STATISTICS

Graphic representation methods

Ackonowledgements: Balogh Péter,
Jeney László, Michael Friendly

- **Aim of the statistical analysis:** making a summary;
- **Objectives of the graphical analysis:** interpreting and presenting;
  - It is often more useful if it includes a visual summary (e.g. fitted curve, data ellipse);
- Different charts for different purposes:
  - sighting (overview);
  - exploration (review templates, trends);
  - model diagnosis (assumptions, outliers);

- Multivariate data require new graphs for presenting relationships between the variables;
  - scatterplot matrices;
  - visual attenuation: less is often more;
  - Low-D view (double drawings);
  - HE drawings to visualize multivariate tests;
  - mosaic drawings to visualize the n-fold frequency tables;

Graphic representation methods

Comparison is a basic method of the statistical analysis. This can be done through the help of various calculations and **graphic representations**.

Its advantages:

- visual;
- it presents also ratios;

Graphic representation methods

- Types of graphic representation methods;
  - General statistical graphical methods;
  - Mapping procedures;
- Functions: Tool and purpose;
  - In research work: analysis tool;
  - In paper, presentation: illustration purpose;
- It is good, if it is expressive, namely it can be used without text additions (particularly in PowerPoint);
  - But: it cannot substitute analysis (each Figure should involve text);

It should involve all relevant information (without repetitions)

- Preferably the title should involve:
  - Study area: e.g. Hungary (regional level: E.g. regions);
  - The phenomenon studied: e.g. regional economic disparities;
  - Indicator: e.g. per capita GDP;
  - Study period (or periods): e.g. 2004 (or 2004-2012);
- Preferably the category axis should involve:
  - Unit, e.g. per capita USD;
- Information must not appear simultaneuously in two places (either in title only, or in category axis only);
- Title is rarely in Excel figure itself (no fault);
  - Word: separate line under the figure (can also easily be changed), PowerPoint: it may already be posted in Excel);
  - Source always should be indicated (in PowerPoint as well);

Figures should always be chosen for the type of phenomenon

TYPES OF REPRESENTATION (1);

- **Using geometric shapes**: mostly in coordinate system, using diagrams;
- **Without coordinate system:** the most common type is pie chart;

Figures should always be chosen for the type of phenomenon

TYPES OF REPRESENTATION (2);

- **Map representation:** the most common way of projection representation; it can be occurred with place discrimination by colour or by applying local charts;
- **Figural representation:** it serves publicity purposes, for statistics it is frivolous, it informs only approximately;
- Easier graphical charting methods;
  - scatter plots;
  - line chart (graph);
  - bar chart;
  - strip chart;
  - pie chart: recommended rarely (only with a few sectors);
  - bubble chart;
  - radar (radiation) diagram;
  - triangle chart;
  - histogram;

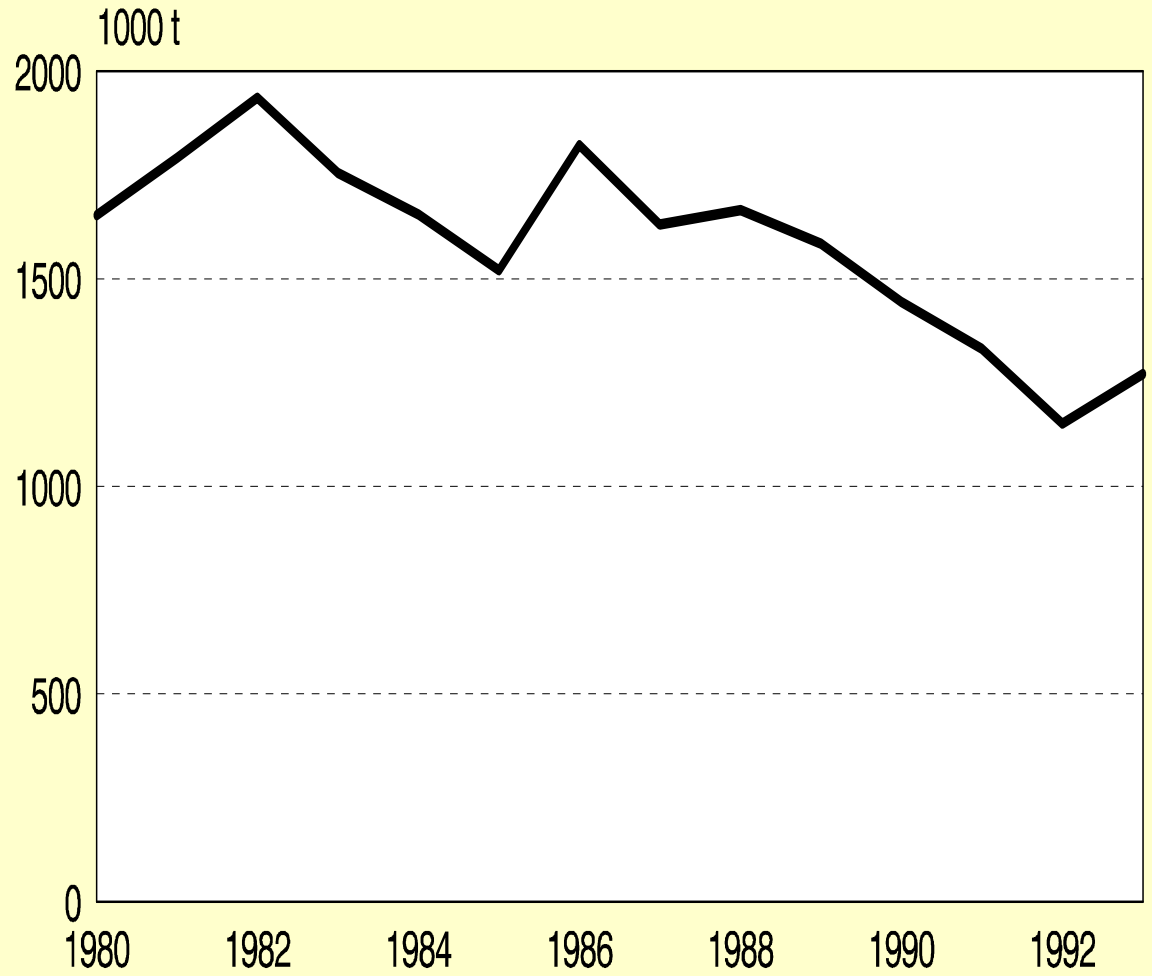
Principles of representation

- The figure should be simple and clear.
- Similar to the statistical table it should have a title; the source of the data should be marked in footnote!
- Let it be a legend! It should comprise all notations applied to the table! Indicate the units used!
- The figure, with all accessories (title, footnotes, legend, names written into a table) should be clear, every detail is understandable, regardless of the context.
- The figure should ensure comparability! If coordinate system is used, be sure to indicate the axis roster, including any breaking off the axis! Grating (grid) can help the reading accuracy.

