x) - cum\ f_j(y)\right)}{\sum_{j=1}^{m-1} cumf_j(x)} =$$

$$= \frac{\sum_{j=1}^{m-1}(p_j - q_j)}{\sum_{j=1}^{m-1} p_j} = (0.10+0.15+0.20+.015)/(0.2+0.4+0.6+0.8) = 0.60/2 = 0.30$$

$$= 1 - \frac{\sum_{j=1}^{m-1} q_j}{\sum_{j=1}^{m-1} p_j} = 1-(0.10+0.25+0.40+0.65)/(0.2+0.4+0.6+0.8) = 1-(1.4/2) = 0.30$$

GI	1-(sumqj/sumpj)	sum(pj-qj)/sumpj
Country A	0.30	0.30
Country B	0.00	0.00
Country C	0.38	0.38

Measures of concentration

- Example of Gini Index using grouped continuous variable: Average monthly earnings of individuals aged 25-30 (hypothetical)

Earnings (y)	Workers: Fj(x)
[250-450[	500
[450, 650[	5000
[650, 850[	5000
[850, 1050[	5000
[1050,1250[	8000
[1250,1450[	2500
[1450,1650[	2500
[1650,1850[	2500
[1850,2050[	2000
[2050,2250[	1250
[2250,2450[	500
[2450,2650]	250
Total	**35000**

Measures of concentration

- Example of Gini Index using grouped continuous variable: Average monthly earnings of individuals aged 25-30 (hypothetical)

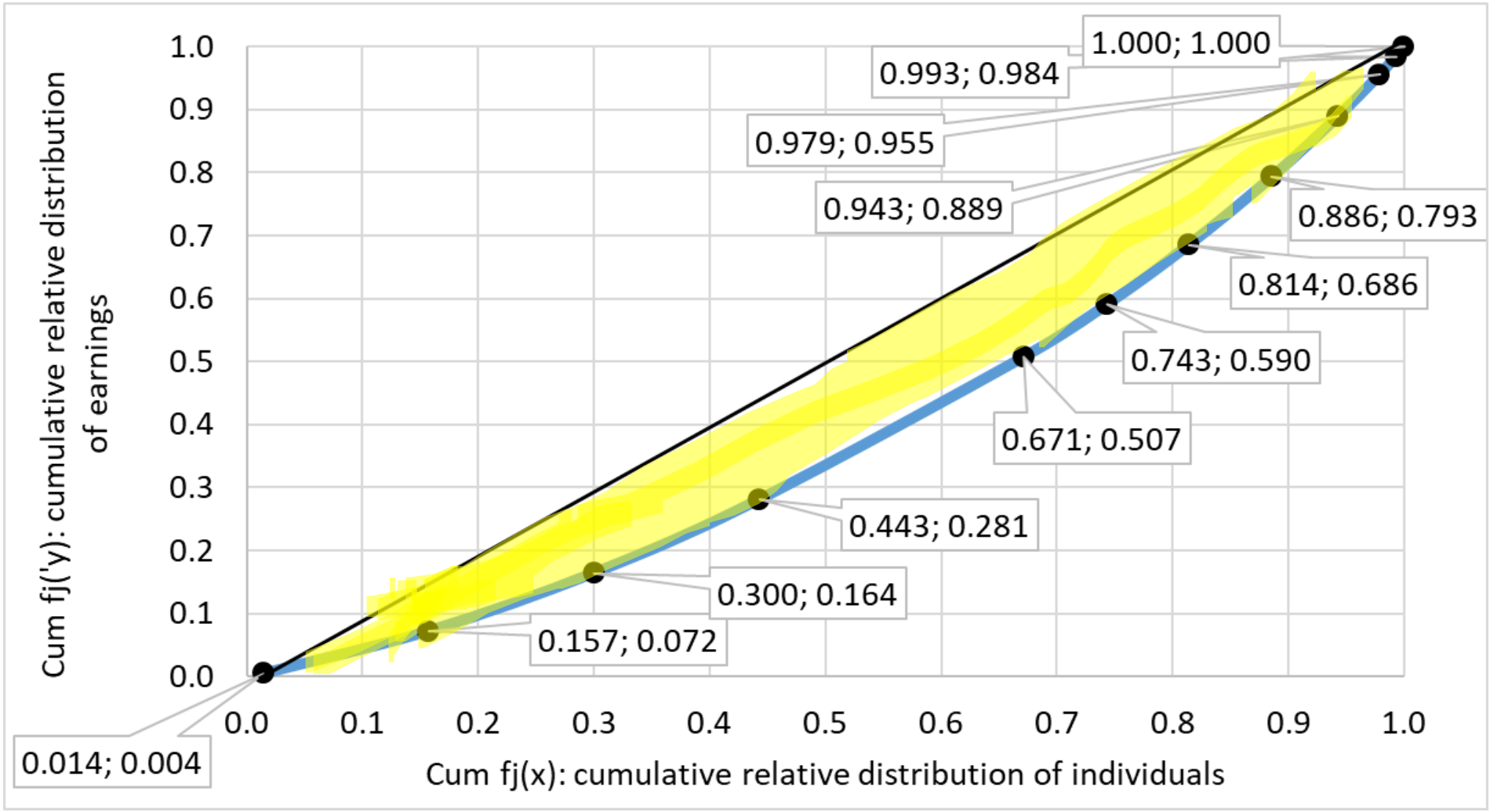
Earnings (y)	Workers: Fj(x)	MPj(y)	y'=Fj(x)*MPj(y)
[250-450[	500	350	175000
[450, 650[	5000	550	2750000
[650, 850[	5000	750	3750000
[850, 1050[	5000	950	4750000
[1050,1250[	8000	1150	9200000
[1250,1450[	2500	1350	3375000
[1450,1650[	2500	1550	3875000
[1650,1850[	2500	1750	4375000
[1850,2050[	2000	1950	3900000
[2050,2250[	1250	2150	2687500
[2250,2450[	500	2350	1175000
[2450,2650]	250	2550	637500
Total	**35000**	-	**40650000**

Measures of concentration

- Example of Gini Index using grouped continuous variable: Average monthly earnings of individuals aged 25-30 (hypothetical)

Earnings (y)	Workers: Fj(x)	MPj(y)	y'=Fj(x)*MPj(y)	fj(y')	fj(x)	cum fj(y')	cum fj(x)
[250-450[	500	350	175000	0,004	0,014	0,004	0,014
[450, 650[	5000	550	2750000	0,068	0,143	0,072	0,157
[650, 850[	5000	750	3750000	0,092	0,143	0,164	0,300
[850, 1050[	5000	950	4750000	0,117	0,143	0,281	0,443
[1050,1250[	8000	1150	9200000	0,226	0,229	0,507	0,671
[1250,1450[	2500	1350	3375000	0,083	0,071	0,590	0,743
[1450,1650[	2500	1550	3875000	0,095	0,071	0,686	0,814
[1650,1850[	2500	1750	4375000	0,108	0,071	0,793	0,886
[1850,2050[	2000	1950	3900000	0,096	0,057	0,889	0,943
[2050,2250[	1250	2150	2687500	0,066	0,036	0,955	0,979
[2250,2450[	500	2350	1175000	0,029	0,014	0,984	0,993
[2450,2650]	250	2550	637500	0,016	0,007	~~1,000~~	~~1,000~~
Total	**35000**	-	**40650000**	**1,000**	**1,000**	**5,927**	**6,943**

Lorenz curve



Gini Index

A	B	C	D=B*C	E	F	G	H	I=H-G
Earnings (y)	Workers: Fj(x)	MPj(y)	y'=Fj(x)*MPj(y)	fj(y')	fj(x)	cum fj(y')	cum fj(x)	cum fj(x)-cum fj(y')
[250-450[	500	350	175000	0,004	0,014	0,004	0,014	0,010
[450, 650[	5000	550	2750000	0,068	0,143	0,072	0,157	0,085
[650, 850[	5000	750	3750000	0,092	0,143	0,164	0,300	0,136
[850, 1050[	5000	950	4750000	0,117	0,143	0,281	0,443	0,162
[1050,1250[	8000	1150	9200000	0,226	0,229	0,507	0,671	0,164
[1250,1450[	2500	1350	3375000	0,083	0,071	0,590	0,743	0,152
[1450,1650[	2500	1550	3875000	0,095	0,071	0,686	0,814	0,129
[1650,1850[	2500	1750	4375000	0,108	0,071	0,793	0,886	0,092
[1850,2050[	2000	1950	3900000	0,096	0,057	0,889	0,943	0,054
[2050,2250[	1250	2150	2687500	0,066	0,036	0,955	0,979	0,023
[2250,2450[	500	2350	1175000	0,029	0,014	0,984	0,993	0,009
[2450,2650]	250	2550	637500	0,016	0,007	~~1,000~~	~~1,000~~	~~0,000~~
Total	**35000**	-	**40650000**	**1,000**	**1,000**	**5,927**	**6,943**	**1,015**

$$GI = \frac{\sum_{j=1}^{m-1}\left(cumf_j(x) - cum\,f_j(y)\right)}{\sum_{j=1}^{m-1} cumf_j(x)} = 1 - \frac{\sum_{j=1}^{m-1} cum\,f_j(y)}{\sum_{j=1}^{m-1} cum\,f_j(x)} = \frac{1{,}015}{6{,}943} = 1 - \frac{5{,}927}{6{,}943} = 0{,}1463$$

Illustration in Excel

Data Analysis for Economics and Business

Lecture 12:
Association between variables
Linear regression analysis

Academic Year 2023/24

Structure of lecture

- Patterns of association between variables
  - Scatter plots
  - Covariance
  - Linear correlation coefficient
- Estimation of the linear regression model (LRM): the least squares (LS) method
- Interpretation of the coefficients: the intercept and the slope

Learning outcomes

- Describe and classify patterns of association between variables
  - Build scatter plots
  - Define, calculate and interpret the covariance
  - Define, calculate and interpret the linear correlation coefficient
- Explain how to apply the least squares (LS) method to estimating the linear regression model (LRM)
- Calculate the coefficients of the LRM: the intercept and the slope
- Explain the meaning of the coefficients of the LRM (intercept and slope)
- Explain how we can evaluate the goodness of fit of the LRM

Association between variables

- We have been dealing with one variable at a time. However, in many situations we are interested in the relationship between two or more variables

- Examples:
  - University tuition fees & level of education: **Do higher tuition fees reduce the number of students going to university? By how much?**
  - Education & earnings: **Does educational attainment influence worker's earnings? What is the wage premium of one additional year of education?**
  - Price of fuel, parking fees & driving: **Do increases in fuel price and parking fees reduce demand for cars? By how much?**
  - Car use & city population density: **Do people drive less in denser cities?**

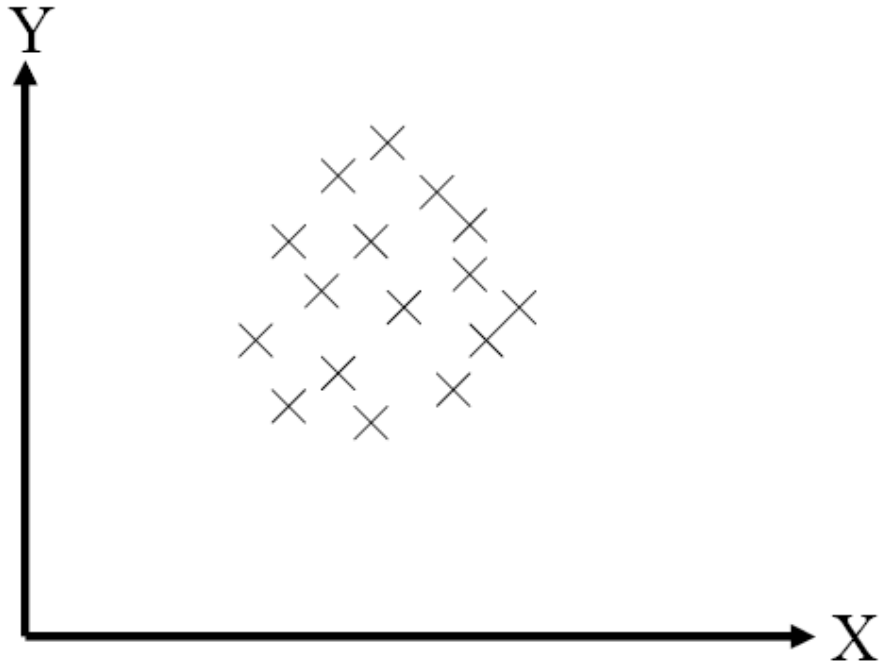
Association between variables

- It is important to analyze the relationship between variables because knowing the behavior of one variable **may help predicting** the other variable's behavior

- But remember: **Correlation is not causation!** (see first lectures)

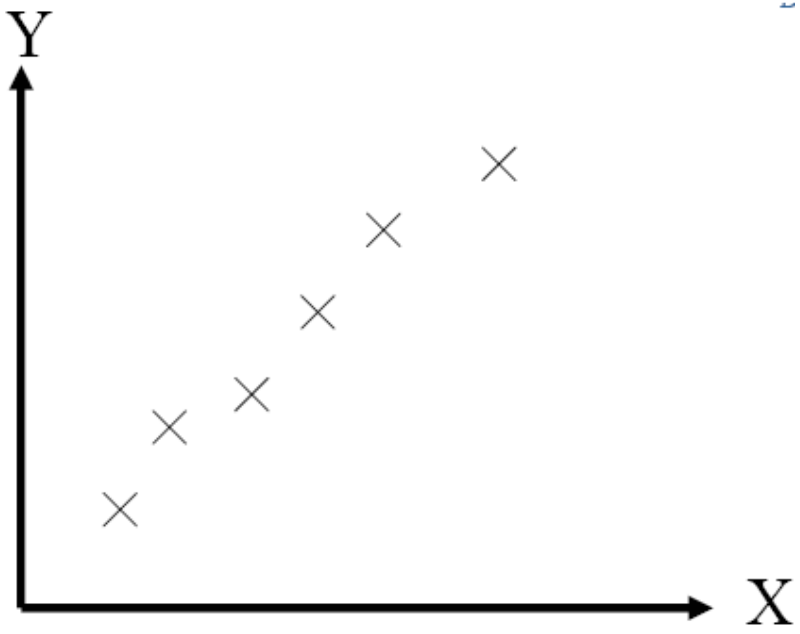
Association between variables

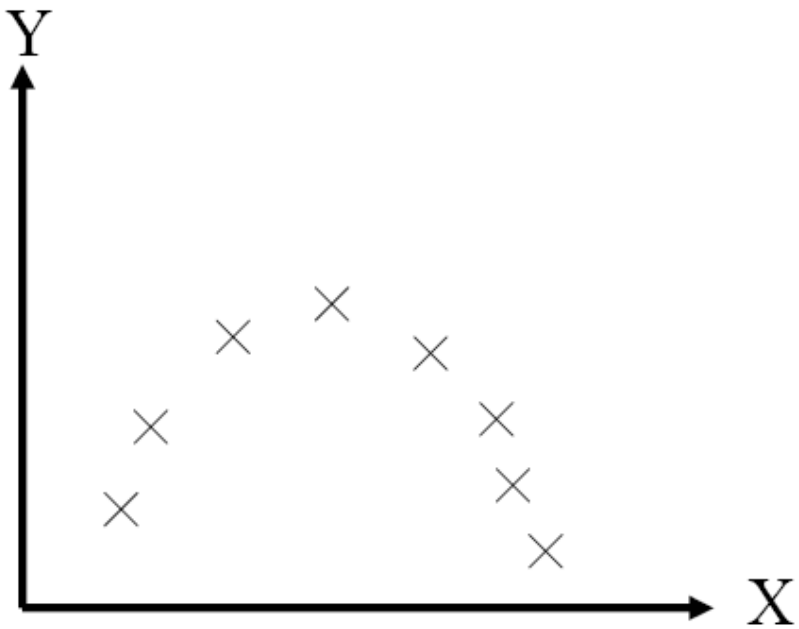
- A starting point to assess the existence of a relationship between two variables is to use a scatter plot, that is, a graphical depiction of the pairs (x,y)



Association between variables

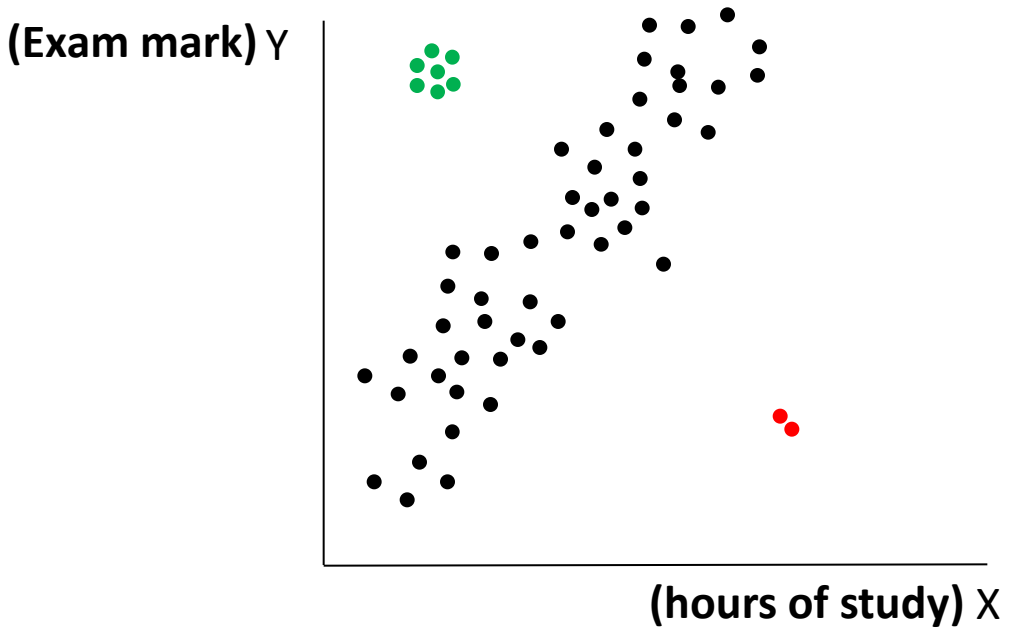
- A starting point to assess the existence of a relationship between two variables is to use a scatter plot, that is, a graphical depiction of the pairs (x,y)