Coordinate systems

- The figures are mostly placed in a coordinate system. This allows us to read the dimensions to be portrayed.
- The axes should indicate scale graduation. If the graduation is not continuous from the intersection of the axis, i.e. from 0, then the break should be indicated by two tildes.
- Mostly Cartesian coordinate system is used.

Fruit production in Hungary

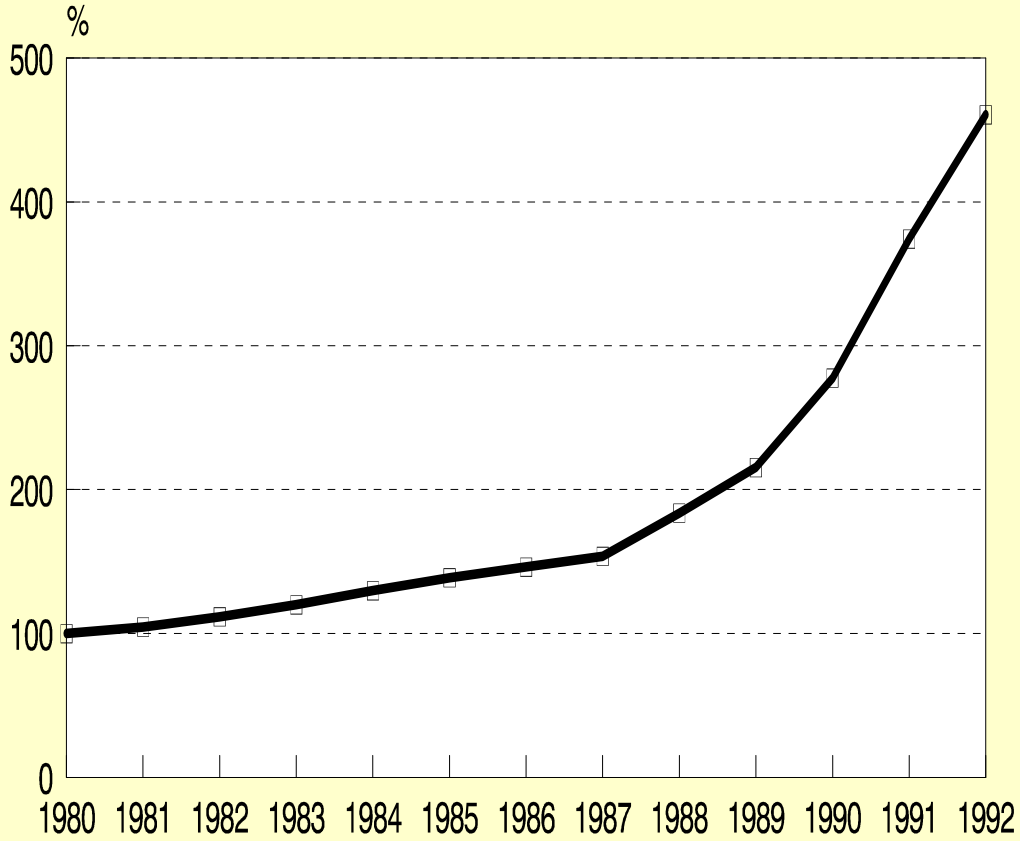


Hungarian Statistical Yearbook, 1993

Consumer price index

Worker – employee, until 1990

1980 = 100



Source: Hungarian Statistical Yearbook

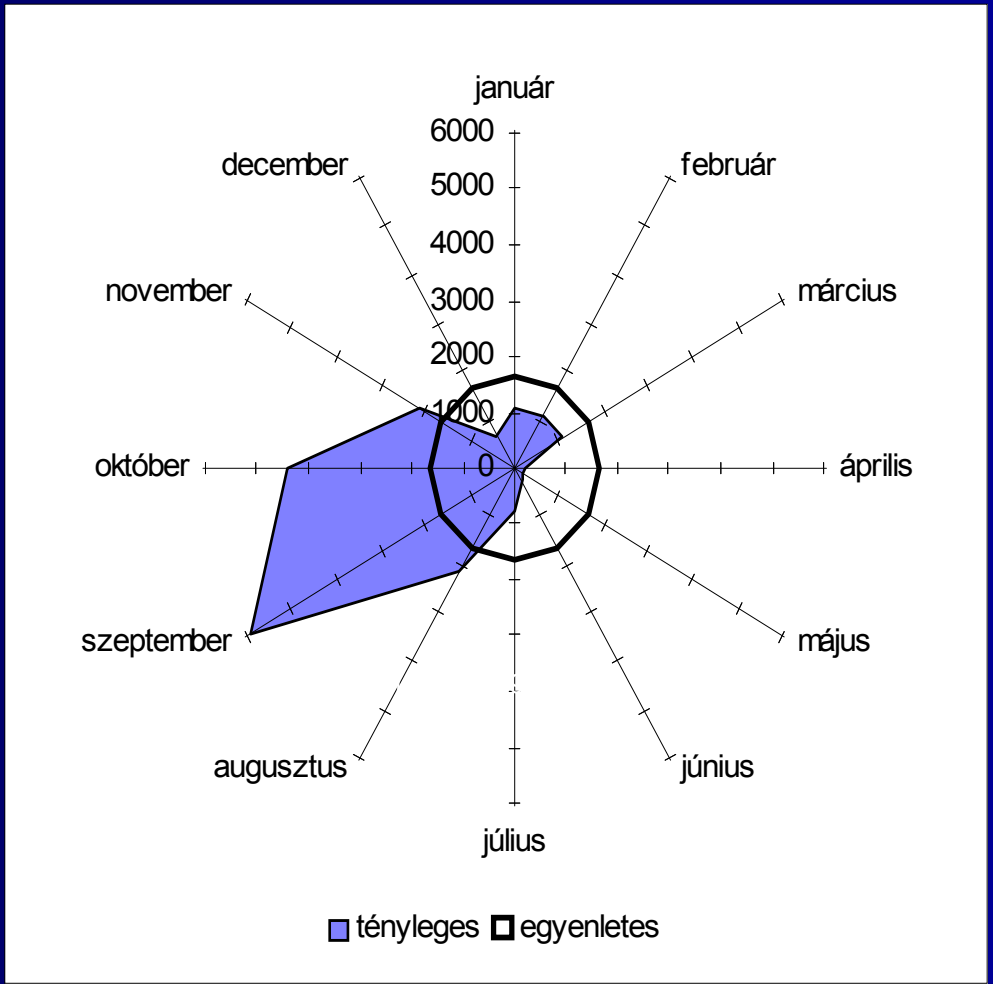
Polar coordinate system

- Polar coordinate system is not Cartesian, the axes meet with different angles to each other (except for 90°), and only positive shafts used to be presented.

- It is used to represent the seasonality within the year, where the 12 months are represented by 12 axes each. If data are located on a circle centered in origo (axis origin), it mean the lack of seasonality. The greater the deviation of the location of data from this circle is, the stronger seasonality.