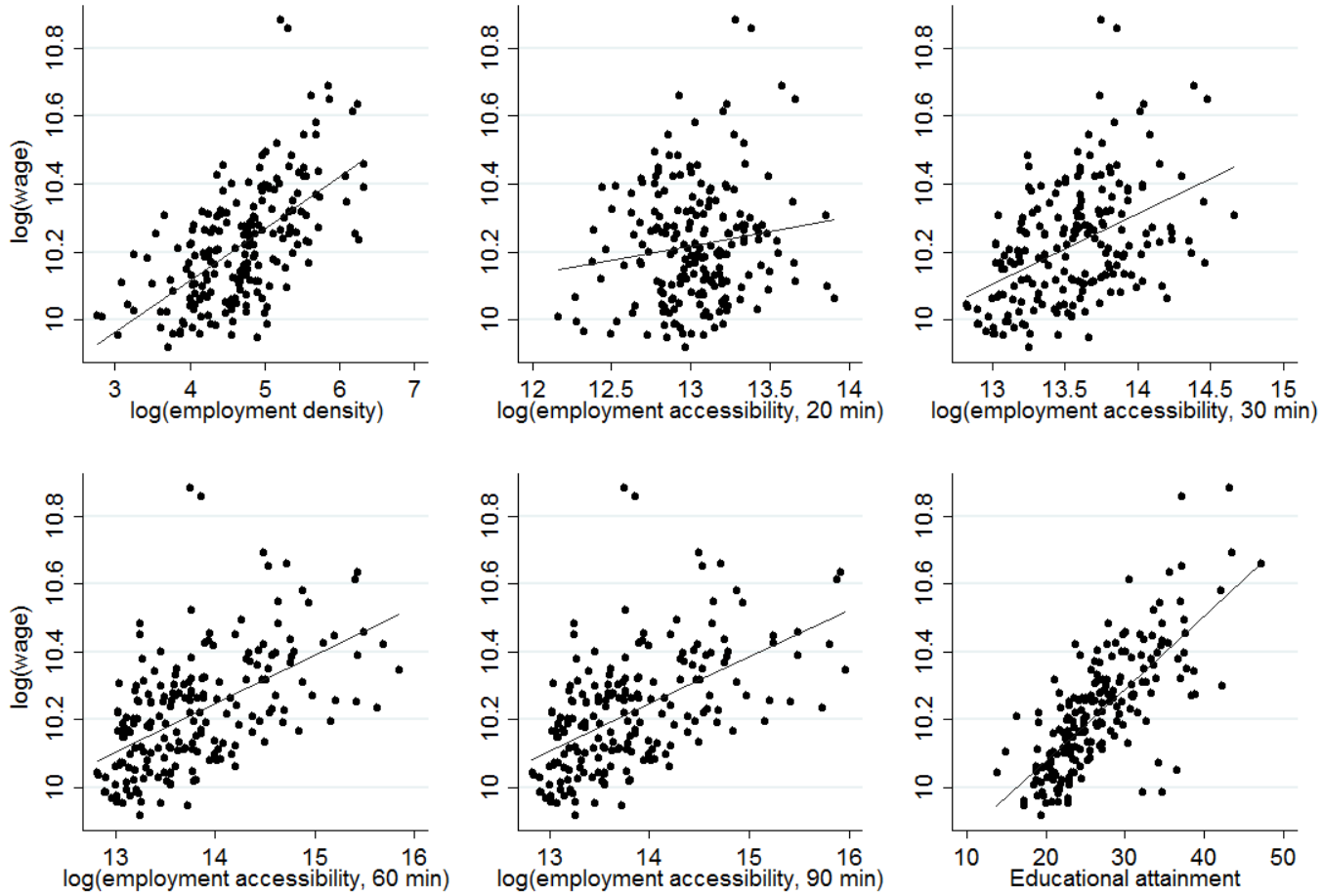
Association between variables

- Association between final exam mark & hours of study: How many hours do you need to study to get a good mark?



Association between variables

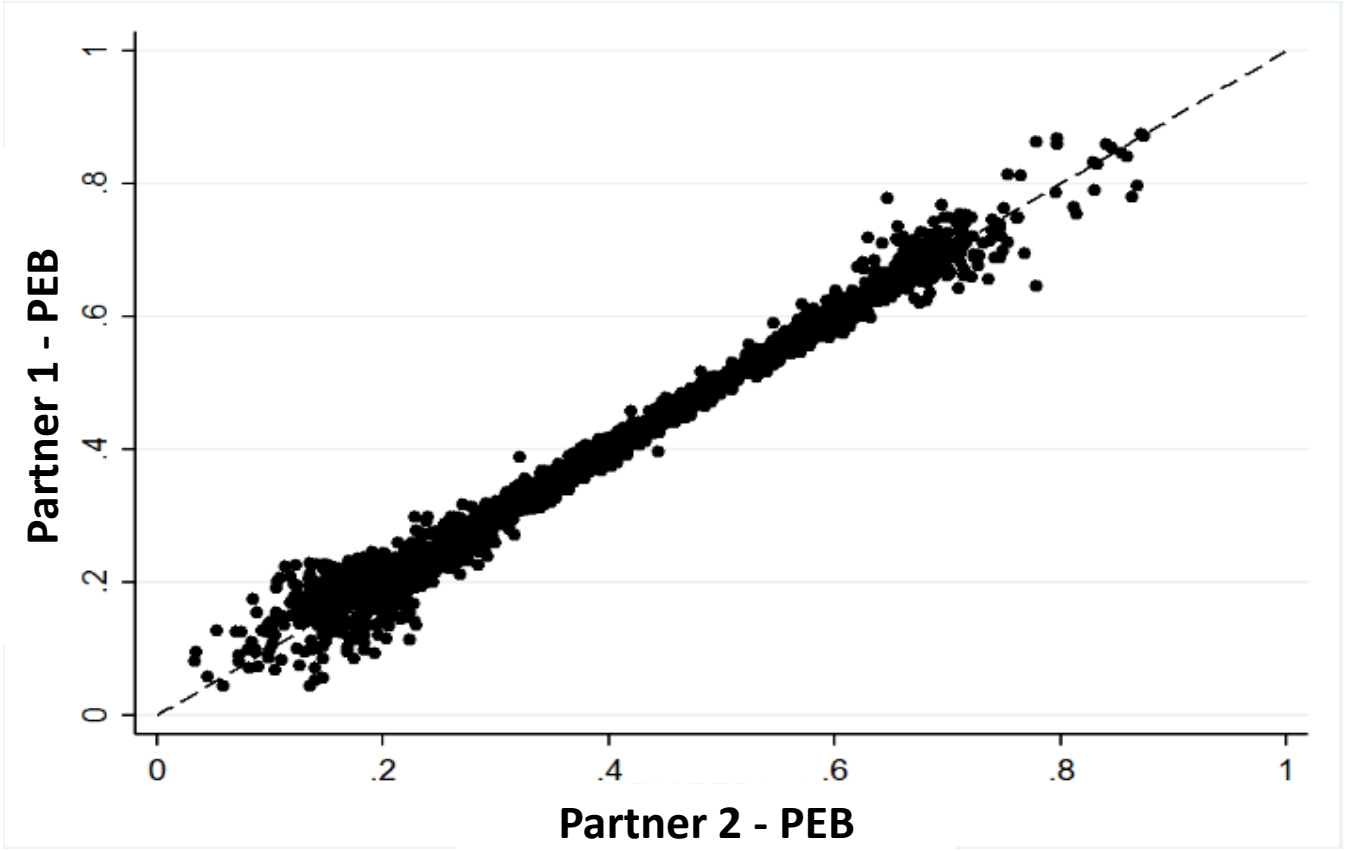
- Scatter plot for relationship between earnings, education, employment density and accessibility



Melo, P.C., Graham, D.J., Levinson, D. and Aarabi, S. 2017. Agglomeration, Accessibility, and Productivity: Evidence for Large Metropolitan Areas in the US. *Urban Studies* Vol. 54(1), pp. 179-195

Association between variables

- Scatter plot for pro-environmental behaviour (PEB) index of partners (couples)



Melo, P.C., Ge, J., Craig, T., Brewer, M., Thronicker, I. 2018. Does work-life balance affect pro-environmental behaviour? Evidence for the UK using longitudinal microdata. *Ecological Economics* Vol. 145, pp. 170-181

Association between variables

- The scatter plot gives a visual indication on the existence of a relationship between the variables – i.e. how much the variables appear to be **correlated** with each other

- Patterns of association:
  - **Positive association**– variables move **together in the same direction**, i.e. increase and decrease together (e.g. temperature and ice cream consumption)
  - **Negative association** – variables move **together in opposite direction**, i.e. one increases and the other decreases (e.g. temperature and hours of study)
  - **No relationship** – variables do note move together in a clear way.

Measuring the association between variables

- We can use the **covariance** to measure the nature (positive or negative) of **linear** association between two variables – i.e. if two variables vary together (i.e. co-vary) in the same direction or in opposite directions.

$$S_{YX} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$$

Measuring the association between variables

- Interpretation of **Covariance**:
  - Positive covariance ($S_{yx}>0$): variables move together in the same direction, there is a positive (linear) association between them.
  - Negative covariance ($S_{yx}<0$): variables move together in opposite direction, there is a negative (linear) association between them.

- Limitation: the magnitude of the covariance tends to increase with the scale of the variables (e.g., dozens vs. millions), which makes its interpretation difficult.

- To overcome this limitation we can use a normalized measure (i.e., a measure that does not depend on the scale of the variables): the **correlation coefficient**

Measuring the association between variables

- The **linear correlation coefficient** measures the nature and magnitude of the **linear** correlation between two variables:

$$r_{YX} = \frac{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 \frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2}} = \frac{S_{YX}}{S_X S_Y}$$

- Interpretation:
  - If r_{YX}= 0, this implies that S_{YX}=0 and there is no linear relation
  - if r_{YX}= 1: perfect positive linear relationship
  - If r_{YX}= -1: perfect negative linear relationship

  The closer r_{YX} is to one in absolute terms the stronger the association

NOTE: again, r_{YX} is a measure of linear relationship, r_{YX} = 0 does not mean there is no relationship between x and y (only there is no **linear** relationship).

Measuring the association between variables

- Advantages of the **linear correlation coefficient:**

  - While the covariance depends on the units of the variables (can take many values), the correlation coefficient is unit free (it is limited between -1 and 1) and thus is not affected by changes in the mean or in the scale of the variables.

  - Hence, compared to covariance, the correlation coefficient has an easier interpretation and using it makes it easier to determine the strength or magnitude of the relationship between two variables.

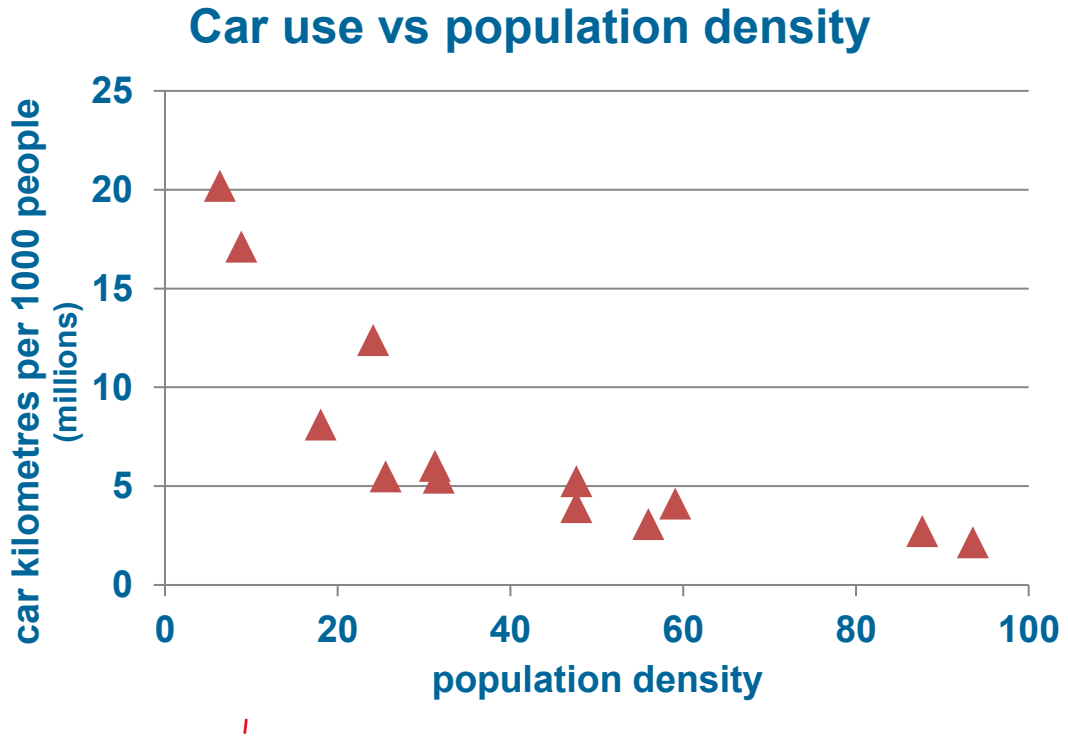
Linear regression model

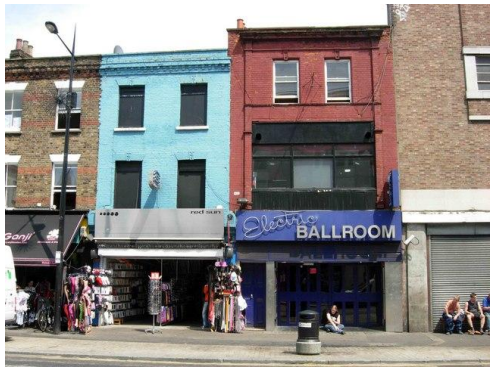
- Suppose we have two variables Y and X and we think that Y can be related to X via a linear relationship.

- Previous examples:
  - University tuition fees & level of education: **Do higher tuition fees reduce the number of students going to university? By how much?**
  - Education & earnings: **Does educational attainment influence worker's earnings? What is the wage premium of one additional year of education?**
  - Price of fuel, parking fees & driving: **Do increases in fuel price and parking fees reduce demand for cars? By how much?**
  - Car use & city population density: **Do people drive less in denser cities?**

Linear regression model

- We start by drawing a scatter plot to inspect visually whether there may be a linear relation, and if so we can fit a regression line that better represents the relationship between the two variables





Source: Millennium Cities Database

Linear regression model

- The relationship for Y and only one X variable can be represented by:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

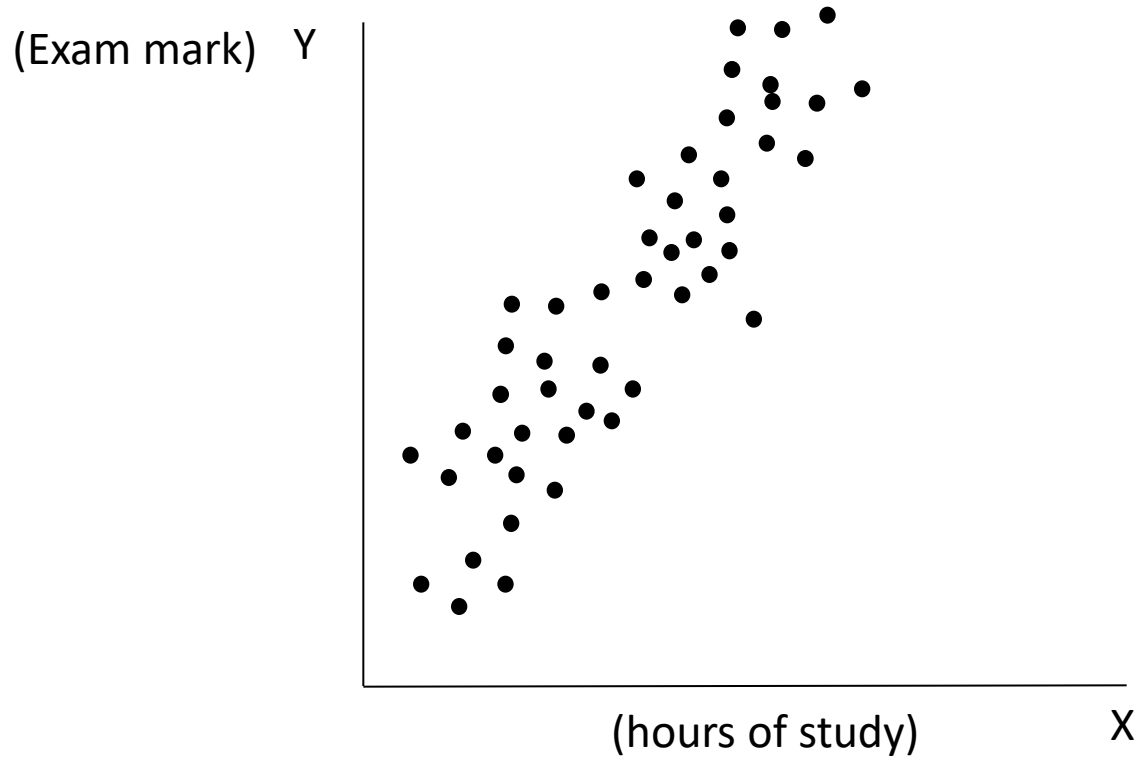
where **ε** is the residual term which allows for random variation between the observed and fitted values for Y (more on that in four slides from now).

- Multiple regression model: when there is more than one explanatory variable X (which in social science is virtually always the case):

$$Y_i = b_0 + \sum_{j=1}^{k} b_j X_i + \varepsilon_i$$

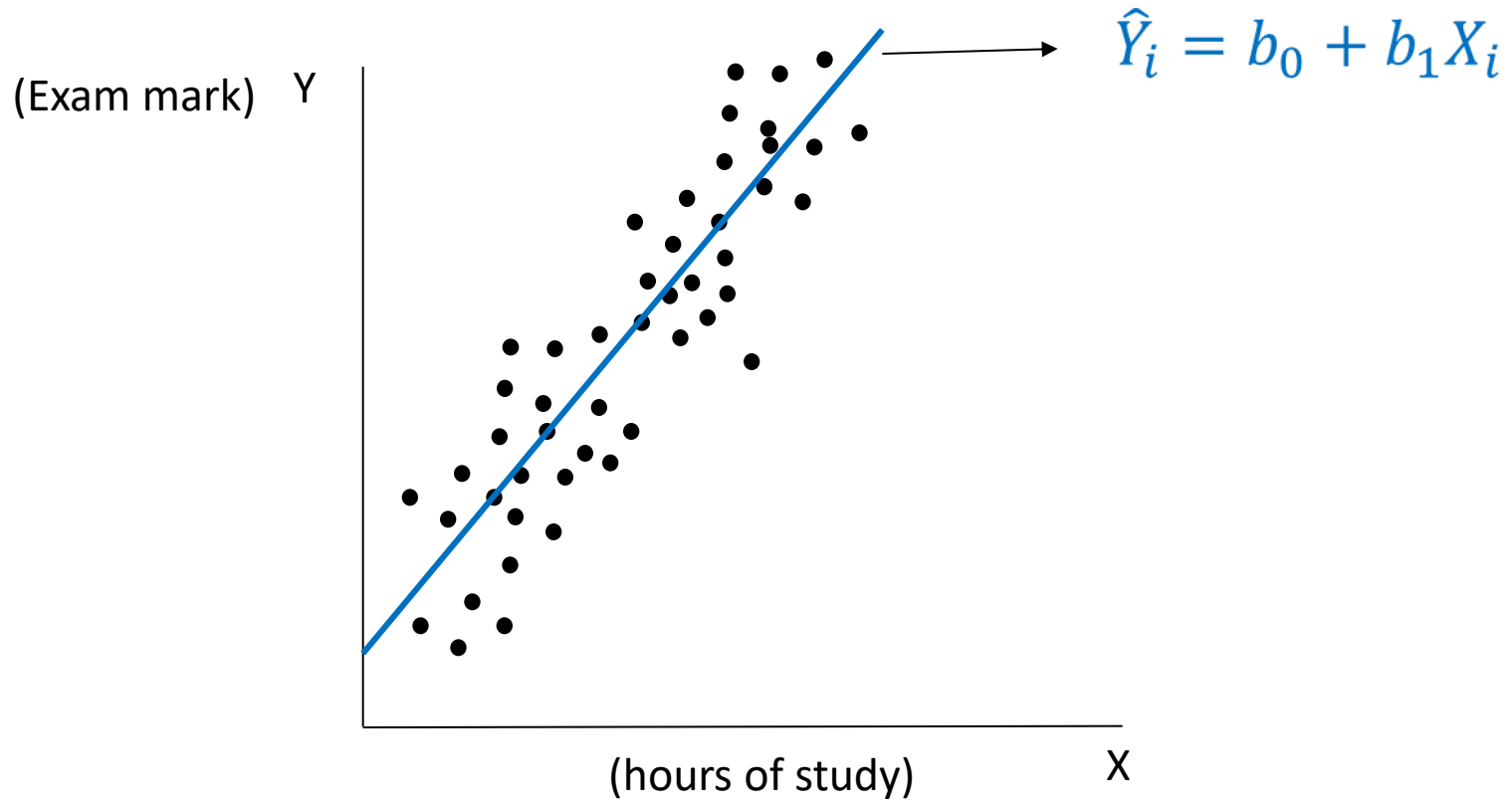
Linear regression model

- Again, the relationship for Y and only one X variable can be represented by:



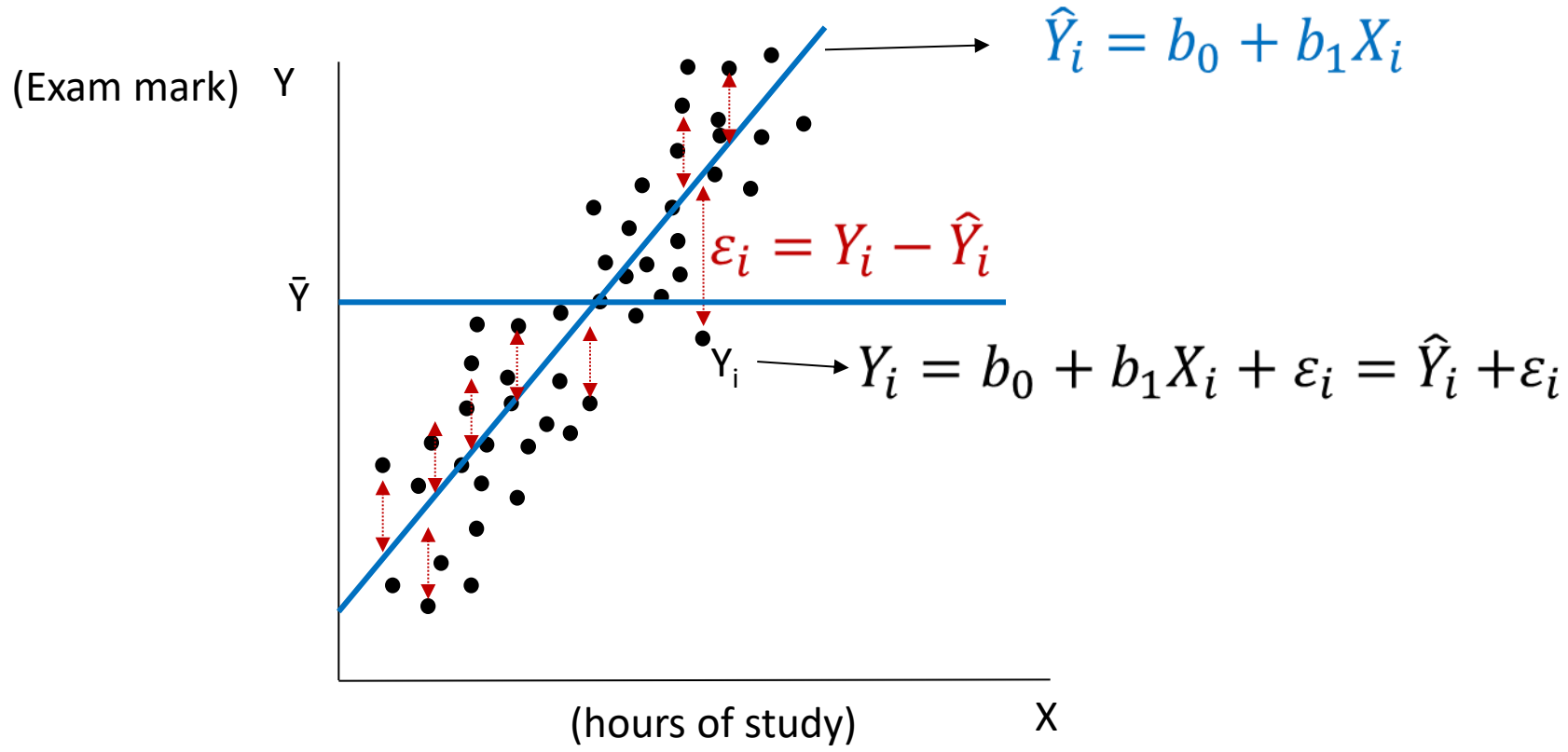
Linear regression model

- Again, the relationship for Y and only one X variable can be represented by:



Linear regression model

- Again, the relationship for Y and only one X variable can be represented by:



Why do we add an error term?

The addition of the error term recognizes that the relationship is not exact / deterministic - there is a **random** element.

Specific reasons for the existence of a **random** error term:

- Unpredictable elements of randomness in behaviour
- Omission of explanatory variables
- Measurement error
- Functional misspecification

Linear regression model

- We want to **estimate b_0 and b_1 that minimise the residuals ε between the observed values Y_i** and **fitted values $\widehat{Y_i}$**.

- The residual is equal to:

$$\varepsilon_i = Y_i - \widehat{Y_i} = Y_i - (b_0 + b_1 X_i)$$

- The fitted value is equal to:

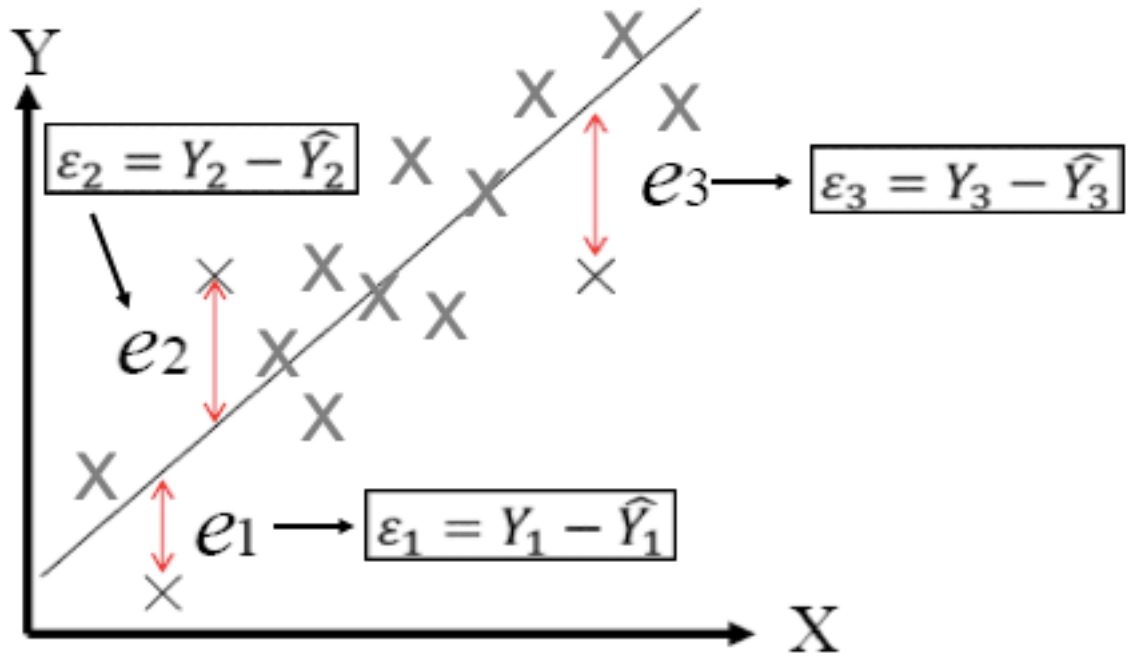
$$\widehat{Y_i} = b_0 + b_1 X_i$$

- **Problem**: It is not possible to minimize the residuals **ε** – they have zero average (they are deviations to the mean value of Y)!

Linear regression model

- **Use the least squares method:**
  it finds the b_0 and b_1 that...
  minimize the sum of the squared residuals (SSR):

b_0 and b_1: $Min \sum \varepsilon_i^2 = Min\,(\varepsilon_1^2+\varepsilon_2^2+\ldots+\varepsilon_n^2)$ $\quad Min \sum (Y_i - \hat{Y}_i)^2 = Min \sum (Y_i - (b_0 + b_1 X_i))^2$



Linear regression model

- And the LR model paramater estimates obtained using the LS method are:

The slope: $$b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{yx}}{S_x^2}$$

The intercept or constant: $$b_0 = \bar{Y} - b_1\bar{X}$$

Linear regression model parameters

Interpretation of the model parameters b_0 and b_1:

- b_0: intercept (or constant) of the regression model – the expected mean value of Y when X=0. It can be zero, but generally the intercept/constant is positive.

- b_1: slope of the regression line - the change in Y for a one-unit change in X (marginal change).
  - $b_1 > 0$: increasing X is associated with an increase in Y. Implies that $S_{YX}>0$.
  - $b_1 < 0$: increasing X is associated with a reduction in Y. Implies that $S_{YX}<0$.
  - $b_1 = 0$: changing X does not affect Y – no relationship between X and Y, $S_{YX}=0$

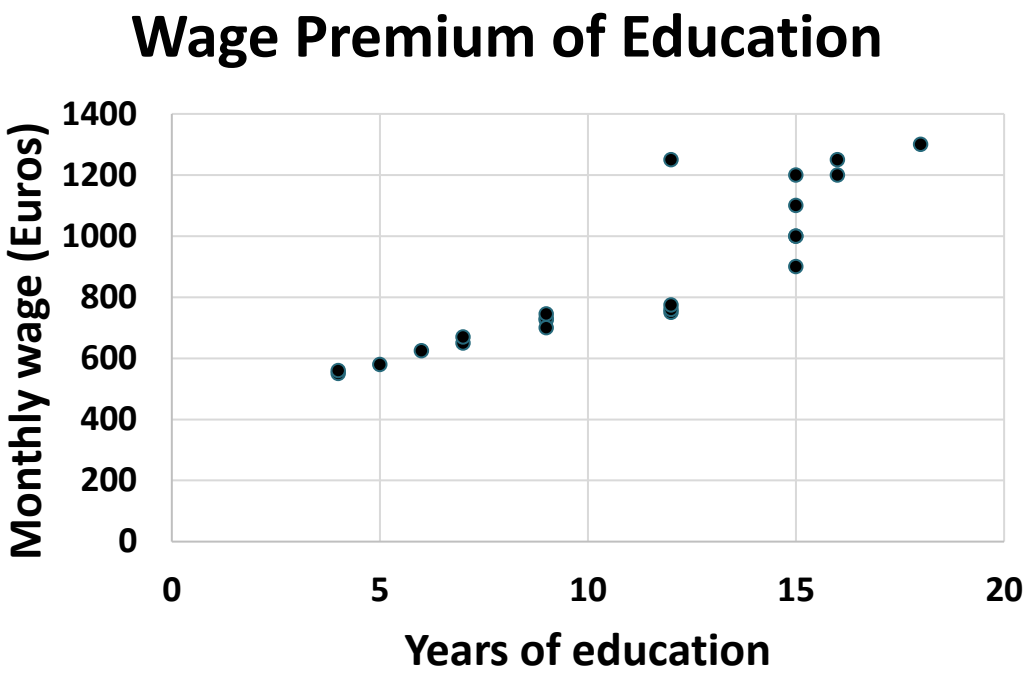
Linear regression model

- What can the model parameters (b_0, b_1, b_j) be used for?
  - Forecast **future** values of Y when we know the future value of X
  - Simulate (or predict) the effect of **alternative** policy variables X (e.g. tax or subsidy) on outcome variables (e.g. demand for public transport, education)

- The **quality** of the forecast or prediction depends on the goodness of fit of the linear regression model.
  - Problem - “rubish in, rubish out”.
  - Problem - correlation is not causation.

- Distinction between “linear in parameters” and “linear in variables”.

What is the wage premium of going to university?



W_i Edu_i

ID	Y	X
	Monthly wage (Eur)	Education (years)
1	550	4
2	560	4
3	580	5
4	650	7
5	670	7
6	725	9
7	625	6
8	750	12
9	760	12
10	730	9
11	745	9
12	775	12
13	900	15
14	1000	15
15	1200	16
16	1300	18
17	1250	16
18	1200	15
19	1250	12
20	700	9
21	1100	15
22	1000	15

What is the wage premium of going to university?

Sample of 22 workers aged between 21-25

- W = monthly wage in Euros
- Edu = number of year of formal education

$$W_i = b_0 + b_1 Edu_i + \varepsilon_i$$

?

ID	W_i → Y	Edu_i → X
	Monthly wage (Eur)	Education (years)
1	550	4
2	560	4
3	580	5
4	650	7
5	670	7
6	725	9
7	625	6
8	750	12
9	760	12
10	730	9
11	745	9
12	775	12
13	900	15
14	1000	15
15	1200	16
16	1300	18
17	1250	16
18	1200	15
19	1250	12
20	700	9
21	1100	15
22	1000	15

What is the wage premium of going to university?

Sample of 22 workers aged between 21-25

- W = monthly wage in Euros
- Edu = number of year of formal education

$$W_i = b_0 + b_1 Edu_i + \varepsilon_i$$

?

$$b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{yx}}{S_x^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

	W_i	Edu_i
ID	Y	X
	Monthly wage (Eur)	Education (years)
1	550	4
2	560	4
3	580	5
4	650	7
5	670	7
6	725	9
7	625	6
8	750	12
9	760	12
10	730	9
11	745	9
12	775	12
13	900	15
14	1000	15
15	1200	16
16	1300	18
17	1250	16
18	1200	15
19	1250	12
20	700	9
21	1100	15
22	1000	15

What is the wage premium of going to university?