Polar coordinate system

Acquisition of onions (t), 1993

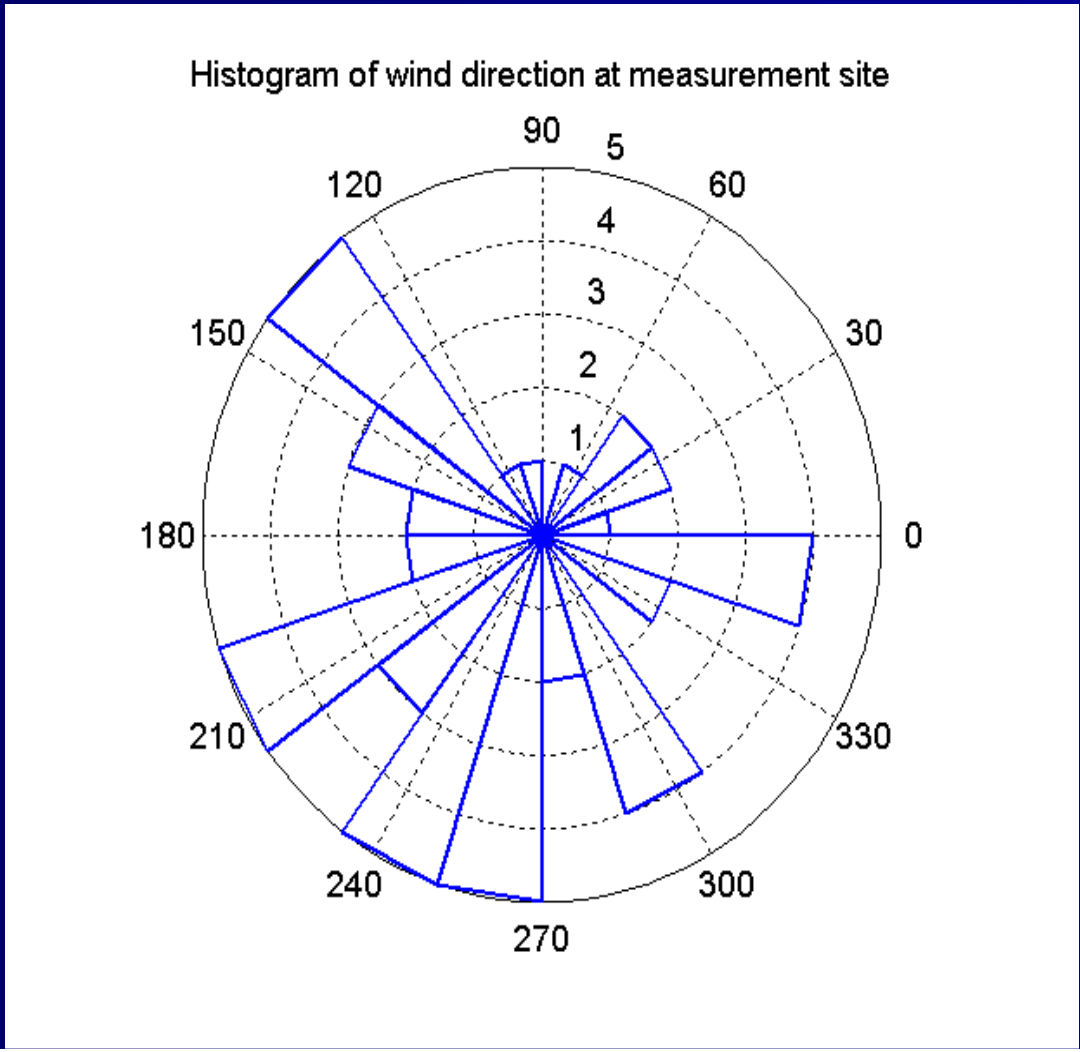


month of the year →

Blue, amporphous: actual

Black circle: uniform

Polar coordinate system - rose diagram



Polar coordinate system

- It can also be used e.g. when communicating of the results of sensory investigation of foods.
- If a product is characterized by its several features, each of them indicating on one axis separately, a polar curve specific to the quality of the product (also known as web chart) is obtained.
- Their comparison facilitates combined consideration of the properties.

Types of graphic figures

- line chart;
- histogram;
- column and strip chart;
- pie chart and other charts;
- cartogram (map);
- pictogram ("image" of a phenomenon);

Histogram

A **histogram** is a bar graph, where the area of the columns is proportional to the **frequency**.

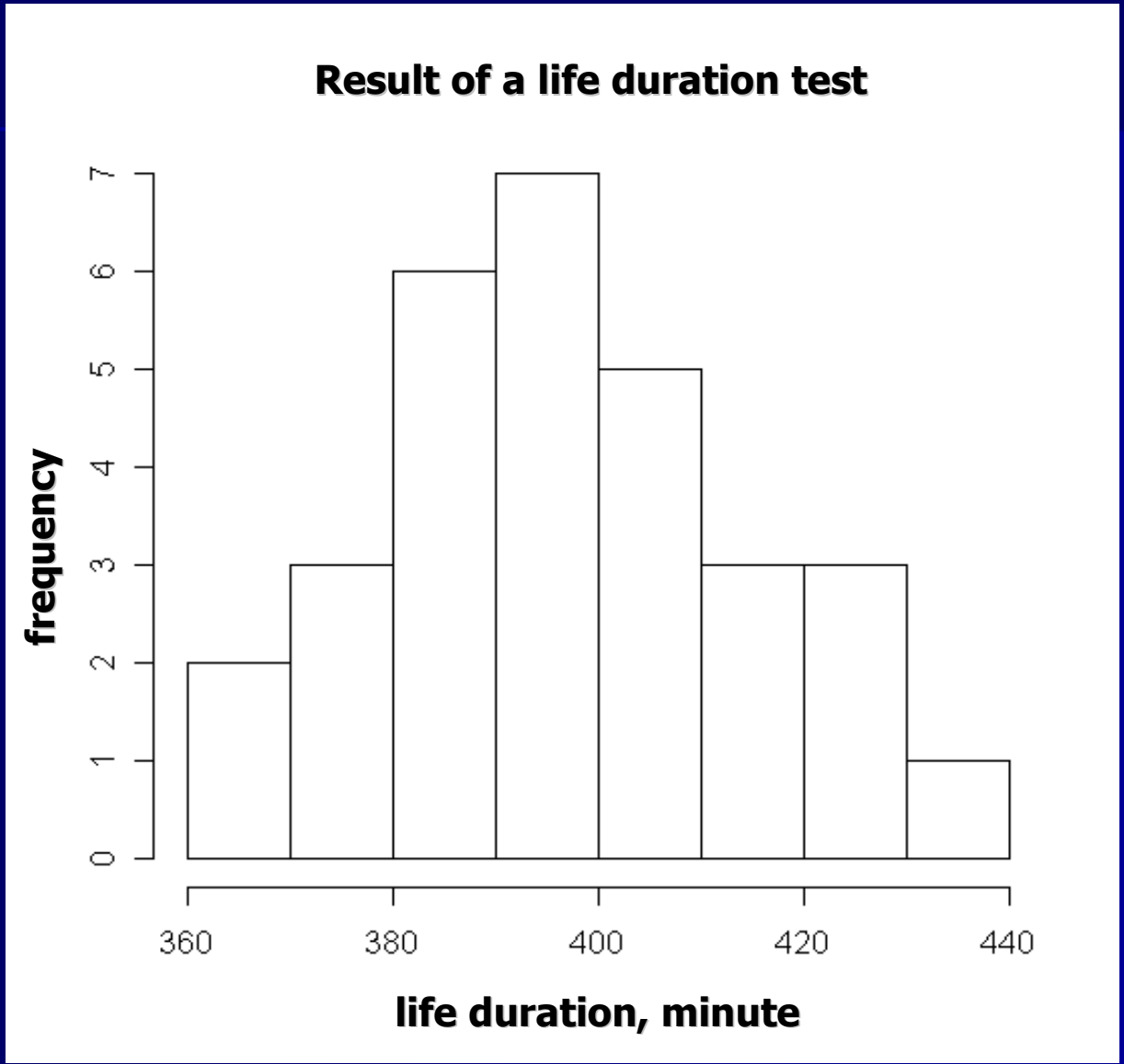
Horizontal axis: class intervals;

Vertical axis: **density:** (relative) frequency *per unit class interval*.

ATTENTION!

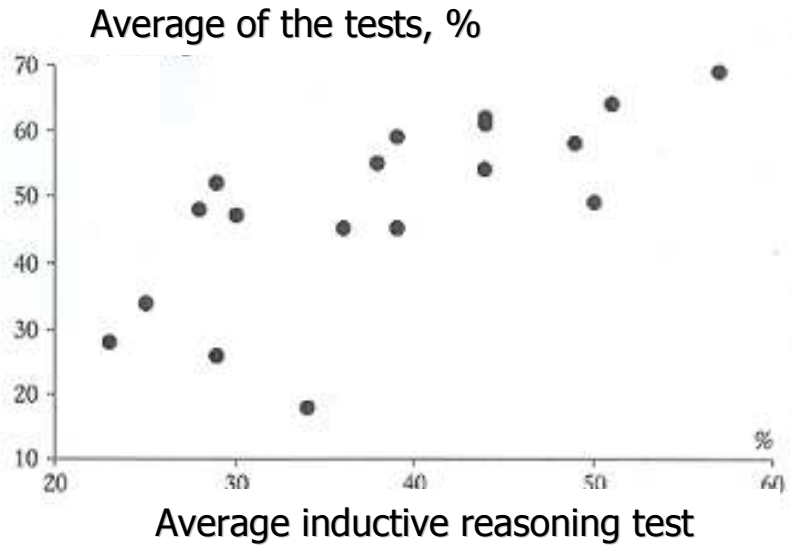
- In case of **non-identical** class intervals only **density** can be interpreted on the vertical axis!
- In case of **identical** class intervals **(relative) frequency** can also be **applicable** since it is **proportional to density**.

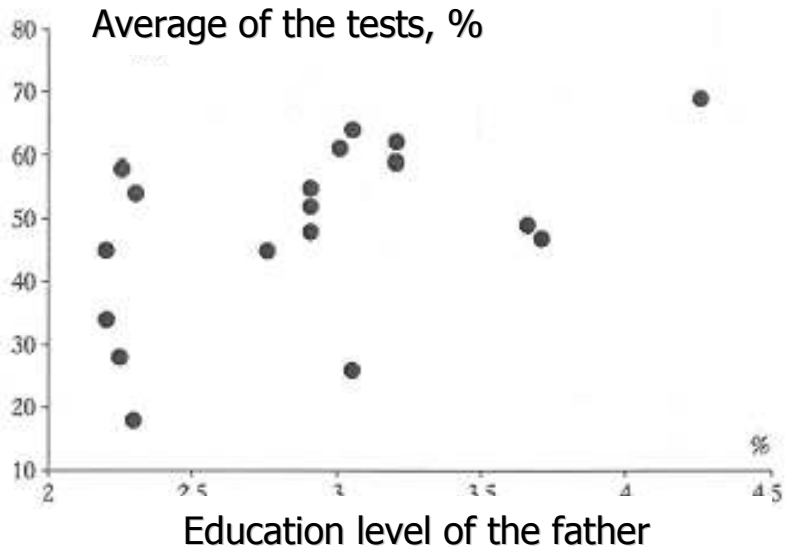
Histogram


Result of a life duration test
frequency
0
1
2
3
4
5
6
7
360
380
400
420
440
life duration, minute