Sample of 22 workers aged between 21-25

- W = monthly wage in Euros
- Edu = number of year of formal education

$$W_i = b_0 + b_1 Edu_i + \varepsilon_i$$

$$b_0 = 294$$

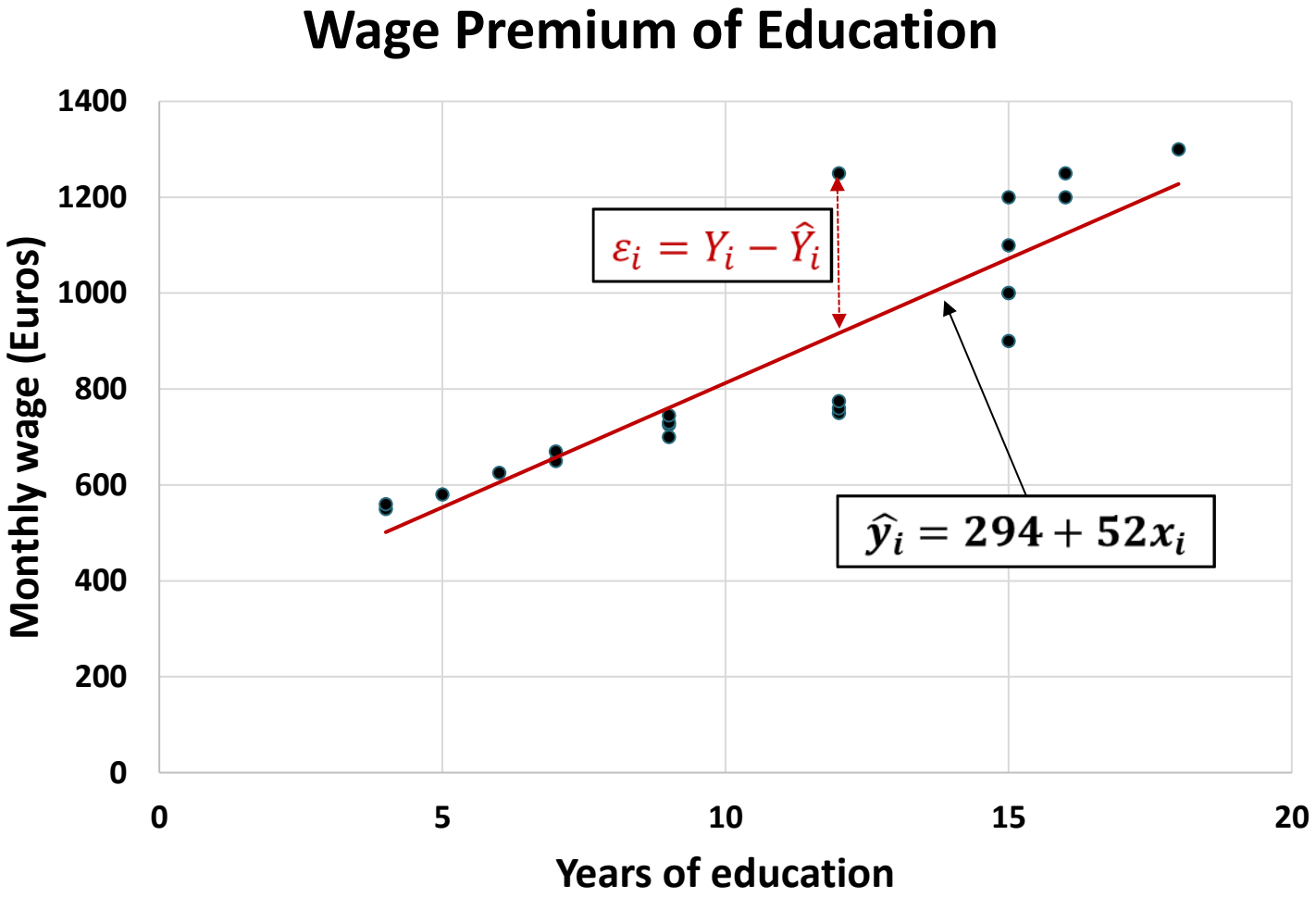
$$b_1 = 52$$

$$\widehat{W_i} = 294 + 52 Edu_i$$

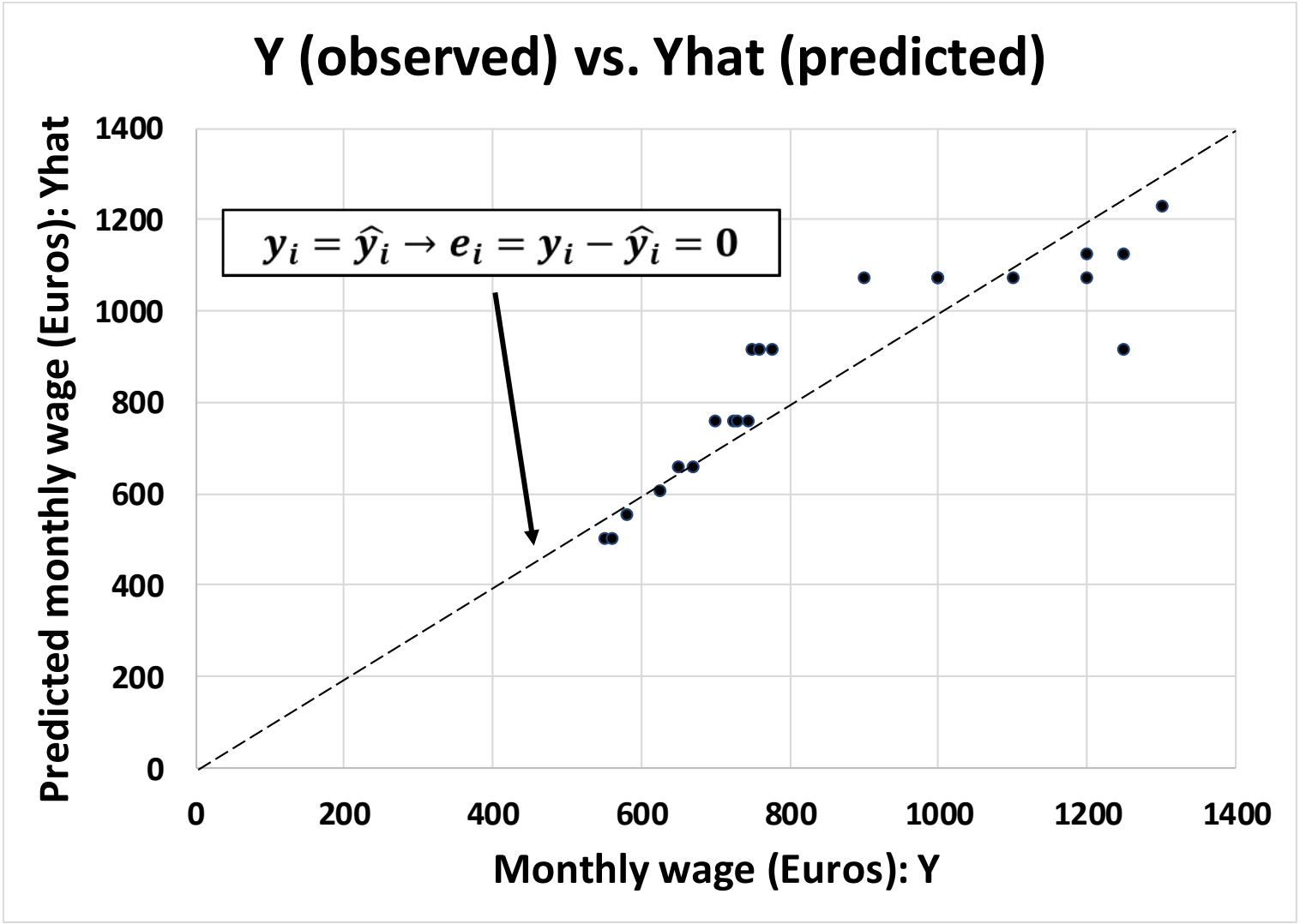
Predicted value of the dependente variable

ID	W_i Y	Edu_i X
	Monthly wage (Eur)	Education (years)
1	550	4
2	560	4
3	580	5
4	650	7
5	670	7
6	725	9
7	625	6
8	750	12
9	760	12
10	730	9
11	745	9
12	775	12
13	900	15
14	1000	15
15	1200	16
16	1300	18
17	1250	16
18	1200	15
19	1250	12
20	700	9
21	1100	15
22	1000	15

What is the wage premium of going to university?



Visualizing model goodness of fit



Visualizing model goodness of fit

W_i Edu_i $\widehat{W}_i = 294 + 52Edu_i$

ID	Y	X	Yhat	Error
	Monthly wage (Eur)	Education (years)	Predicted wage (Eur)	
1	550	4	501	49
2	560	4	501	59
3	580	5	553	27
4	650	7	657	-7
5	670	7	657	13
6	725	9	761	-36
7	625	6	605	20
8	750	12	916	-166
9	760	12	916	-156
10	730	9	761	-31
11	745	9	761	-16
12	775	12	916	-141
13	900	15	1072	-172
14	1000	15	1072	-72
15	1200	16	1124	76
16	1300	18	1228	72
17	1250	16	1124	126
18	1200	15	1072	128
19	1250	12	916	334
20	700	9	761	-61
21	1100	15	1072	28
22	1000	15	1072	-72

$e_i = W_i - \widehat{W}_i$

$0 =< R^2 <= 1$

$R2 = 1- SSR/SST =$

$=1-[(y_i-\hat{y}_i)^2/(y_i- \bar{Y})^2]$

Data Analysis for Economics and Business

Lectures 14 & 15:
Time series: major components of time series analysis;
Measures of change over time
(absolute change, relative change, rates of change)

Academic Year 2023/24

Structure of lecture

1. Time series data
    - Main components
    - Rates of change

2. Measures of change over time:
    - Absolute change
    - Mean absolute change
    - Relative change: rate of change
    - Percentage change vs. percentage points

Learning outcomes

- Understand and explain the main components of time series data
- Define and calculate absolute & relative changes
- Explain the differences between absolute & relative changes
- Explain the difference between percentage changes & percentage points

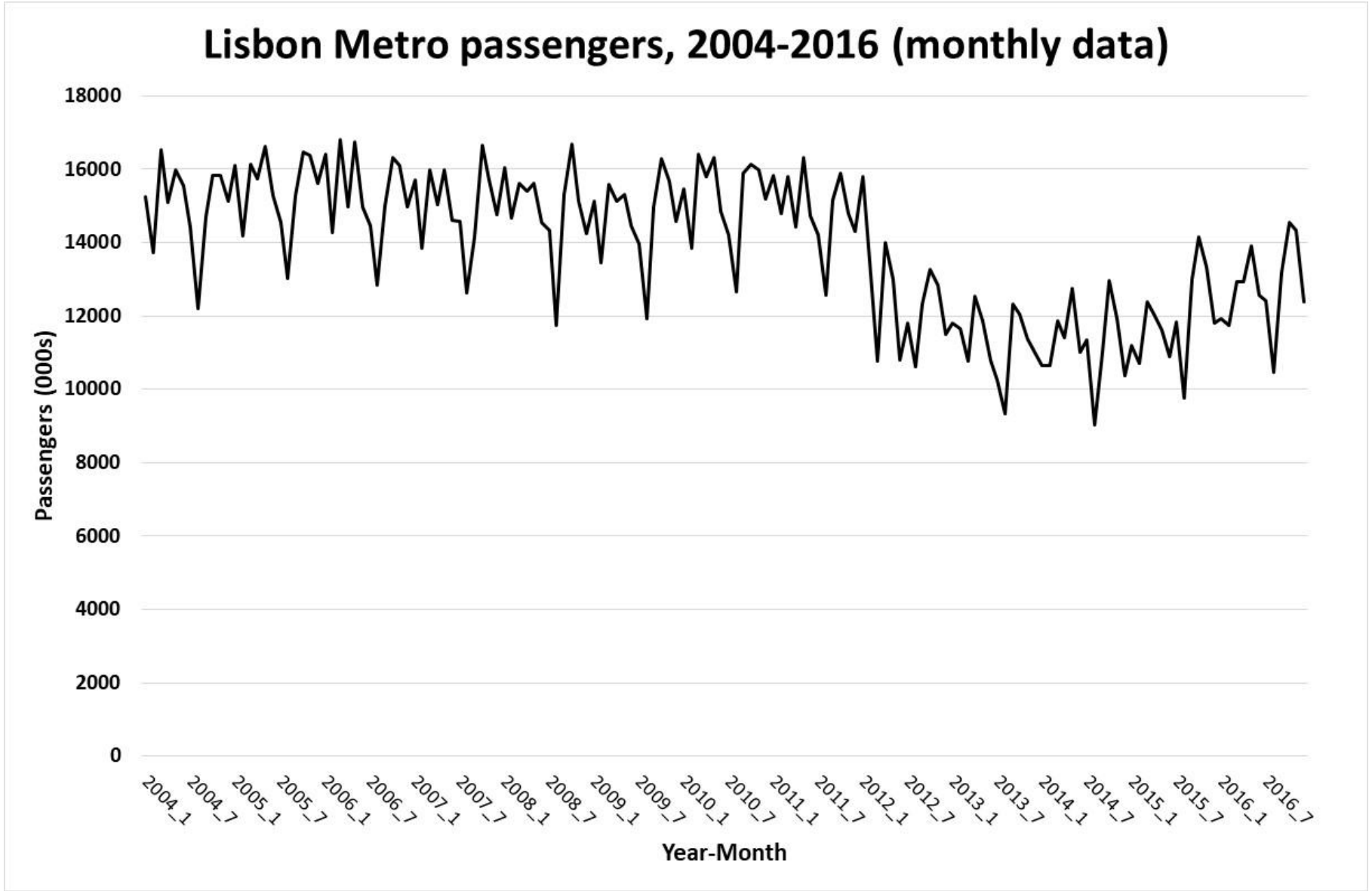
Types of data: Time series data

- **Time series (TS) data**: Multiple time periods for one unit (e.g. Portugal in 1986-2016)
- **Cross-sectional (CS) data**: Multiple units observed only once (e.g. EU countries in 2016)
- **Panel data or longitudinal (PD) data**: Multiple units observed multiple times (e.g. EU countries in 1986-2016)

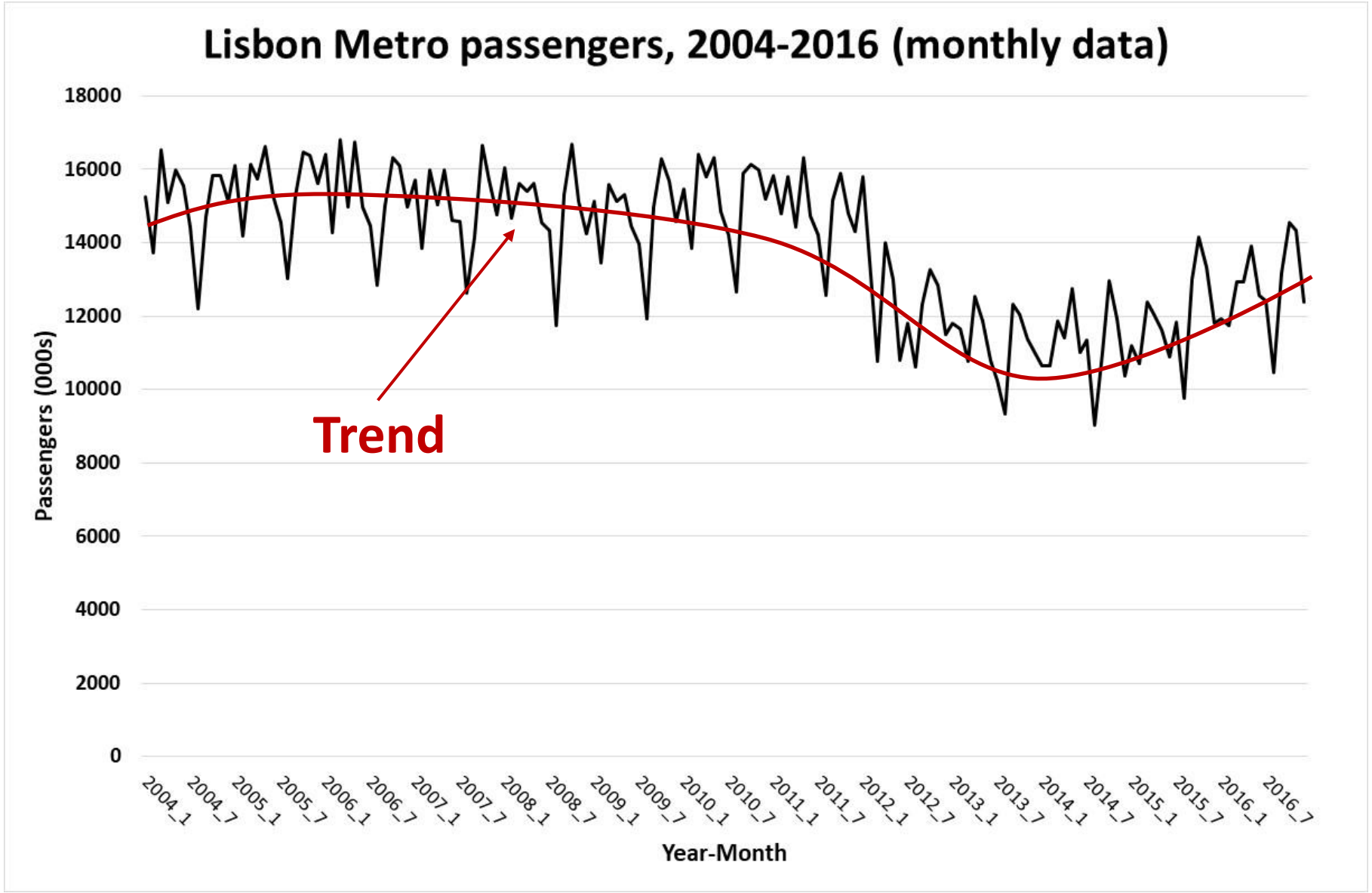
Unemployment rate (%) for people aged 20-24 years old

Country/Year	1986	2016
Germany	8,3	6,7
Spain	44,2	41,4
France	21,1	22,7
Greece	22,9	46,2
Italy	29,6	35,1
Portugal	19,5	26,1
United Kingdom	17,0	9,8

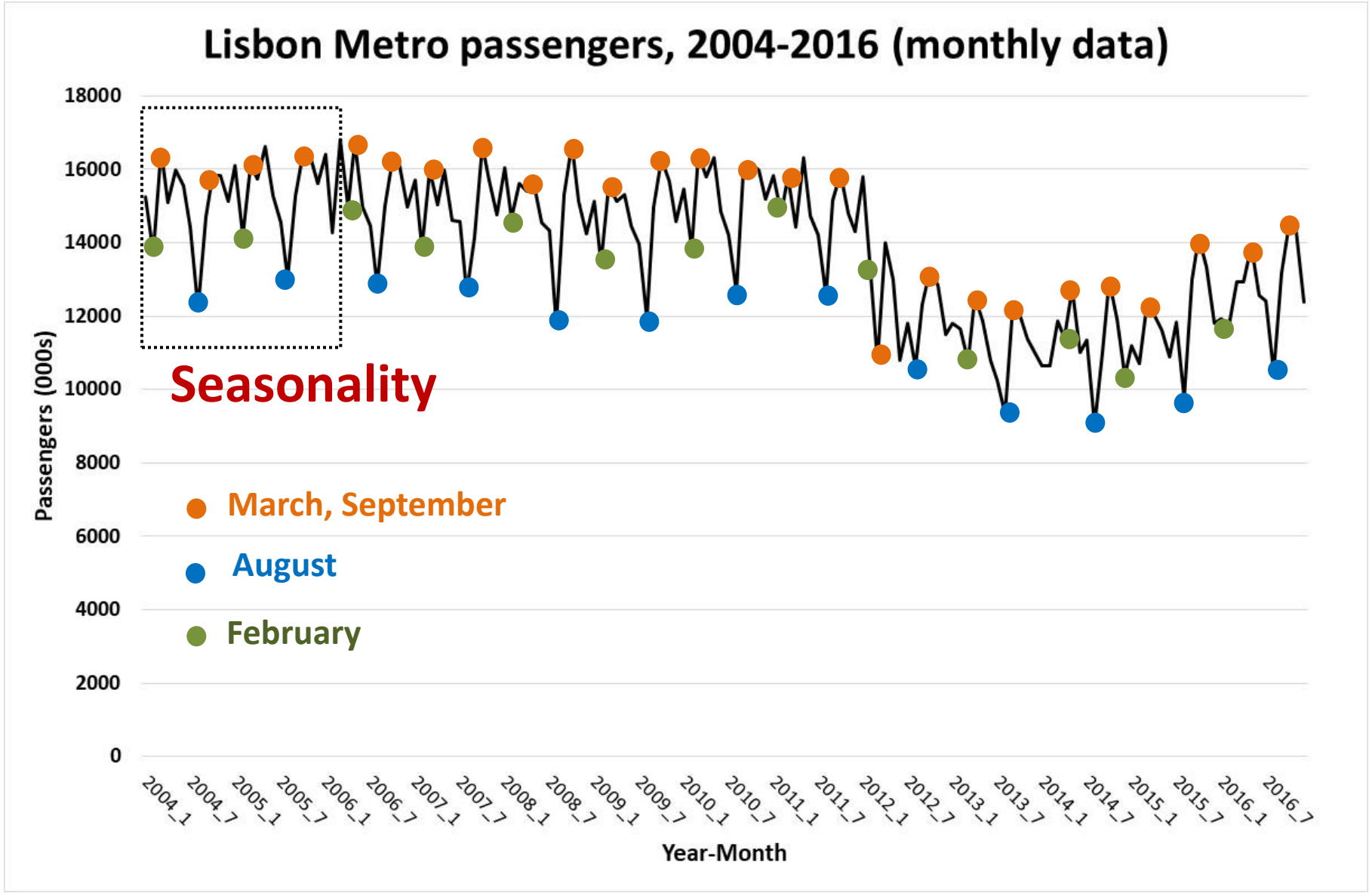
Components of time series data



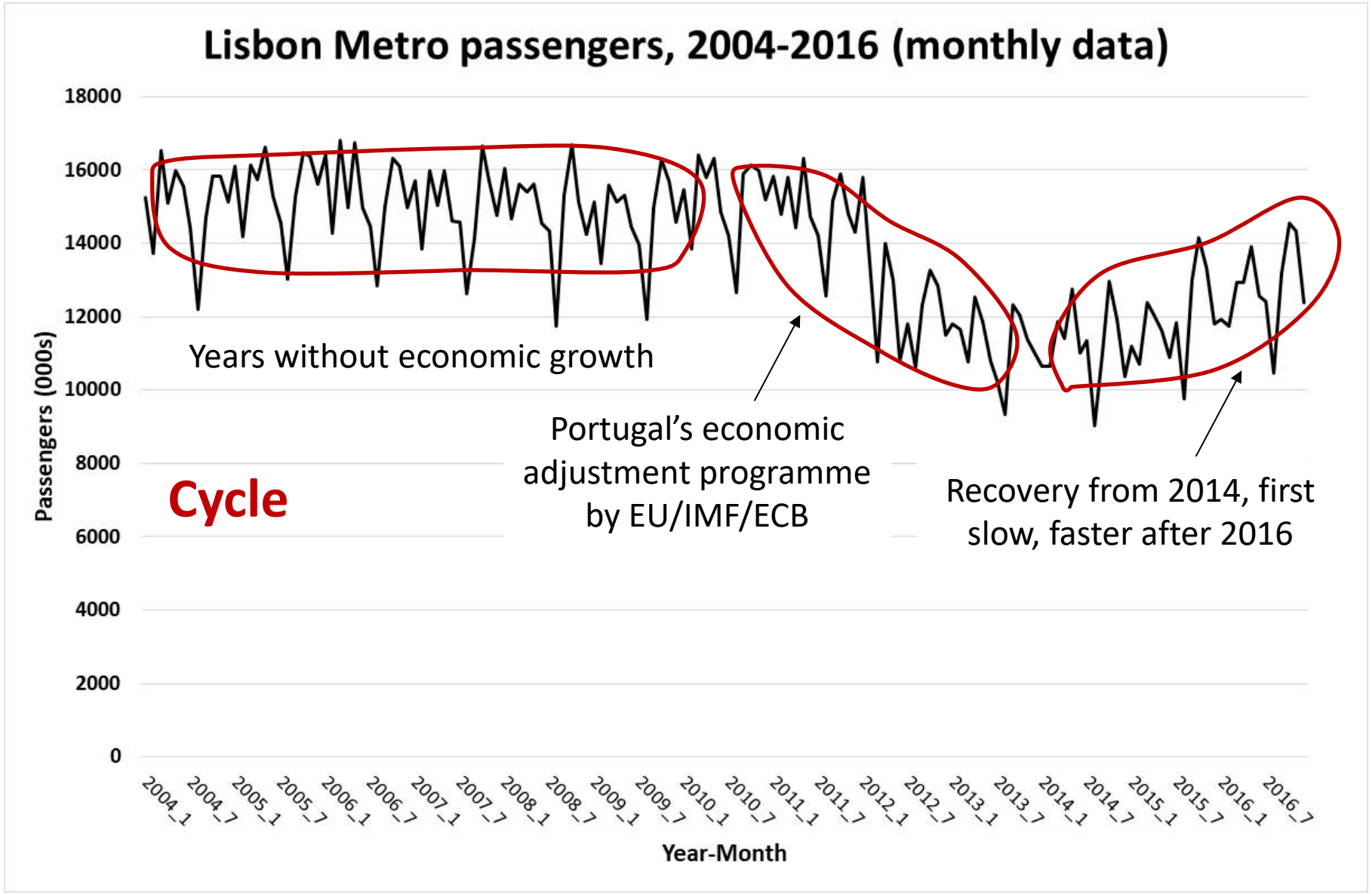
Components of time series data



Components of time series data



Components of time series data



Components of time series data

Four main componentes:

- **Trend** – long term pattern or general trend in the level of the data. Can be linear or non-linear
- **Seasonality** – periodic or regular variation with constant length/duration, generally less than a year; e.g. calendar effects relating to seasons, school holidays
- **Cycle** – periodic variation with length/duration over one year; e.g. business cycles or phases of expansion and recession of economic activity. Length/duration of cycle generally is irregular/not constant
- **Irregular or random** – fluctuations in the data which are not systematic and consist of deviations of the observed data from the other components

Modelling time series components

- There are different models and methods to estimate time series data and using the results to predict or forecast future values. However, this is not part of our course.

Time series rates of change

- **Infra-annual time series:**
  - Chain rate of change – Percentage change compared to the previous period

$$r_{t,s} = \frac{x_{t,s} - x_{t,s-1}}{x_{t,s-1}}, \qquad where\ t:\ year\ \ and\ \ \ s:\ period$$

  - Year-on-year rate of change - Percentage change compared to the same period of the previous year

$$h_{t,s} = \frac{x_{t,s} - x_{t-1,s}}{x_{t-1,s}}, \qquad where\ t:\ year\ \ and\ \ \ s:\ period$$

  - Cumulative year-on-year rate of change:

$$R_{t,s} = \left(\frac{x_{t,s} + x_{t,s-1} + x_{t,s-2} + \cdots}{x_{t-1,s} + x_{t-1,s-1} + x_{t-1,s-2} + \cdots} \right) - 1$$

- Each of the rates allows analysing different aspects of the time series data

Absolute change over time

- Absolute change - change in levels between two points in time

$$\Delta X_{t+k,t} = X_{t+k} - X_t, \qquad with\ k = 1, 2, \ldots \text{(time periods)}$$

Population of Lisbon and Lisbon Metropolitan Area

Year	Lisbon	Lisbon Metropolitan Area (AML)
2016	504 964	2 821 349
2011	537 412	2 827 050
2001	563 149	2 678 695
1992	643 466	2 539 817
1991	656 002	2 539 520

$$\Delta P^{Lis}_{1992,1991} = P^{Lis}_{1992} - P^{Lis}_{1991} \longrightarrow k = 1\ period$$

$$\Delta P^{Lis}_{2016,1991} = P^{Lis}_{2016} - P^{Lis}_{1991} \longrightarrow k = 25\ periods$$

Absolute change over time

- Absolute change - change in levels between two points in time

$$\Delta X_{t+k,t} = X_{t+k} - X_t, \qquad with\ k = 1, 2, \ldots \text{(time periods)}$$

Population of Lisbon and Lisbon Metropolitan Area

Year	Lisbon	Lisbon Metropolitan Area (AML)
2016	504 964	2 821 349
2011	537 412	2 827 050
2001	563 149	2 678 695
1992	643 466	2 539 817
1991	656 002	2 539 520

$\Delta P^{Lis}_{1992,1991} = P^{Lis}_{1992} - P^{Lis}_{1991}$ =643 466-656 002= -12 536 people

$\Delta P^{Lis}_{2016,1991} = P^{Lis}_{2016} - P^{Lis}_{1991}$ =504 964-656 002= -151 038 people

$\Delta P^{AML}_{1992,1991} = P^{AML}_{1992} - P^{AML}_{1991}$ =2 539 817-2 539 520= 297 people

$\Delta P^{AML}_{2016,1991} = P^{AML}_{2016} - P^{AML}_{1991}$ =2 821 349-2 539 520= 281 829 people

Mean absolute change

- Mean absolute change (per unit of time) – Total absolute change divided by the number of periods

$$m\Delta X_{t+k,t} = \frac{X_{t+k} - X_t}{k}, \qquad with\ k = 1, 2, \ldots \text{(time periods)}$$

Population of Lisbon and Lisbon Metropolitan Area

Year	Lisbon	Lisbon Metropolitan Area (AML)
2016	504 964	2 821 349
2011	537 412	2 827 050
2001	563 149	2 678 695
1992	643 466	2 539 817
1991	656 002	2 539 520

$$m\Delta P^{Lis}_{1992,1991} = \frac{P^{Lis}_{1992} - P^{Lis}_{1991}}{1} \longrightarrow k = 1\ period$$

$$m\Delta P^{Lis}_{2016,1991} = \frac{P^{Lis}_{2016} - P^{Lis}_{1991}}{25} \longrightarrow k = 25\ periods$$

Pros & cons of absolute changes

- Advantages:
  - Easy to calculate

- Disadvantages:
  - Different units for different variables (population, employment)
  - The same magnitude of absolute change can have a very different meaning

Year	Country A		Country B	
	GDPpc (Eur)	$\Delta GDPpc_{t+k,t}$	GDPpc (Eur)	$\Delta GDPpc_{t+k,t}$
2014	15 000	-	35 000	-
2015	20 000	5 000	40 000	5 000
2016	25 000	5 000	45 000	5 000

Relative change and rate of change

- **Relative change** between two points in time:

$$r_{t+k,t} = \frac{X_{t+k} - X_t}{X_t} = \frac{\Delta X_{t+k,t}}{X_t} = \frac{X_{t+k}}{X_t} - 1$$

- **Rate of change** generally refers to the relative change **expressed in percentage**:

$$r_{t+k,t}(\%) = \left(\frac{X_{t+k} - X_t}{X_t}\right) * 100 = \left(\frac{\Delta X_{t+k,t}}{X_t}\right) * 100 = \left(\frac{X_{t+k}}{X_t} - 1\right) * 100$$

Relative change and rate of change

Population of Lisbon and Lisbon Metropolitan Area

Year	Lisbon	Lisbon Metropolitan Area (AML)
2016	504 964	2 821 349
2011	537 412	2 827 050
2001	563 149	2 678 695
1992	643 466	2 539 817
1991	656 002	2 539 520

$rP^{Lis}_{1992,1991} = (P^{Lis}_{1992} - P^{Lis}_{1991})/P^{Lis}_{1991}$ = -12 536/656 002= -0.019

The rate of change is -1.9% and is also the annual rate of change

$rP^{Lis}_{2016,1991} = (P^{Lis}_{2016} - P^{Lis}_{1991})/P^{Lis}_{1991}$ = -151 038/656 002= -0.230

The rate of change is -23%

Pros & cons of relative changes

- Disadvantages :
  - Not as easy to calculate as absolute changes

- Advantages:
  - Units are the same for different variables (%)
  - Magnitudes take into account the initial/starting point and are directly comparable

Year	Country A			Country B		
	GDPpc (Eur)	$\Delta_{t+k,t}$	$r_{t+k,t}$ (%)	GDPpc (Eur)	$\Delta_{t+k,t}$	$r_{t+k,t}$ (%)
2014	15 000	-	-	35 000	-	
2015	20 000	5 000	(5 000/15 000)*100=33.3%	40 000	5 000	(5 000/35 000)*100=14.3%
2016	25 000	5 000	(5 000/20 000)*100=25%	45 000	5 000	(5 000/40 000)*100=12.5%

NB! Percentage change vs. percentage points

- **Percentage points (p.p)**: the absolute difference between two rates of change (or any two measures expressed in %)

Year	Country A		
	GDPpc (Eur)	$_{rt+k,t}$ (%)	$\Delta r_{t+k,t}$ (p.p)
2014	15 000	-	-
2015	20 000	33.3%	-
2016	25 000	25%	-8.3

NB! Percentage change vs. percentage points

Change of a value in percentage

20%  30%

50% growth

(relative change)

10 p.p. growth

(absolute change)

Annual rate of change

- Let X_{t+k-1} and X_{t+k} be the values of a given variable X in periods t+k-1 and t+k, respectively.

- note that
  - when k=1 then t+k-1 = t + 1 - 1 = t and t+k = t + 1
  - when k=2 then t+k-1 = t + 2 - 1 = t + 1 and t+k = t + 2
  - when k=3 then t+k-1 = t + 3 - 1 = t + 2 and t+k = t + 3
  - Etc.

- *How do we calculate the annual rate of change between two consecutive years t and t+1, $r_{t+k,t+k-1}$?*

Annual rate of change

- **Annual rate of change** refers to the rate of change between two consecutive periods (years) t and t+1:

$$r_{t+k,t+k-1} = \left(\frac{X_{t+k} - X_{t+k-1}}{X_{t+k-1}}\right) = \left(\frac{\Delta X_{t+k,t+k-1}}{X_{t+k-1}}\right) = \frac{X_{t+k}}{X_{t+k-1}} - 1$$

$$with\ k = 1,2,3, \ldots T\ years$$

Hence any given value X_{t+k} can be written as a function of the value in the previous period X_{t+k-1}. How?

Annual rate of change

- **Annual rate of change** refers to the rate of change between two consecutive periods (years) t and t+1:

$$r_{t+k,t+k-1} = \left(\frac{X_{t+k} - X_{t+k-1}}{X_{t+k-1}}\right) = \left(\frac{\Delta X_{t+k,t+k-1}}{X_{t+k-1}}\right) = \frac{X_{t+k}}{X_{t+k-1}} - 1$$

$$with\ k = 1,2,3, \ldots T\ years$$

Hence any given value X_{t+k} can be written as a function of the value in the previous period X_{t+k-1}. How?

$$X_{t+k} = X_{t+k-1} * (1 + r_{t+k,t+k-1})$$

Annual rate of change

- Let’s see an example…

Annual rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2007-2016

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GDP pc	17181.70	17191.10	16663.20	16971.80	16686.30	16079.20	15985.00	16214.90	16578.90	16887.20
$r_{t+k,t+k-1}$	-	0.00055	-0.03071	0.01852	-0.01682	-0.03638	-0.00586	0.01438	0.02245	0.01860
$r_{t+k,t+k-1}$ (in %)	-	0.055	-3.071	1.852	-1.682	-3.638	-0.586	1.438	2.245	1.860

$X_{08} =$

$X_{09} =$

$X_{10} =$

...

$X_{16} =$

Annual rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2007-2016

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GDP pc	17181.70	17191.10	16663.20	16971.80	16686.30	16079.20	15985.00	16214.90	16578.90	16887.20
$r_{t+k,t+k-1}$	-	0.00055	-0.03071	0.01852	-0.01682	-0.03638	-0.00586	0.01438	0.02245	0.01860
$r_{t+k,t+k-1}$ (in %)	-	0.055	-3.071	1.852	-1.682	-3.638	-0.586	1.438	2.245	1.860

$X_{08} = X_{07} * \left(1 + r_{08/07}\right)$=17 181.70*(1 + 0.00055)=17 191.10

$X_{09} =$

$X_{10} =$

...

$X_{16} =$

Annual rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2007-2016

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GDP pc	17181.70	17191.10	16663.20	16971.80	16686.30	16079.20	15985.00	16214.90	16578.90	16887.20
$r_{t+k,t+k-1}$	-	0.00055	-0.03071	0.01852	-0.01682	-0.03638	-0.00586	0.01438	0.02245	0.01860
$r_{t+k,t+k-1}$ (in %)	-	0.055	-3.071	1.852	-1.682	-3.638	-0.586	1.438	2.245	1.860

$X_{08} = X_{07} * (1 + r_{08/07})$=17 181.70*(1 + 0.00055)=17 191.10

$X_{09} = X_{08} * (1 + r_{09,08})$=

$X_{10} =$

...

$X_{16} =$

Annual rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2007-2016

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GDP pc	17181.70	17191.10	16663.20	16971.80	16686.30	16079.20	15985.00	16214.90	16578.90	16887.20
$r_{t+k,t+k-1}$	-	0.00055	-0.03071	0.01852	-0.01682	-0.03638	-0.00586	0.01438	0.02245	0.01860
$r_{t+k,t+k-1}$ (in %)	-	0.055	-3.071	1.852	-1.682	-3.638	-0.586	1.438	2.245	1.860

$X_{08} = X_{07} * (1 + r_{08/07})$=17 181.70*(1 + 0.00055)=17 191.10

$X_{09} = X_{08} * (1 + r_{09,08})$=$X_{07} * (1 + r_{08,07}) * (1 + r_{09,08})$=

$X_{10} =$

...

$X_{16} =$

Annual rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2007-2016

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GDP pc	17181.70	17191.10	16663.20	16971.80	16686.30	16079.20	15985.00	16214.90	16578.90	16887.20
$r_{t+k,t+k-1}$	-	0.00055	-0.03071	0.01852	-0.01682	-0.03638	-0.00586	0.01438	0.02245	0.01860
$r_{t+k,t+k-1}$ (in %)	-	0.055	-3.071	1.852	-1.682	-3.638	-0.586	1.438	2.245	1.860

$X_{08} = X_{07} * (1 + r_{08/07})$=17 181.70*(1 + 0.00055)=17 191.10

$X_{09} = X_{08} * (1 + r_{09,08})$=$X_{07} * (1 + r_{08,07}) * (1 + r_{09,08})$= 17 181.70*(1 + 0.000555)*(1-0.03071)
=16 663.20

$X_{10} =$

...

$X_{16} =$

Annual rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2007-2016

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GDP pc	17181.70	17191.10	16663.20	16971.80	16686.30	16079.20	15985.00	16214.90	16578.90	16887.20
$r_{t+k,t+k-1}$	-	0.00055	-0.03071	0.01852	-0.01682	-0.03638	-0.00586	0.01438	0.02245	0.01860
$r_{t+k,t+k-1}$ (in %)	-	0.055	-3.071	1.852	-1.682	-3.638	-0.586	1.438	2.245	1.860

$X_{08} = X_{07} * (1 + r_{08/07})$=17 181.70*(1 + 0.00055)=17 191.10

$X_{09} = X_{08} * (1 + r_{09,08})$=$X_{07} * (1 + r_{08,07}) * (1 + r_{09,08})$= 17 181.70*(1 + 0.000555)*(1-0.03071)
=16 663.20

$X_{10} = X_{09} * (1 + r_{10,09})$=

...

$X_{16} =$

Annual rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2007-2016

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GDP pc	17181.70	17191.10	16663.20	16971.80	16686.30	16079.20	15985.00	16214.90	16578.90	16887.20
$r_{t+k,t+k-1}$	-	0.00055	-0.03071	0.01852	-0.01682	-0.03638	-0.00586	0.01438	0.02245	0.01860
$r_{t+k,t+k-1}$ (in %)	-	0.055	-3.071	1.852	-1.682	-3.638	-0.586	1.438	2.245	1.860

$$X_{08} = X_{07} * \left(1 + r_{08/07}\right) = 17\,181.70 * (1 + 0.00055) = 17\,191.10$$

$$X_{09} = X_{08} * \left(1 + r_{09,08}\right) = X_{07} * \left(1 + r_{08,07}\right) * \left(1 + r_{09,08}\right) = 17\,181.70 * (1 + 0.000555) * (1 - 0.03071) = 16\,663.20$$

$$X_{10} = X_{09} * \left(1 + r_{10,09}\right) = X_{07} * \left(1 + r_{08,07}\right) * \left(1 + r_{09,08}\right) * \left(1 + r_{10,09}\right) =$$

...

$$X_{16} =$$

Annual rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2007-2016

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GDP pc	17181.70	17191.10	16663.20	16971.80	16686.30	16079.20	15985.00	16214.90	16578.90	16887.20
$r_{t+k,t+k-1}$	-	0.00055	-0.03071	0.01852	-0.01682	-0.03638	-0.00586	0.01438	0.02245	0.01860
$r_{t+k,t+k-1}$ (in %)	-	0.055	-3.071	1.852	-1.682	-3.638	-0.586	1.438	2.245	1.860

$X_{08} = X_{07} * (1 + r_{08/07})$=17 181.70*(1 + 0.00055)=17 191.10

$X_{09} = X_{08} * (1 + r_{09,08})$=$X_{07} * (1 + r_{08,07}) * (1 + r_{09,08})$= 17 181.70*(1 + 0.000555)*(1-0.03071)
=16 663.20

$X_{10} = X_{09} * (1 + r_{10,09})$=$X_{07} * (1 + r_{08,07}) * (1 + r_{09,08}) * (1 + r_{10,09}) =$
= 17 181.70*(1+0.000555)*(1-0.03071)*(1+0.01852)=16 971.80

…

$X_{16} =$

Annual rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2007-2016

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GDP pc	17181.70	17191.10	16663.20	16971.80	16686.30	16079.20	15985.00	16214.90	16578.90	16887.20
$r_{t+k,t+k-1}$	-	0.00055	-0.03071	0.01852	-0.01682	-0.03638	-0.00586	0.01438	0.02245	0.01860
$r_{t+k,t+k-1}$ (in %)	-	0.055	-3.071	1.852	-1.682	-3.638	-0.586	1.438	2.245	1.860

$X_{08} = X_{07} * (1 + r_{08/07})$=17 181.70*(1 + 0.00055)=17 191.10

$X_{09} = X_{08} * (1 + r_{09,08})$=$X_{07} * (1 + r_{08,07}) * (1 + r_{09,08})$= 17 181.70*(1 + 0.000555)*(1-0.03071)
=16 663.20