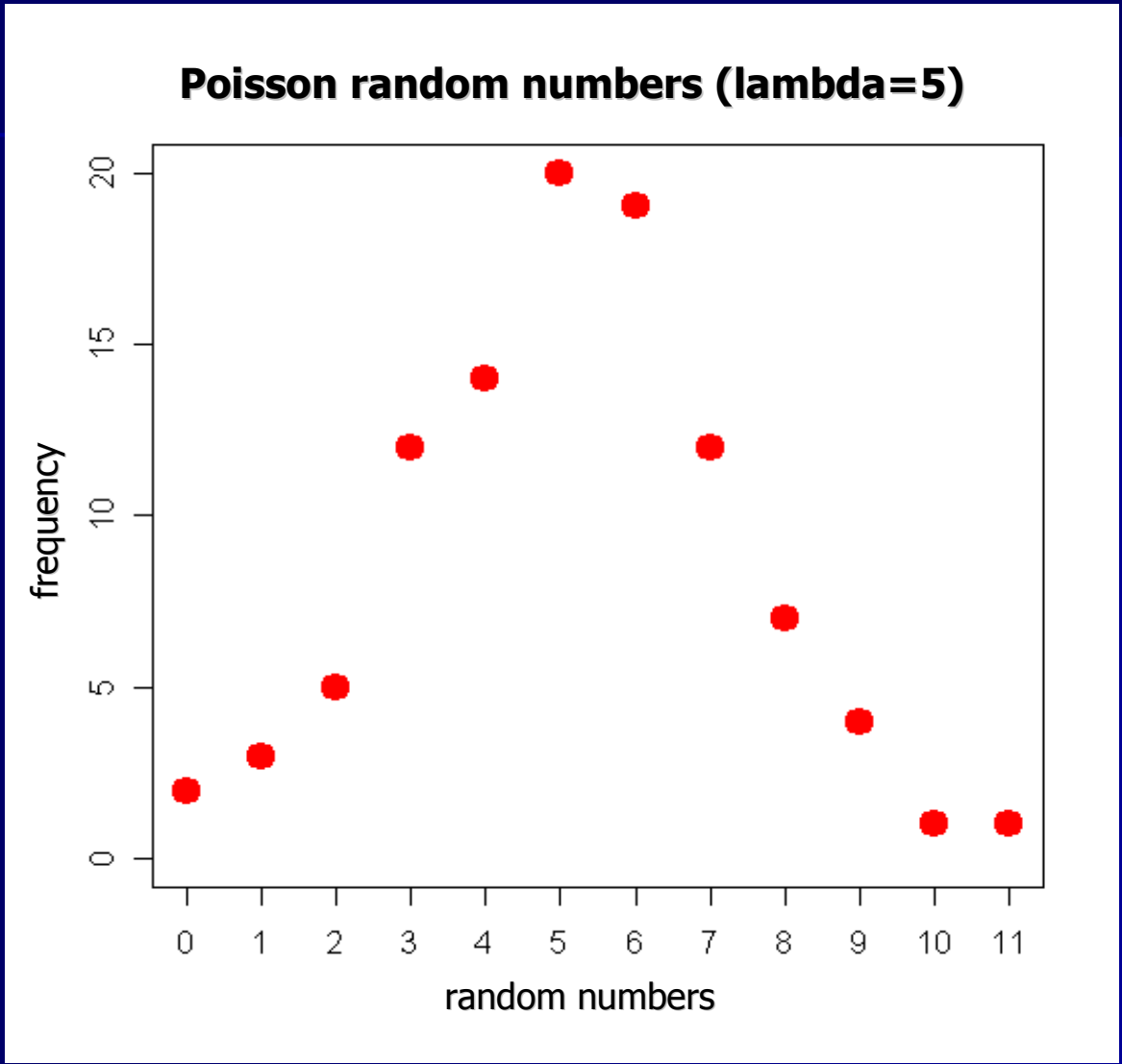
Scatter plots

- Scatter plots: representation of two interrelated quantitative criteria in a coordinate system (temporal and quantitative time series)