$X_{10} = X_{09} * (1 + r_{10,09})$=$X_{07} * (1 + r_{08,07}) * (1 + r_{09,08}) * (1 + r_{10,09}) =$
= 17 181.70*(1+0.000555)*(1-0.03071)*(1+0.01852)=16 971.80

…

$X_{16} = X_{15} * (1 + r_{16,15}) =$

Annual rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2007-2016

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GDP pc	17181.70	17191.10	16663.20	16971.80	16686.30	16079.20	15985.00	16214.90	16578.90	16887.20
$r_{t+k,t+k-1}$	-	0.00055	-0.03071	0.01852	-0.01682	-0.03638	-0.00586	0.01438	0.02245	0.01860
$r_{t+k,t+k-1}$ (in %)	-	0.055	-3.071	1.852	-1.682	-3.638	-0.586	1.438	2.245	1.860

$X_{08} = X_{07} * (1 + r_{08/07})$=17 181.70*(1 + 0.00055)=17 191.10

$X_{09} = X_{08} * (1 + r_{09,08})$=$X_{07} * (1 + r_{08,07}) * (1 + r_{09,08})$= 17 181.70*(1 + 0.000555)*(1-0.03071)
=16 663.20

$X_{10} = X_{09} * (1 + r_{10,09})$=$X_{07} * (1 + r_{08,07}) * (1 + r_{09,08}) * (1 + r_{10,09}) =$
= 17 181.70*(1+0.000555)*(1-0.03071)*(1+0.01852)=16 971.80

…

$X_{16} = X_{15} * (1 + r_{16,15}) = X_{07} * (1 + r_{08,07}) * (1 + r_{09,08}) * (1 + r_{10,09})$*…*$(1 + r_{16,15})$=

Annual rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2007-2016

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
GDP pc	17181.70	17191.10	16663.20	16971.80	16686.30	16079.20	15985.00	16214.90	16578.90	16887.20
$r_{t+k,t+k-1}$	-	0.00055	-0.03071	0.01852	-0.01682	-0.03638	-0.00586	0.01438	0.02245	0.01860
$r_{t+k,t+k-1}$ (in %)	-	0.055	-3.071	1.852	-1.682	-3.638	-0.586	1.438	2.245	1.860

$X_{08} = X_{07} * (1 + r_{08/07})$=17 181.70*(1 + 0.00055)=17 191.10

$X_{09} = X_{08} * (1 + r_{09,08})$=$X_{07} * (1 + r_{08,07}) * (1 + r_{09,08})$= 17 181.70*(1 + 0.000555)*(1-0.03071)
=16 663.20

$X_{10} = X_{09} * (1 + r_{10,09})$=$X_{07} * (1 + r_{08,07}) * (1 + r_{09,08}) * (1 + r_{10,09}) =$
= 17 181.70*(1+0.000555)*(1-0.03071)*(1+0.01852)=16 971.80

…

$X_{16} = X_{15} * (1 + r_{16,15}) = X_{07} * (1 + r_{08,07}) * (1 + r_{09,08}) * (1 + r_{10,09})$*…*$(1 + r_{16,15})$=
= 17 181.70*(1+0.00055)*(1-0.03071)*(1+0.01852)*…*(1+0.01860)=16 887.20

Annual rate of change

- Generalizing (no numbers, only letters)?

Annual rate of change

- Generalizing (no numbers, only letters):

$$X_{t+1} = X_t * (1 + r_{t+1,t})$$

$$X_{t+2} = X_{t+1} * (1 + r_{t+2,t+1}) = X_t * (1 + r_{t+1,t}) * (1 + r_{t+2,t+1})$$

$$X_{t+3} = X_{t+2} * (1 + r_{t+3,t+2}) = X_t * (1 + r_{t+1,t}) * (1 + r_{t+2,t+1}) * (1 + r_{t+3,t+2})$$

...

$$X_{t+k} = X_{t+k-1} * (1 + r_{t+k,t+k-1}) = X_t * (1 + r_{t+1,t}) * (1 + r_{t+2,t+1}) * \cdots * (1 + r_{t+k,t+k-1})$$

$$\boxed{So:\ X_{t+k} = X_t * (1 + r_{t+1,t}) * (1 + r_{t+2,t+1}) * \cdots * (1 + r_{t+k,t+k-1})}$$ Equation (1)

Average rate of change

- Average relative change across multiple time periods, $mr_{t+k,t}$,
  is the rate which if applied *k* times to x_t gives the value of x_{t+k}?

Average rate of change

- Average relative change across multiple time periods, $mr_{t+k,t}$, **is the rate which if applied *k* times to x_t gives the value of x_{t+k}?**

$$X_{t+k} = X_t * \left(1 + mr_{t+k,t}\right)^k$$ Equation (2)

- Re-arranging for $mr_{t+k,t}$?

Average rate of change

- Average relative change across multiple time periods, $mr_{t+k,t}$, **is the rate which if applied *k* times to x_t gives the value of x_{t+k}?**

$$\boxed{X_{t+k} = X_t * \left(1 + mr_{t+k,t}\right)^k}$$ Equation (2)

- Re-arranging for $mr_{t+k,t}$?

$$mr_{t+k,t} = \left(\frac{X_{t+k}}{X_t}\right)^{1/k} - 1$$

Average rate of change

- Average relative change across multiple time periods, $mr_{t+k,t}$, **is the rate which if applied *k* times to x_t gives the value of x_{t+k}?**

$$X_{t+k} = X_t * \left(1 + mr_{t+k,t}\right)^k$$ Equation (2)

- Re-arranging for $mr_{t+k,t}$?

$$mr_{t+k,t} = \left(\frac{X_{t+k}}{X_t}\right)^{1/k} - 1$$

- Example: GDP per capita for Portugal, at 2011 constant prices, 1960-2015

	1960	1981	2001	2011	2015
GDP pc	3463.20	9016.00	16398.30	16686.30	16578.90
$mr_{t+k,t}$	-	0.047	0.039	0.031	0.029
k periods	-	21	41	51	55

$$mr_{2011,1960} =$$

note that t = 1960 and t+k = 2011 hence k = 2011 – 1960 = 51

Average rate of change

- Average relative change across multiple time periods, $mr_{t+k,t}$, **is the rate which if applied *k* times to x_t gives the value of x_{t+k}?**

$$X_{t+k} = X_t * \left(1 + mr_{t+k,t}\right)^k$$ Equation (2)

- Re-arranging for $mr_{t+k,t}$?

$$mr_{t+k,t} = \left(\frac{X_{t+k}}{X_t}\right)^{1/k} - 1$$

- Example: GDP per capita for Portugal, at 2011 constant prices, 1960-2015

	1960	1981	2001	2011	2015
GDP pc	3463.20	9016.00	16398.30	16686.30	16578.90
$mr_{t+k,t}$	-	0.047	0.039	0.031	0.029
k periods	-	21	41	51	55

$$mr_{2011,1960} = \left(\frac{GDP_{2011}}{GDP_{1960}}\right)^{1/51} - 1 = \left(\frac{16\,686.30}{3\,463.20}\right)^{\frac{1}{51}} - 1 = 0.031 \quad \text{or} \quad 3.1\%$$

note that t = 1960 and t+k = 2011 hence k = 2011 – 1960 = 51

Average rate of change

- Setting Eq. (1) = Eq. (2) and solving for $mr_{t+k,t}$

Eq (2): $X_{t+k} = X_t * \left(1 + mr_{t+k,t}\right)^k$

Eq (1): $X_{t+k} = X_t * \left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)$

Average rate of change

- Setting Eq. (1) = Eq. (2) and solving for $mr_{t+k,t}$

Eq (2): $X_{t+k} = X_t * \left(1 + mr_{t+k,t}\right)^k$

Eq (1): $X_{t+k} = X_t * \left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)$

$$\left(1 + mr_{t+k,t}\right)^k = \left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)$$

Average rate of change

- Setting Eq. (1) = Eq. (2) and solving for $mr_{t+k,t}$

Eq (2): $X_{t+k} = X_t * \left(1 + mr_{t+k,t}\right)^k$

Eq (1): $X_{t+k} = X_t * \left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)$

$$\left(1 + mr_{t+k,t}\right)^k = \left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)$$

$$\left(1 + mr_{t+k,t}\right) = \left[\left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)\right]^{1/k}$$

Average rate of change

- Setting Eq. (1) = Eq. (2) and solving for $mr_{t+k,t}$

Eq (2): $X_{t+k} = X_t * \left(1 + mr_{t+k,t}\right)^k$

Eq (1): $X_{t+k} = X_t * \left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)$

$$\left(1 + mr_{t+k,t}\right)^k = \left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)$$

$$\left(1 + mr_{t+k,t}\right) = \left[\left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)\right]^{1/k}$$

$$mr_{t+k,t} = \left[\left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)\right]^{1/k} - 1$$

Average rate of change

- Setting Eq. (1) = Eq. (2) and solving for $mr_{t+k,t}$

Eq (2): $X_{t+k} = X_t * \left(1 + mr_{t+k,t}\right)^k$

Eq (1): $X_{t+k} = X_t * \left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)$

$$\left(1 + mr_{t+k,t}\right)^k = \left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)$$

$$\left(1 + mr_{t+k,t}\right) = \left[\left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)\right]^{1/k}$$

$$mr_{t+k,t} = \left[\left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)\right]^{1/k} - 1$$

$$mr_{t+k,t} = \left[\frac{X_{t+k}}{X_t}\right]^{1/k} - 1 = \left[\frac{X_t * \left(1 + r_{t+k,t}\right)}{X_t}\right]^{1/k} - 1 = \left(1 + r_{t+k,t}\right)^{1/k} - 1$$

Average rate of change

- Setting Eq. (1) = Eq. (2) and solving for $mr_{t+k,t}$

Eq (2): $X_{t+k} = X_t * \left(1 + mr_{t+k,t}\right)^k$

Eq (1): $X_{t+k} = X_t * \left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)$

$$\left(1 + mr_{t+k,t}\right)^k = \left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)$$

$$\left(1 + mr_{t+k,t}\right) = \left[\left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)\right]^{1/k}$$

$$mr_{t+k,t} = \left[\left(1 + r_{t+1,t}\right) * \left(1 + r_{t+2,t+1}\right) * \cdots * \left(1 + r_{t+k,t+k-1}\right)\right]^{1/k} - 1$$

$$mr_{t+k,t} = \left[\frac{X_{t+k}}{X_t}\right]^{1/k} - 1 = \left[\frac{X_t * \left(1 + r_{t+k,t}\right)}{X_t}\right]^{1/k} - 1 = \left(1 + r_{t+k,t}\right)^{1/k} - 1$$

$$so: \left(1 + mr_{t+k,t}\right) = \left(1 + r_{t+k,t}\right)^{1/k}$$

$$\left(1 + mr_{t+k,t}\right)^k = \left(1 + r_{t+k,t}\right)$$

Average rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2014-2016

	2014	2015	2016
GDP pc (euros)	16 214.90	16 578.90	16 887.20
$r_{t+k,t+k-1}$	-		
$r_{t+2,t} = r_{16,14}$	-	-	
$mr_{16,14}=$			
$mr_{16,14}=$			

Average rate of change

- Example: GDP per capita for Portugal, at 2011 constant prices, 2014-2016

	2014	**2015**	**2016**
GDP pc (euros)	16 214.90	16 578.90	16 887.20
$r_{t+k,t+k-1}$	-	0.022	0.019
$r_{t+2,t} = r_{16,14}$	-	-	0.041
$mr_{16,14}=(1+r_{16,14})^{(1/2)}-1$	0.0205		
$mr_{16,14}=((1+r_{16,15})(1+r_{15,14}))^{(1/2)}-1$	0.0205		

Data Analysis for Economics and Business

Lecture 17: Index numbers: simple indices; chain indices, fixed base indices; operations with indices

Academic Year 2023/24

Structure of lecture

- Index numbers
- Chain index and fixed base index
- Operations with index numbers (reversibility, rebasing, circularity)

Learning outcomes

- Define index number
- Relate index numbers with rates of change
- Define and calculate chain indices and fixed base indices
- Explain the **main properties and operations with index numbers (reversibility, rebasing, circularity)**
- Apply the properties of index numbers using chain and fixed base indices (reversibility, rebasing, circularity)

Index number

- **Index number**:

Number that shows the relationship between the value observed in a given time period *t* (for time series data) or statistical unit *i* (for cross-sectional data) and the value observed for that variable in the time period / statistical unit chosen as the reference (reference value or base)

$$i_{i,0} = \frac{X_i}{X_0}$$

$$i_{i,0} = \frac{X_i}{X_0} * 100$$

$$i_{t,0} = \frac{X_t}{X_0}$$

$$i_{t,0} = \frac{X_t}{X_0} * 100$$

- ✓ Values above 1 or 100 if *100 means that $x_t > x_0$
- ✓ Values below 1 or 100 if *100 means that $x_t < x_0$
- ✓ The base of the index is the value taken as reference

Index number

- **What is the relative change between GDP per capita of Lisbon and Portugal?**

$$i_{i,0} = \frac{X_i}{X_0} * 100$$

GDP per capita in 2004, current prices

Regiões	Milhares de euros	Portugal =100
Continente	13,6	99,3
Norte	10,7	78,1
Centro	11,7	85,4
Lisboa	19,3	140,9
Alentejo	12,8	93,4
Algarve	14,1	102,9
R. A. Açores	12,0	87,6
R. A. Madeira	16,6	121,2
PORTUGAL	13,7	100,0

Fonte: INE

Index number

Evolution of Portuguese population

	2000	2001	2002	2003	2004	2005
Population (000s)	10 256,7	10 329,3	10 407,5	10 474,7	10 529,3	10 569,6
Index number (2000=100)	100,0	100,7	101,5	102,1	102,7	103,1

Fonte: INE

- **What is the index number for population between 2000 and 2005?**

Index number

Evolution of Portuguese population

	2000	2001	2002	2003	2004	2005
Population (000s)	10 256,7	10 329,3	10 407,5	10 474,7	10 529,3	10 569,6
Index number (2000=100)	100,0	100,7	101,5	102,1	102,7	103,1

Fonte: INE

- **What is the index number for population between 2000 and 2005?**

$$i_{2005,2000} = \frac{X_{2005}}{X_{2000}} = \frac{10\,569.6}{10\,256.7} = 1.031 \rightarrow so\ it\ is\ 103.1$$

Index number

- **Relationship between an index number and the relative change $r_{t,0}$,** where 0 denotes the reference (or base) period:

$$i_{t,0} = \frac{X_t}{X_0} = \left(1 + r_{t,0}\right)$$

So:

$$\Longmapsto i_{t,0} = 1 + r_{t,0}$$

$$\Longmapsto r_{t,0} = i_{t,0} - 1$$

Index number

Evolution of Portuguese population

	2000	2001	2002	2003	2004	2005
Population (000s)	10 256,7	10 329,3	10 407,5	10 474,7	10 529,3	10 569,6
Index number (2000=100)	100,0	100,7	101,5	102,1	102,7	103,1

Fonte: INE

- **What is the index number for population between 2000 and 2005?**

$$i_{2005,2000} = \frac{X_{2005}}{X_{2000}} = \frac{10\ 569.6}{10\ 256.7} = 1.031 \rightarrow so\ 103.1$$

- **What is the rate of change of population between 2000 and 2005? Can you obtain it directly from the index number?**

Index number

Evolution of Portuguese population

	2000	2001	2002	2003	2004	2005
Population (000s)	10 256,7	10 329,3	10 407,5	10 474,7	10 529,3	10 569,6
Index number (2000=100)	100,0	100,7	101,5	102,1	102,7	103,1

Fonte: INE

- **What is the index number for population between 2000 and 2005?**

$$i_{2005,2000} = \frac{X_{2005}}{X_{2000}} = \frac{10\ 569.6}{10\ 256.7} = 1.031 \rightarrow so\ 103.1$$

- **What is the rate of change of population between 2000 and 2005? Can you obtain it directly from the index number?**

$$i_{2005,2000} = \left(1 + r_{2005,2000}\right) \Rightarrow r_{2005,2000} = 1.031 - 1 = 0.031 \rightarrow 3.1\%$$

Chain Indices and Fixed Base Indices

- **Chain Index**: an index number in which the value at any given period is related to a base in the previous period – **reference value changes over time**

$$i_{1,0}, i_{2,1}, i_{3,2}, \ldots, i_{n,n-1}$$

- **Fixed base index**: an index number for which the base period for the calculations is selected and remains unchanged (i.e. fixed) during the lifetime of the index – **the reference value is always the same**

$$i_{1,0}, i_{2,0}, i_{3,0}, \ldots, i_{n,0}$$

Chain Indices and Fixed Base Indices

Evolution of Portuguese population

	2000	2001	2002	2003	2004	2005
Population (000s)	10 256,7	10 329,3	10 407,5	10 474,7	10 529,3	10 569,6
Index number (2000=100)	100,0	100,7	101,5	102,1	102,7	103,1

Fonte: INE

Chain Index:

$$i_{05,04} = \frac{Pop_{05}}{Pop_{04}} * 100$$

$$i_{04,03} = \frac{Pop_{04}}{Pop_{03}} * 100$$

...

$$i_{01,00} = \frac{Pop_{01}}{Pop_{00}} * 100$$

Fixed base Index:

$$i_{05,00} = \frac{Pop_{05}}{Pop_{00}} * 100$$

$$i_{04,00} = \frac{Pop_{04}}{Pop_{00}} * 100$$

...

$$i_{00,00} = \frac{Pop_{00}}{Pop_{00}} * 100$$

Chain Indices and Fixed Base Indices

Variation in Lisbon metro demand between 2005-2017

year	Passengers (millions)	Passengers - Chain Index	Passengers - Fixed base Index (2005=100)
2017	161.5	105.4	87.1
2016	153.2	107.3	82.6
2015	142.7	101.9	77.0
2014	140.1	99.8	75.6
2013	140.4	91.2	75.7
2012	154.0	86.1	83.1
2011	178.8	97.9	96.4
2010	182.6	103.3	98.5
2009	176.7	99.0	95.3
2008	178.4	99.3	96.2
2007	179.7	97.7	96.9
2006	184.0	99.2	99.2
2005	185.4		100.0

Operations with Indices

- **Indices reversibility**:

$$i_{t,0} = \frac{1}{i_{0,t}}$$

Example:

$$i_{98,94} = \frac{GDP_{98}}{GDP_{94}} * 100 = \frac{1}{\frac{GDP_{94}}{GDP_{98}}} * 100 = \frac{1}{i_{94,98}} * 100$$

Operations with Indices

- **Rebasing**:

$$i_{t,a} = \frac{i_{t,b}}{i_{a,b}}$$

Example:

$$i_{98,94} = \frac{GDP_{98}}{GDP_{94}} * 100 = \frac{\frac{GDP_{98}}{GDP_{90}}}{\frac{GDP_{94}}{GDP_{90}}} * 100 = \frac{i_{98,90}}{i_{94,90}} * 100$$

Operations with Indices

- **Circularity**: a fixed base index can be computed as a product of chain indices

$$i_{2,1} \times i_{1,0} = i_{2,0}$$

$$i_{3,2} \times i_{2,1} \times i_{1,0} = i_{3,0}$$

$$i_{t,t-1} \times \ldots \times i_{3,2} \times i_{2,1} \times i_{1,0} = i_{t,0}$$

Example:

$$i_{98,94} = i_{98,97} * i_{97,96} * i_{96,95} * i_{95,94}$$

$$\frac{GDP_{98}}{GDP_{94}} * 100 = \left(\frac{GDP_{98}}{GDP_{97}} * \frac{GDP_{97}}{GDP_{96}} * \frac{GDP_{96}}{GDP_{95}} * \frac{GDP_{95}}{GDP_{94}}\right) * 100$$

Data Analysis for Economics and Business

Lecture 18: Index numbers – Composite Indices
Current prices vs. Constant prices, Adjusting for inflation

Academic Year 2023/24

Structure of lecture

- Composite indices: why do we need them?
- Laspeyres and Paasche composite indices
- Budget coefficients (or budget weights)
- Constructing the Consumer Price Index (CPI)
- Adjusting for inflation

Learning outcomes

- Explain the need for composite indices in economics and business
- Define & compare Laspeyres & Paasche indices of prices & quantities
- Compute Laspeyres and Paasche indices
- Explain the meaning of budget weights and how to compute them
- Compute, and explain, the consumer price index (CPI)
- Understand how to adjust for inflation and how to relate nominal (i.e. current) prices to real (i.e. constant) prices

Index numbers – recap…

- Index numbers are ***relative numbers*** which allow comparing values of the same variable over time (or/and between regions). We say they are ***relative numbers*** because we take each value of a given variable in relation to a reference or baseline value (e.g. the initial year)

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

- Relation between ***index numbers*** and ***relative change/rate of change***:

$$i_{t,0} = \left(1 + r_{t,0}\right) * 100$$

- Recall distinction ***chain*** vs. ***fixed-base*** index numbers.
- Operations with index numbers: ***reversibility***, ***rebasing***, ***circularity***.

Simple vs. Composite Indices

- We have used simple index numbers to study one single variable or phenomenon, e.g. prices, metro passengers, population, etc.

- To study ***composite*** variables/phenomena, we can use ***composite*** index numbers.

- Popular examples of composite indices include:
  - Consumer price index (CPI) – for a sample of consumer goods and services
  - Retail price index (RPI) – for a sample of retail goods and services
  - Industrial production index (IPI) – for output from manufacturing, mining, electric and gas industries

- **Composite indices** enable us to assess **the evolution of:**
  - **quantities** consumed of different goods
  - **prices** of different goods

Simple vs. Composite Indices

Example: consumer expenditure in food goods i.

i	units		period 0		period 1		period 2	
	quant	price	price	quant	price	quant	price	quant
apples	Kg apples	EUR/Kg apple	1,00	5	0,95	4	0,85	3
milk	litre milk	EUR/Litre milk	0,75	3	0,75	4	0,75	5
meat	Kg meat	EUR/Kg meat	9,98	2	10,47	2,1	10,97	2,2
expenditure		EUR		27,18		28,78		30,43

How can we assess the evolution of different goods expressed in different units and with varying unit prices?

Simple vs. Composite Indices

- Consider the notation below, where:
  q= quantities, **p**= prices
  j=1, 2, …, m denotes the goods or variables to be aggregated

  In period *0*:

  - Quantities consumed: $q_0^1,\ldots,q_0^j,\ldots,q_0^m$

  - Prices: $p_0^1,\ldots,p_0^j,\ldots,p_0^m$

  In period *t*:

  -Quantities consumed: $q_t^1,\ldots,q_t^j,\ldots,q_t^m$

  -Prices: $p_t^1,\ldots,p_t^j,\ldots,p_t^m$

- How to compare the quantities/prices consumed for different goods across different periods? **We can use aggregate composite indices to bring together vectors of quantities and prices**

Simple vs. Composite Indices

- How to transform vectors of quantities and prices into scalars?
  Computing expenditure in a given basket of goods:

  Expenditure in *period 0:* $\sum_{j=1}^{m}\left(q_0^j \times p_0^j\right) = (q_0^1 * p_0^1) + (q_0^2 * p_0^2) + \ \ldots + (q_0^m * p_0^m)$

  Expenditure in *period t:* $\sum_{j=1}^{m}\left(q_t^j \times p_t^j\right) = (q_t^1 * p_t^1) + (q_t^2 * p_t^2) + \ \ldots + (q_t^m * p_t^m)$

- We can compute the index for expenditure evolution – a ***value index***, which reflects both changes of prices and in quantities

$$\frac{\sum_{j=1}^{m}\left(q_t^j \times p_t^j\right)}{\sum_{j=1}^{m}\left(q_0^j \times p_0^j\right)}$$

Simple vs. Composite Indices

- How to compare values from different variables, in different units?

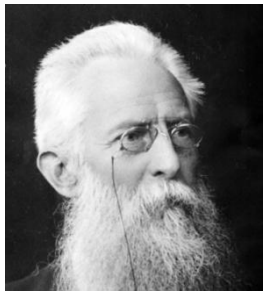
  Example: consumer expenditure in food goods i.

i	units		period 0		period 1		period 2	
	quant	price	price	quant	price	quant	price	quant
apples	Kg apples	EUR/Kg apple	1,00	5	0,95	4	0,85	3
milk	litre milk	EUR/Litre milk	0,75	3	0,75	4	0,75	5
meat	Kg meat	EUR/Kg meat	9,98	2	10,47	2,1	10,97	2,2
expenditure		EUR		27,18		28,78		30,43

- Simple index of prices - for apples (Eur per kilogram): $\left(P_a^1/P_a^0\right)$*100
- Simple index of quantity - for apples (kg): $\left(Q_a^1/Q_a^0\right)$*100
- Index of value or expenditure - for apples $\left(Q_a^1 * P_a^1\right)/Q_a^0 * P_a^0\right)$*100
(expenditure in eur=kg*Eur):
- Composite value or expenditure index for all foods $\left[\left(\sum_i Q_i^1 * P_i^1\right)/\left(\sum_i Q_i^0 * P_i^0\right)\right]$*100
(expenditure in eur):

Laspeyres and Paasche indices

- There are two very popular composite indices in economics: Laspeyres and Paasche indices.



Etienne Laspeyres

Hermann Paasche

- They are used to monitor changes in prices, and quantities, over time and allow separating out the effect of inflation (nominal vs. real prices).

- They are calculated in a very similar way, the only BIG difference is the time period considered as the base or reference.
  - Laspeyres indices use period 0 (initial period) as the base
  - Paasche indices use period *t* (current period) as the base

Laspeyres Price Index (i)

- **Laspeyres indices use period 0 (initial period) as the base:**

  Laspeyres composite index of prices measures the **change in the prices** of the goods in a given basket between period 0 and period t, assuming quantitites consumed remain the same as in period 0.

$$L^{P}_{t,0} = \frac{\sum_{j=1}^{m}\left(q^{j}_{0} \times p^{j}_{t}\right)}{\sum_{j=1}^{m}\left(q^{j}_{0} \times p^{j}_{0}\right)}$$

Expenditure buying the basket consumed in period *0* at the prices of period *t*

Expenditure buying the basket consumed in period *0* at the prices of period 0

The **Laspeyres composite index of prices uses q_0 quantities as the base**.

Laspeyres Price Index (ii)

- **Laspeyres composite index of prices** can be computed as the **weighted average of the simple indices of prices**:

$$L_{t,0}^{P} = \frac{\sum_{j=1}^{m} p_t^j \cdot q_0^j}{\sum_{j=1}^{m} p_0^j \cdot q_0^j} = \frac{\sum_{j=1}^{m} p_0^j \cdot q_0^j \frac{p_t^j}{p_0^j}}{\sum_{j=1}^{m} p_0^j \cdot q_0^j} = \sum_{j=1}^{m} \frac{p_0^j \cdot q_0^j \left(\frac{p_t^j}{p_0^j}\right)}{\sum_{j=1}^{m} p_0^j \cdot q_0^j} = \sum_{j=1}^{m} w_0^j \left(\frac{p_t^j}{p_0^j}\right)$$

with: $w_0^j = \frac{p_0^j \cdot q_0^j}{\sum_{j=1}^{m} p_0^j \cdot q_0^j}$ and $\sum_{j=1}^{m} w_0^j = 1$

- The **weight w_0 of each simple index has an economic meaning** – it is the **budget coefficient or budget weight** – **how much each good weighs in total expenditure in period 0**

Paasche Price Index (i)

- **Paasche indices use period t (current period) as the base period:**

  Paasche composite index of prices measures the **change in the prices of the goods** in a given basket between period 0 and period t, assuming the quantitites consumed in period t.

$$P_{t,0}^{P} = \frac{\sum_{j=1}^{m}\left(q_t^j \times p_t^j\right)}{\sum_{j=1}^{m}\left(q_t^j \times p_0^j\right)}$$

Expenditure buying the basket consumed in period *t* at the prices of period *t*

Expenditure buying the basket consumed in period *t* at the prices of period 0

The **Paasche composite index of prices uses q_t quantities as the base**.

Paasche Price Index (ii)

- The **Paasche composite index of prices** can be computed as the **weighted average of the simple indices of prices**:

$$P_{t,0}^{P} = \frac{\sum_{j=1}^{m} p_t^j . q_t^j}{\sum_{j=1}^{m} p_0^j . q_t^j} = \frac{\sum_{j=1}^{m} p_t^j . q_t^j . \frac{p_0^j}{p_0^j}}{\sum_{j=1}^{m} p_0^j . q_t^j} = \sum_{j=1}^{m} \frac{p_0^j . q_t^j \left(\frac{p_t^j}{p_0^j}\right)}{\sum_{j=1}^{m} p_0^j . q_t^j} = \sum_{j=1}^{m} w_t^j \left(\frac{p_t^j}{p_0^j}\right)$$

with: $w_t^j = \frac{p_0^j . q_t^j}{\sum_{j=1}^{m} p_0^j . q_t^j}$ and $\sum_{j=1}^{m} w_t^j = 1$

- The **weight w_t of each simple index has an economic meaning** – it is the **budget coefficient or budget weight** – **how much each good weights in total expenditure at quantities of current period t and prices p_0**

Laspeyres Quantity Index (i)

- **Laspeyres indices use period 0 (initial period) as the base period:**

  Laspeyres composite index of quantities measures the **change in the quantities** of goods consumed between period 0 and period t, assuming prices remain the same as in period 0.

$$L^{Q}_{t,0} = \frac{\sum_{j=1}^{m}\left(q_t^j \times p_0^j\right)}{\sum_{j=1}^{m}\left(q_0^j \times p_0^j\right)}$$

Expenditure buying the basket consumed in period *t* at prices of period *0*

Expenditure buying the basket consumed in period *0* at prices of period 0

The **Laspeyres composite index of quantities uses p_0 prices as the base**.

Laspeyres Quantity Index (ii)

- **Laspeyres composite index of quantities** can be computed as the **weighted average of the simple indices of quantities**:

$$L_{t,0}^{Q} = \frac{\sum_{j=1}^{m} p_0^j \cdot q_t^j}{\sum_{j=1}^{m} p_0^j \cdot q_0^j} = \frac{\sum_{j=1}^{m} p_0^j \cdot q_t^j \cdot \frac{q_0^j}{q_0^j}}{\sum_{j=1}^{m} p_0^j \cdot q_0^j} = \frac{\sum_{j=1}^{m} p_0^j \cdot q_0^j \left(\frac{q_t^j}{q_0^j}\right)}{\sum_{j=1}^{m} p_0^j \cdot q_0^j} = \sum_{j=1}^{m} w_0^j \left(\frac{q_t^j}{q_0^j}\right)$$

with: $w_0^j = \frac{p_0^j \cdot q_0^j}{\sum_{j=1}^{m} p_0^j \cdot q_0^j}$ and $\sum_{j=1}^{m} w_0^j = 1$

- The **weight w_0 of each simple index has an economic meaning** – it is the **budget coefficient or budget weight** – **how much each good weights in total expenditure in period 0**

Paasche Quantity Index (i)

- **Paasche indices use period t (current period) as the base period:**

  Paasche composite index of quantities measures the **change in the quantities** of goods consumed between period 0 and period t, assuming the prices of goods in period t.

$$P_{t,0}^{Q} = \frac{\sum_{j=1}^{m}\left(q_t^j \times p_t^j\right)}{\sum_{j=1}^{m}\left(q_0^j \times p_t^j\right)}$$

Expenditure buying the basket consumed in period *t* at the prices of period *t*

Expenditure buying the basket consumed in period *0* at the prices of period t

The **Paasche composite index of quantities uses p_t prices as the base**.

Paasche Quantity Index (ii)

- The **Paasche composite index of quantities** can be computed as the **weighted average of the simple indices of quantities**:

$$P_{t,0}^{Q} = \frac{\sum_{j=1}^{m} p_t^j . q_t^j}{\sum_{j=1}^{m} p_t^j . q_0^j} = \frac{\sum_{j=1}^{m} p_t^j . q_t^j . \frac{q_0^j}{q_0^j}}{\sum_{j=1}^{m} p_t^j . q_0^j} = \sum_{j=1}^{m} \frac{p_t^j . q_0^j . \left(\frac{q_t^j}{q_0^j}\right)}{\sum_{j=1}^{m} p_t^j . q_0^j} = \sum_{j=1}^{m} w_t^j . \left(\frac{q_t^j}{q_0^j}\right)$$

with: $w_t^j = \frac{p_t^j . q_0^j}{\sum_{j=1}^{m} p_t^j . q_0^j}$ and $\sum_{j=1}^{m} w_t^j = 1$

- The **weight w_t of each simple index has an economic meaning** – it is the **budget coefficient or budget weight** – **how much each good weights in total expenditure at prices of current period t and quantities q_0**

Tutorial

Goods	Units	2013		2014		2015		2016	
		price	quantity	price	quantity	price	quantity	price	quantity
Apples	eur/kg	1,00	10	1,15	11	1,25	12	1,35	14
Milk	eur/litre	0,50	100	0,65	105	0,70	110	0,80	115
Shoes	eur/pair	20,00	4	25,00	4	30,00	5	40,00	5
Meat	eur/kg	9,00	20	8,50	21	9,50	22	10,00	24
Total expenditure		**320,00**		**359,40**		**451,00**		**550,90**	

- The variables (goods) are expressed in different units (kg, litres, pairs) so we cannot add up the values across goods.
- However, we can aggregate the expenditure on each good and all goods – value index.
- We can use the value indices to study the change in prices and quantities over time: Laspeyres & Paasche composite indices

Tutorial

Goods	Units	2013		2014		2015		2016	
		price	quantity	price	quantity	price	quantity	price	quantity
Apples	eur/kg	1,00	10	1,15	11	1,25	12	1,35	14
Milk	eur/litre	0,50	100	0,65	105	0,70	110	0,80	115
Shoes	eur/pair	20,00	4	25,00	4	30,00	5	40,00	5
Meat	eur/kg	9,00	20	8,50	21	9,50	22	10,00	24
Total expenditure		**320,00**		**359,40**		**451,00**		**550,90**	

Laspeyres composite index of prices				
Expenditure at current prices and initial quantities: **Sum(q_0*p_t)**	320,00	346,50	392,50	453,50
Value or expenditure index: **Sum(q_0*p_t)/Sum(q_0*p_0)*100**	100,00	108,28	122,66	141,72
Percentage change	-	8,28	22,66	41,72

Laspeyres composite index of quantities				
Expenditure at initial prices and period t quantities: **Sum(p_0*qt)**	320,00	332,50	365,00	387,50
Value or expenditure index: **Sum(p_0*q_t)/Sum(q_0*p_0)*100**	100,00	103,91	114,06	121,09
Percentage change	-	3,91	14,06	21,09

Prices in 2014 increased 8,28% in relation to 2013.

Quantities in 2014 increased 3,91% in relation to 2013.

Tutorial

Goods	Units	2013		2014		2015		2016	
		price	quantity	price	quantity	price	quantity	price	quantity
Apples	eur/kg	1,00	10	1,15	11	1,25	12	1,35	14
Milk	eur/litre	0,50	100	0,65	105	0,70	110	0,80	115
Shoes	eur/pair	20,00	4	25,00	4	30,00	5	40,00	5
Meat	eur/kg	9,00	20	8,50	21	9,50	22	10,00	24
Total expenditure		**320,00**		**359,40**		**451,00**		**550,90**	

Paasche composite index of prices				
Expenditure at initial prices and current period quantities: **Sum($qt*p_0$)**	320,00	332,50	365,00	387,50
Value or expenditure index: **Sum($qt*pt$)/Sum($qt*p_0$)*100**	100,00	108,09	123,56	142,17
Percentage change	-	8,09	23,56	42,17

Paasche composite index of quantities				
Expenditure at period t prices (current period) and initial quantities: **Sum(q_0*pt)**	320,00	346,50	392,50	453,50
Value or expenditure index: **Sum($qt*pt$)/Sum(q_0*p_t)*100**	100,00	103,72	114,90	121,48
Percentage change	-	3,72	14,90	21,48

Prices in 2014 increased 8,09% in relation to 2013.

Quantities in 2014 increased 3,72% in relation to 2013.

summary

Laspeyres and Paasche Indices

	Laspeyres	Paasche
Price Index	$L_{t,0}^{P} = \frac{\sum_{j=1}^{m} p_t^j \cdot q_0^j}{\sum_{j=1}^{m} p_0^j \cdot q_0^j} = \sum_{j=1}^{m} w_0^j \cdot \left(\frac{p_t^j}{p_0^j}\right)$	$P_{t,0}^{P} = \frac{\sum_{j=1}^{m} p_t^j \cdot q_t^j}{\sum_{j=1}^{m} p_0^j \cdot q_t^j} = \sum_{j=1}^{m} w_t^j \cdot \left(\frac{p_t^j}{p_0^j}\right)$
Quantity Index	$L_{t,0}^{Q} = \frac{\sum_{j=1}^{m} p_0^j \cdot q_t^j}{\sum_{j=1}^{m} p_0^j \cdot q_0^j} = \sum_{j=1}^{m} w_0^j \cdot \left(\frac{q_t^j}{q_0^j}\right)$	$P_{t,0}^{Q} = \frac{\sum_{j=1}^{m} p_t^j \cdot q_t^j}{\sum_{j=1}^{m} p_t^j \cdot q_0^j} = \sum_{j=1}^{m} w_t^j \cdot \left(\frac{q_t^j}{q_0^j}\right)$

Laspeyres and Paasche Indices

- **Budget weights:**

Laspeyres Indices: with: $w_0^j = \frac{p_0^j \cdot q_0^j}{\sum_{j=1}^{m} p_0^j \cdot q_0^j}$ and $\sum_{j=1}^{m} w_0^j = 1$

Paasche Price Index: with: $w_t^j = \frac{p_0^j \cdot q_t^j}{\sum_{j=1}^{m} p_0^j \cdot q_t^j}$ and $\sum_{j=1}^{m} w_t^j = 1$

Paasche Quantity Index: with: $w_t^j = \frac{p_t^j \cdot q_0^j}{\sum_{j=1}^{m} p_t^j \cdot q_0^j}$ and $\sum_{j=1}^{m} w_t^j = 1$

Laspeyres vs. Paasche indices

- **Advantages of Laspeyres index**:
  - weights w_0 are only needed for one year, the base or initial year; hence it is “cheaper” to compute;
  - indices for each year can be compared directly because the base year w_0 is fixed.
- **Disadvantages of Laspeyres index**:
  - base year weights w_0 can become quickly outdated, especially for goods/services for which demand can change fast;
  - tends to overestimate price increases.