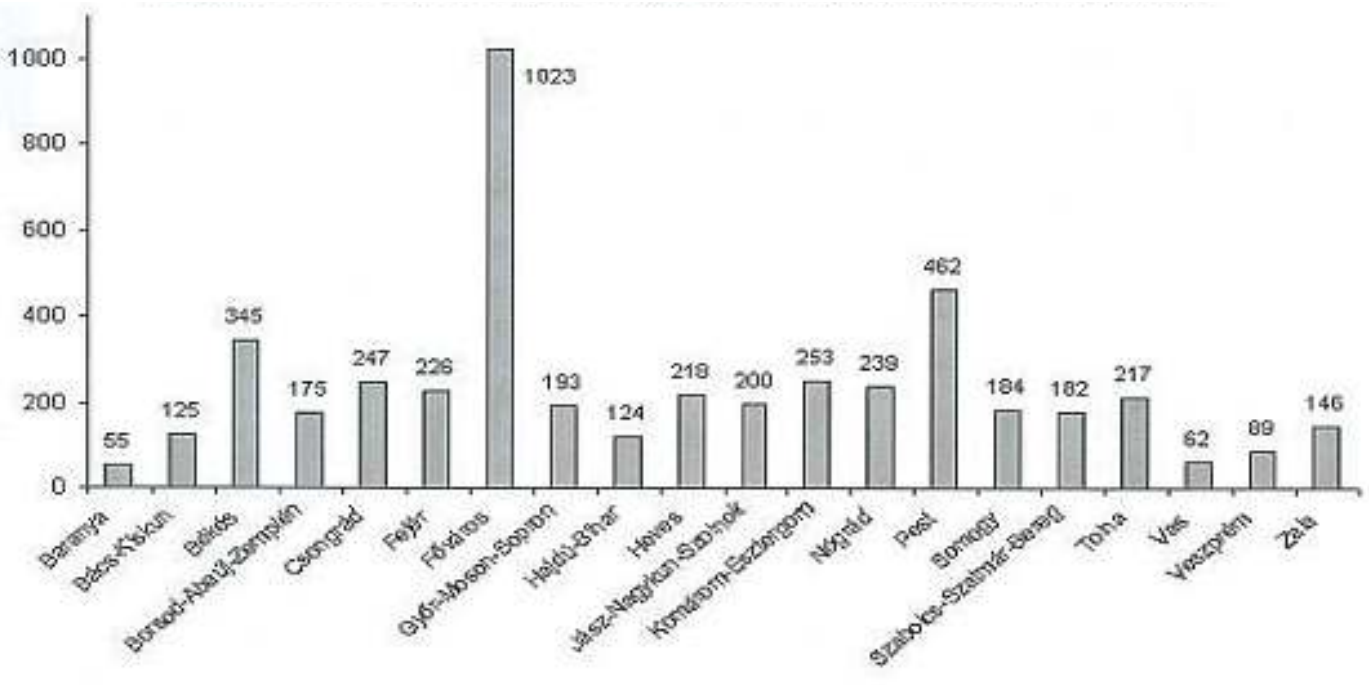
Scatter plots



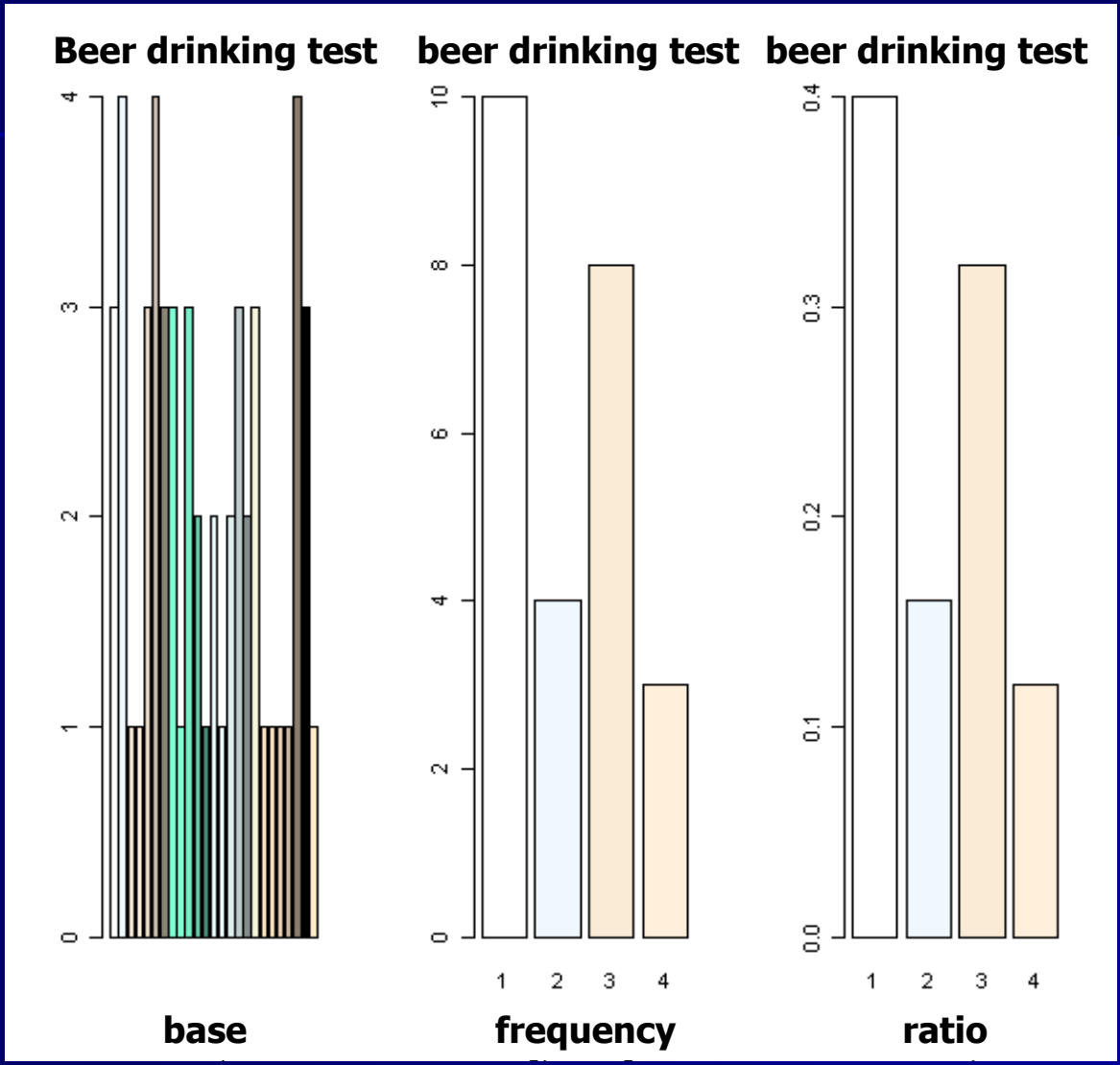
Bar chart

- **Bar chart:** comparison with the height of the columns

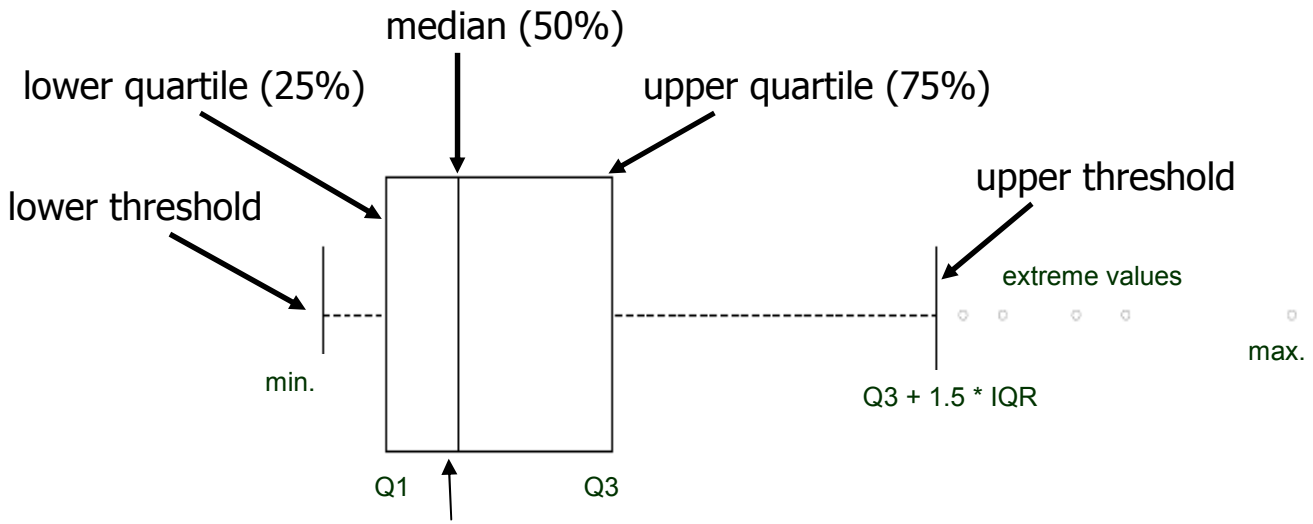
Number of communications by county (capital city) victim assistance services