- **Advantages of Paasche index**:
  - by using current-period weights w_t it takes account of changes in patterns of consumption.
- **Disadvantages of Paasche index**:
  - more time and cost demanding to update weights (each year);
  - indices cannot be compared directly each year because weights change;
  - tends to underestimate price increases.

the CPI & inflation

Constructing the CPI

- The consumer price index (CPI) measures the price of goods and services of a “typical” consumer or household “shopping basket”.

- It is **based on the Laspeyres price index**.

- By comparing the cost of the typical basket in current periot t with some base period 0 or t-1 we measure the change in prices – **inflation, change in purchasing power** (and partly the cost of living)

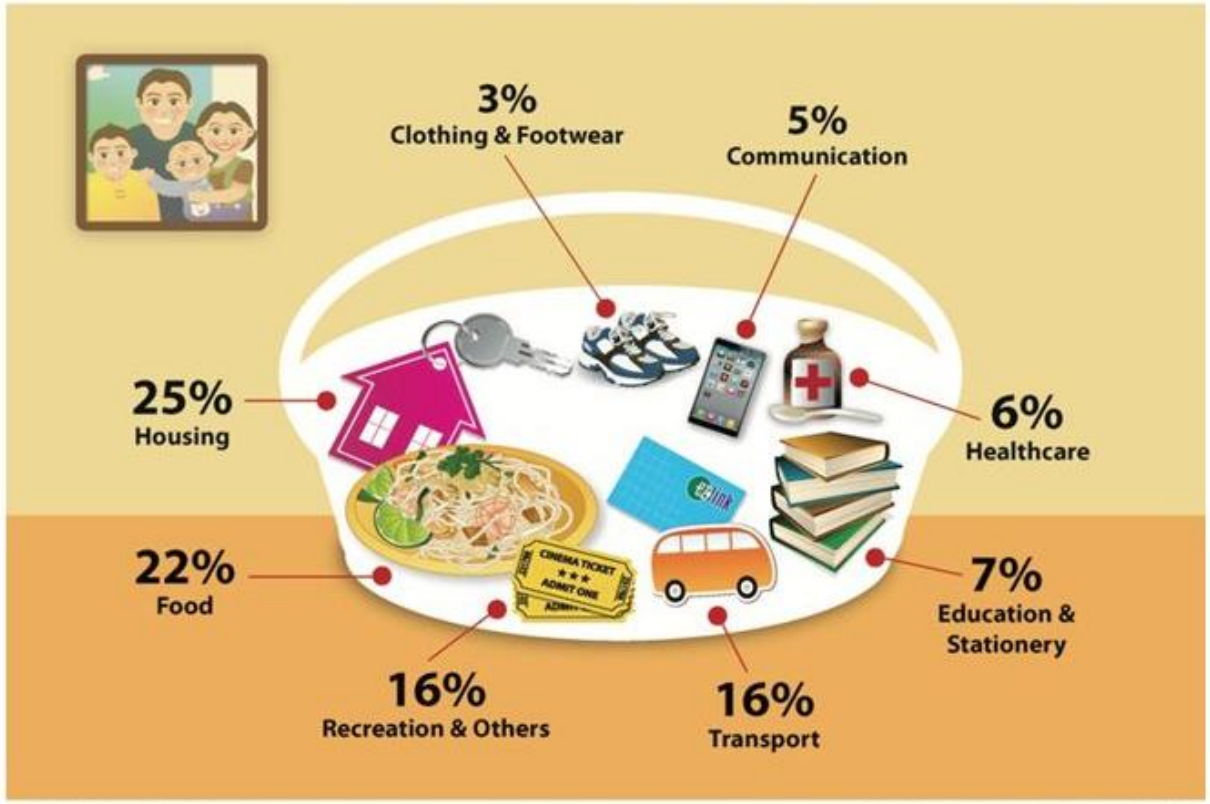
Laspeyres and Paasche Indices

	Laspeyres	Paasche
Price Index	$L^P_{t,0} = \frac{\sum_{j=1}^{m} p_t^j \cdot q_0^j}{\sum_{j=1}^{m} p_0^j \cdot q_0^j} = \sum_{j=1}^{m} w_0^j \cdot \left(\frac{p_t^j}{p_0^j}\right)$	$P^P_{t,0} = \frac{\sum_{j=1}^{m} p_t^j \cdot q_t^j}{\sum_{j=1}^{m} p_0^j \cdot q_t^j} = \sum_{j=1}^{m} w_t^j \cdot \left(\frac{p_t^j}{p_0^j}\right)$
Quantity Index	$L^Q_{t,0} = \frac{\sum_{j=1}^{m} p_0^j \cdot q_t^j}{\sum_{j=1}^{m} p_0^j \cdot q_0^j} = \sum_{j=1}^{m} w_0^j \cdot \left(\frac{q_t^j}{q_0^j}\right)$	$P^Q_{t,0} = \frac{\sum_{j=1}^{m} p_t^j \cdot q_t^j}{\sum_{j=1}^{m} p_t^j \cdot q_0^j} = \sum_{j=1}^{m} w_t^j \cdot \left(\frac{q_t^j}{q_0^j}\right)$

Constructing the CPI

- Brief overview of the CPI procedure:
  1. Select a base year or starting year
  2. Select the goods and services in the “typical shopping basket”
  3. Collect price data for items in basket
  4. Compute the budget weights for each item in the basket
  5. Multiply the price of each item by the quantity consumed and aggregate for all items in the basket
  6. Compare the price of each item, or for the total basket, between the current year and base year (or month) to get the CPI for that period

Constructing the CPI



Constructing the CPI

- **Some important aspects to consider in the construction of the CPI**
  - What's in the basket?
  - Does the basket change over time? if so, how often should it be updated?
  - How do we get the weights for the different items in the basket?
  - How do we collect prices on the items in the basket?
  - How often should we collect and update the prices of different goods?

Constructing the CPI

- CPI for Portugal built by INE (Office for National Statistics).
- Measured monthly!
- Shopping basket: +700 items, +70,000 prices
- CPI bias: criticism that CPI overstates the effect of price inrceases / inflation, because of:
  - Substitution effect
  - Quality changes of goods
  - Delay in new products entering the shopping basket

Using the CPI to adjust for inflation

- Minimum wage in Portugal at **current prices (or nominal values)**:
  - 1975: 20 Euros
  - 2015: 505 Euros
  - Nominal rate of change is 2425% over 40 years

Using the CPI to adjust for inflation

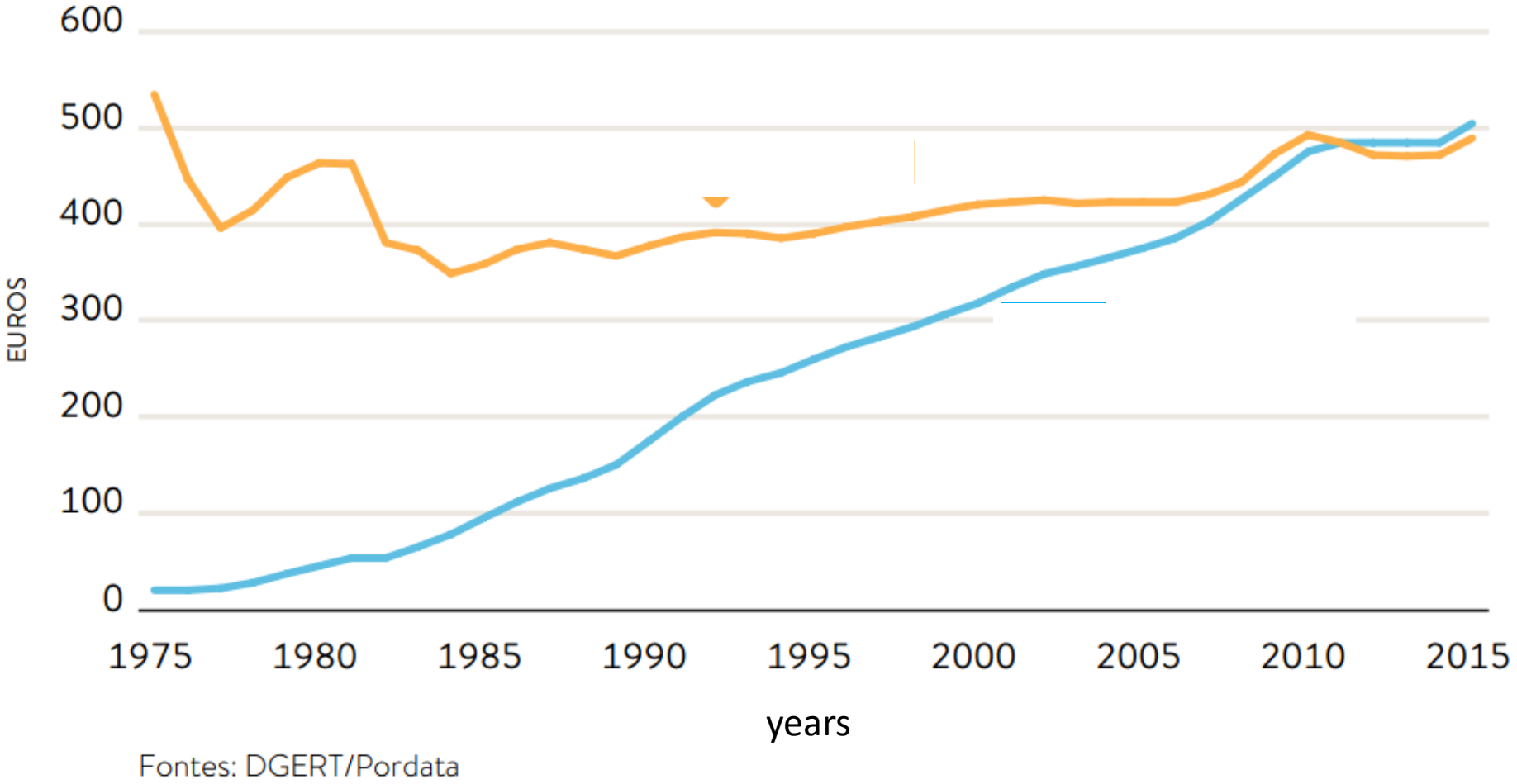
- Minimum wage in Portugal at **current prices (or nominal values)**:
  - 1975: 20 Euros
  - 2015: 505 Euros
  - Nominal rate of change is 2425% over 40 years

- **Problem**! 1 euro in 1975 could buy a lot more than 1 euro in 2015 - We need to adjust for inflation (i.e. increase in prices)

- Therefore, we need to compare the minimum wage in **constant prices or real values**. We use the CPI to adjust for inflation.

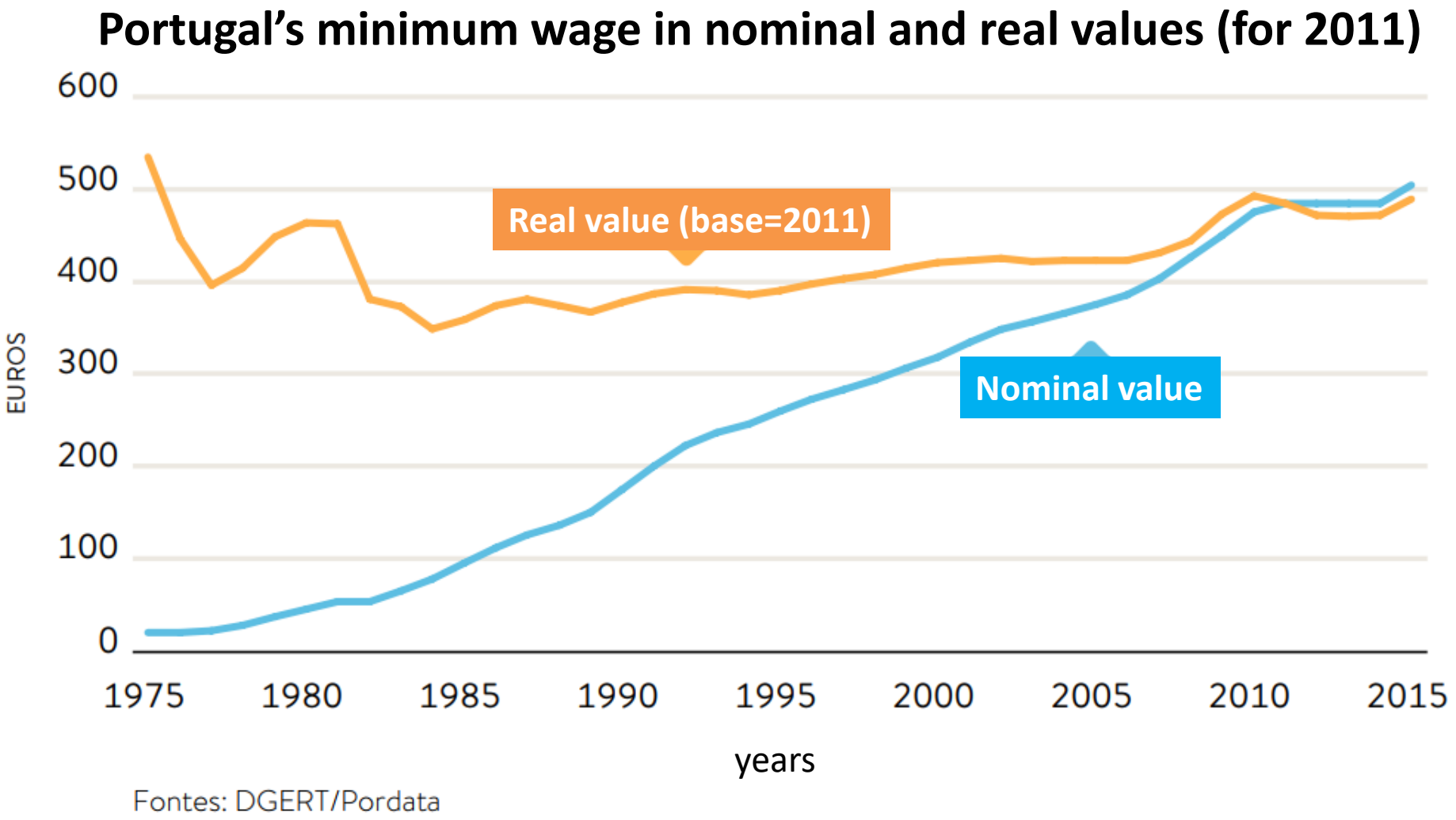
Using the CPI to adjust for inflation

Portugal's minimum wage in nominal and real values (for 2011)



Source: https://www.ffms.pt/publicacoes/detalhe/1963/que-numero-e-este

Using the CPI to adjust for inflation



Source: https://www.ffms.pt/publicacoes/detalhe/1963/que-numero-e-este

Using the CPI to adjust for inflation

- Minimum wage in Portugal - nominal vs real values:
  - 1975: 20 Euros at current 1975 prices worth 535 Euros at 2011 prices
  - 2015: 505 Euros at current 2015 prices worth 489 Euros at 2011 prices
  - When considering the minimum wage in real values, we see that it was higher in 1975 than in 2015!

- In fact, when using real values (i.e. after adjusting for inflation) we can see that the minimum wage was the highest in 1975!

	1975	**2009**	**2011**	**2015**
Minimum wage at current prices	20 Eur	450 Eur	485 Eur	505 Eur
CPI (base=2011)	0.037	0.951	1.00	1.033
Minimum wage at 2011 constant prices	(20/0.037) ≈535 Eur	(450/0.951) ≈473 Eur	(485/1.00) ≈485 Eur	(505/1.033) ≈489 Eur

Source: https://www.ffms.pt/publicacoes/detalhe/1963/que-numero-e-este

Using the CPI to adjust for inflation

- Change the CPI base from 2011 to 2009 and calculate the new minimum wage at constant prices for 2009. **Which operation or property of index numbers should you use?**

- Minimum wage in Portugal - nominal vs real values for 2009:
  - 1975: 20 Euros at current 1975 prices worth **?? Euros at 2009 prices**
  - 2015: 505 Euros at current 2015 prices worth **?? Euros at 2009 prices**

	1975	**2009**	**2011**	**2015**
Minimum wage at current prices	20 Eur	450 Eur	485 Eur	505 Eur
CPI (base=2011)	0.037	0.951	1.00	1.033
CPI (base=2009)				
Minimum wage at 2009 constant prices				

Using the CPI to adjust for inflation

- Change the CPI base from 2011 to 2009 and calculate the new minimum wage at constant prices for 2009. **Which operation or property of index numbers should you use? rebasing**

- Minimum wage in Portugal - nominal vs real values for 2009:
  - 1975: 20 Euros at current 1975 prices worth **513 Euros at 2009 prices**
  - 2015: 505 Euros at current 2015 prices worth **465 Euros at 2009 prices**

	1975	**2009**	**2011**	**2015**
Minimum wage at current prices	20 Eur	450 Eur	485 Eur	505 Eur
CPI (base=2011)	0.037	0.951	1.00	1.033
CPI (base=2009)	0.037/0.951= 0.039	1.00	1.00/0.951= 1.052	1.033/0.951 =1.086
Minimum wage at 2009 constant prices	(20/0.039) ≈513 Eur	(450/1) ≈450 Eur	(485/1.052) ≈461 Eur	(505/1.086) ≈465 Eur

Using the CPI to adjust for inflation

- Change the CPI base from 2011 to 2015 and calculate the new minimum wage at constant prices for 2015. **Which operation or property of index numbers should you use?**

- Minimum wage in Portugal - nominal vs real values for 2015:
  - 1975: 20 Euros at current 1975 prices worth **?? Euros at 2015 prices**
  - 2015: 505 Euros at current 2015 prices worth **?? Euros at 2015 prices**

	1975	**2009**	**2011**	**2015**
Minimum wage at current prices	20 Eur	450 Eur	485 Eur	505 Eur
CPI (base=2011)	0.037	0.951	1.00	1.033
CPI (base=2015)				
Minimum wage at 2015 constant prices				

Using the CPI to adjust for inflation

- Change the CPI base from 2011 to 2015 and calculate the new minimum wage at constant prices for 2015. **Which operation or property of index numbers should you use? rebasing**

- Minimum wage in Portugal - nominal vs real values for 2015:
  - 1975: 20 Euros at current 1975 prices worth **556 Euros at 2015 prices**
  - 2015: 505 Euros at current 2015 prices worth **505 Euros at 2015 prices**

	1975	**2009**	**2011**	**2015**
Minimum wage at current prices	20 Eur	450 Eur	485 Eur	505 Eur
CPI (base=2011)	0.037	0.951	1.00	1.033
CPI (base=2015)	0.037/1.033= 0.036	0.951/1.033 =0.921	1.00/1.033= 0.968	1.00
Minimum wage at 2015 constant prices	(20/0.036) ≈556 Eur	(450/0.921) ≈489 Eur	(485/0.968) ≈501 Eur	(505/1) ≈505 Eur