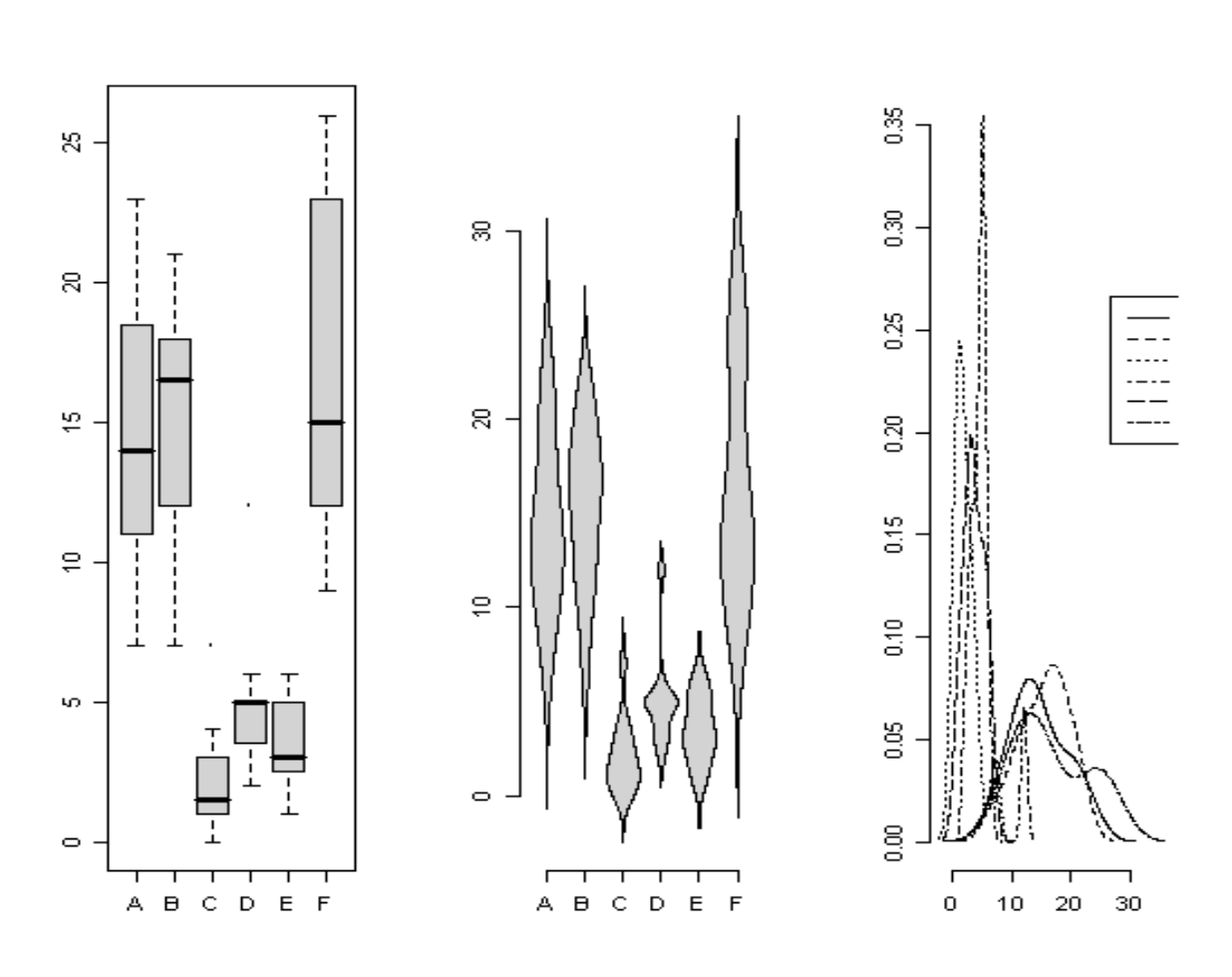
Bar chart


Beer drinking test
beer drinking test
beer drinking test
0
1
2
3
4
0
2
4
6
8
10
0.0
0.1
0.2
0.3
0.4
1
2
3
4
1
2
3
4
base
frequency
ratio

Box plot representation



Representation of the violin diagram in the company of the box plot and density chart



Editing rules of graphic illustration

- You have to choose the most appropriate mode of representation;
- The selected graphic picture should be simple, clear, transparent and expressive;
- The selected picture should promote correct understanding of the rates and relationships of the phenomena it has expressed;
- Each graph should consist of two parts :
  - figure;
  - explanatory symbols (title, scale, legend);
  - any footnotes;

Line chart

- It serves for representing seamless, equidistant time series and quantity time series of identical cross-volume.
- It is important choosing the right scale.
- Several interrelated phenomena are often analyzed using graphs.
- It is more appropriate to start from a common base instead of absolute figures.

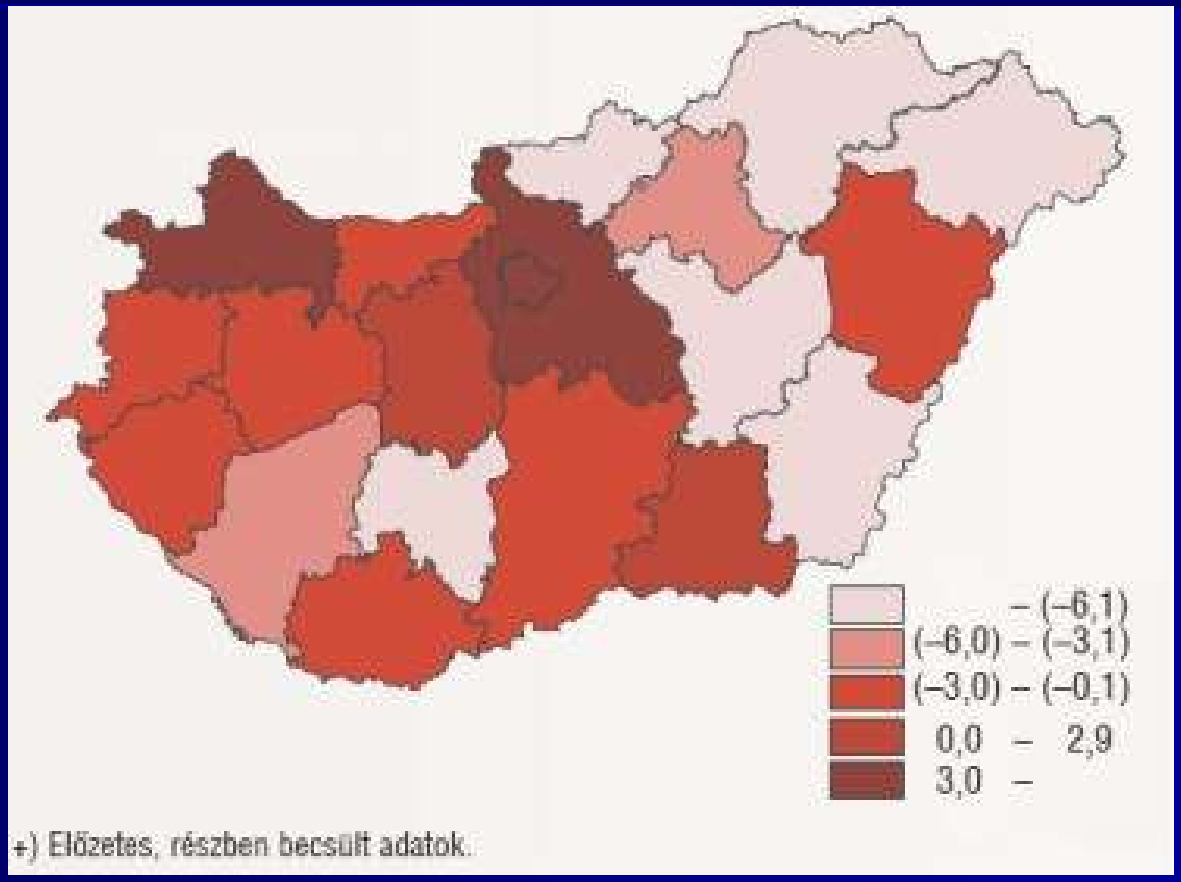
Column and strip chart

- Except for descriptive statistical rows, any statistical row can be represented in this way.
- It is advisable to carry out the analysis with columns of equidistant basis, as the ratio among the data is illustrated by the height of the columns in this way.
- When representing time series, this form is used instead of line chart if:
  - only data for some periods are available;
  - representation is carried out on the basis of seamless time series;
  - comparing data on unequal intervals;

Chartogram and chart diagram

- This kind of spatial visualization is used for analyzing regional time series if regional data are complete.
- The fact of the representation is that some features of territorial units are represented on the map with different colouring and line making.
- In case of **chart diagram** quantitative data are defined to a surface, a given area of the map. The reference area can be of any size: in general, a statistical, or administrative territorial unit. The chart should be positioned within the area that its location should be clear, easily recognizable.
- It should be placed
  - for small surfaces in proximity of the location, while
  - for major surface in the gravity centre of the location.

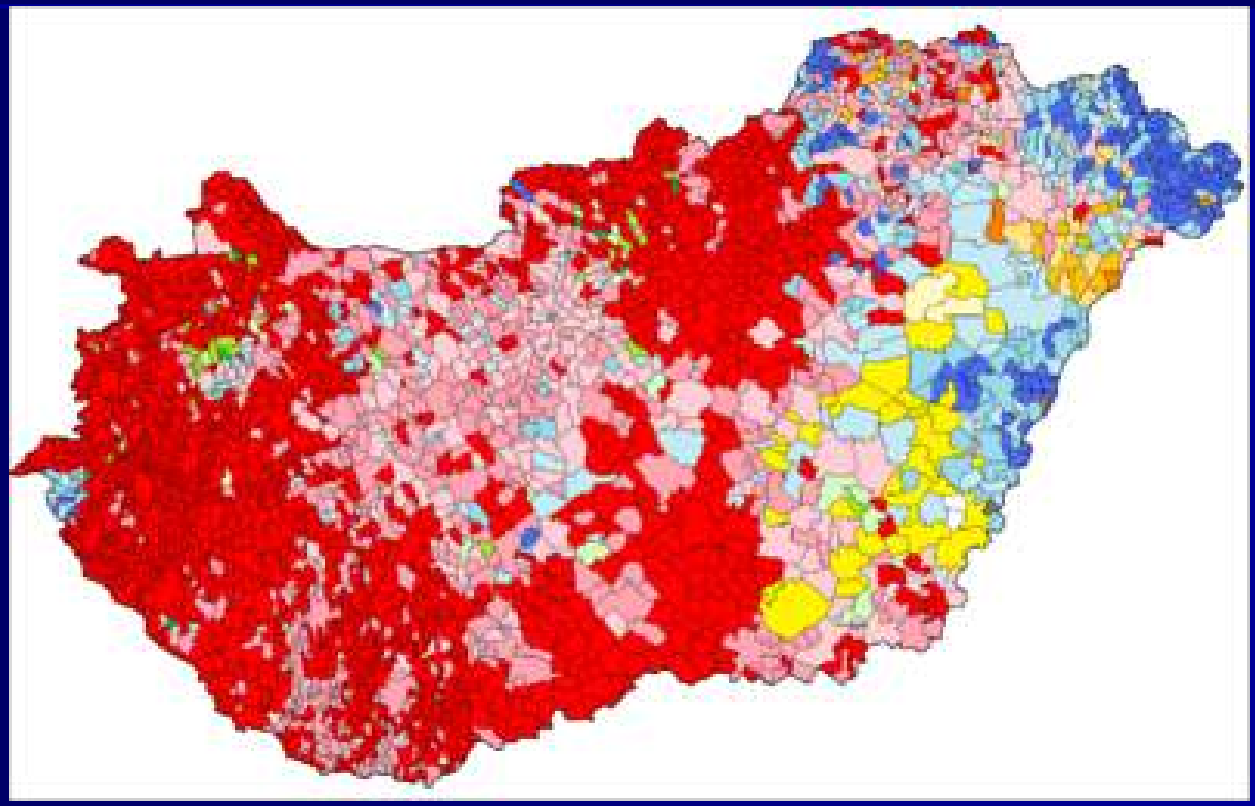
- **Chartogram** – colouring indicates the size of the given indicator;



Domestic migration difference per thousand inhabitants in 2009
(Source: Statistical reflection 2010/42)

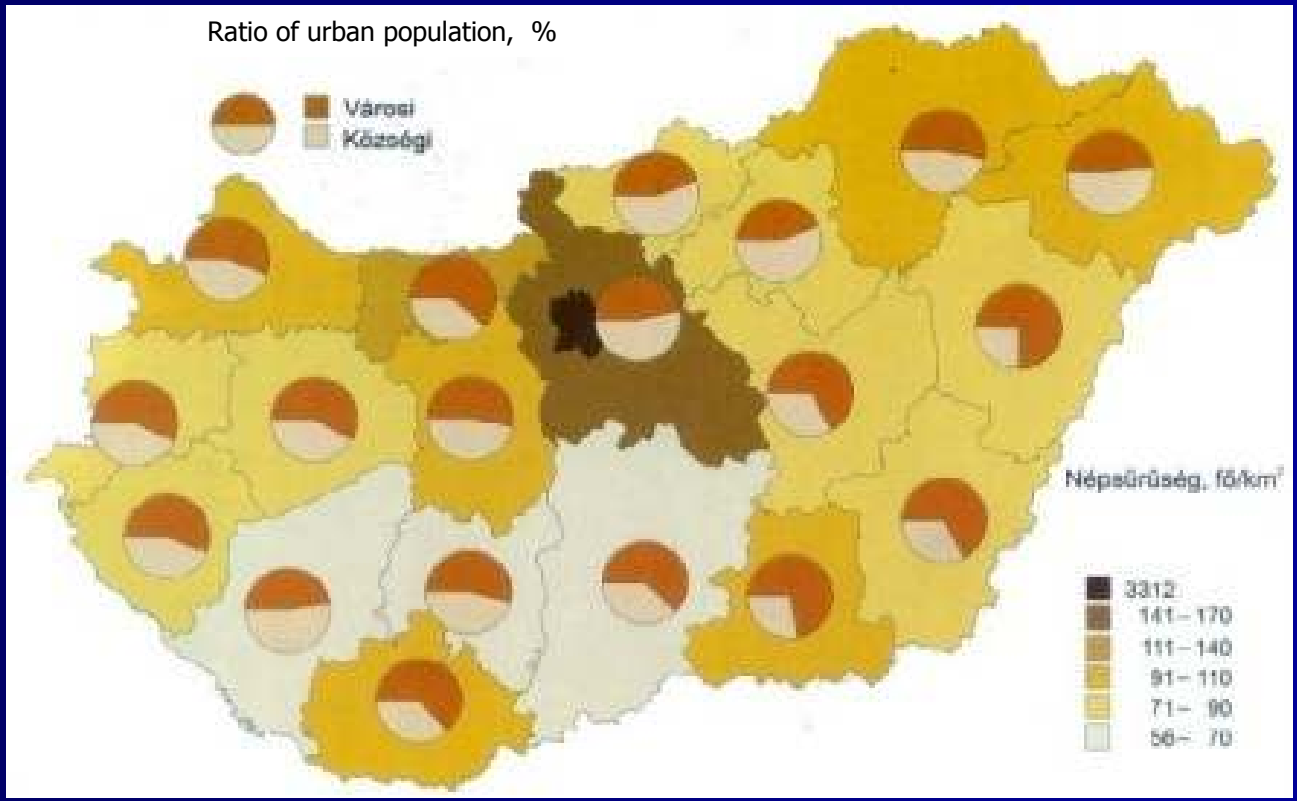
- **Chartogram** – colouring indicates the quality of the given indicator

).



The largest religious groups in Hungary. Red and pink: the Roman Catholic; blue: Reformed; green: Lutheran; brown or orange: Greek Catholic; yellow: the majority do not belong to any religious community (2001).

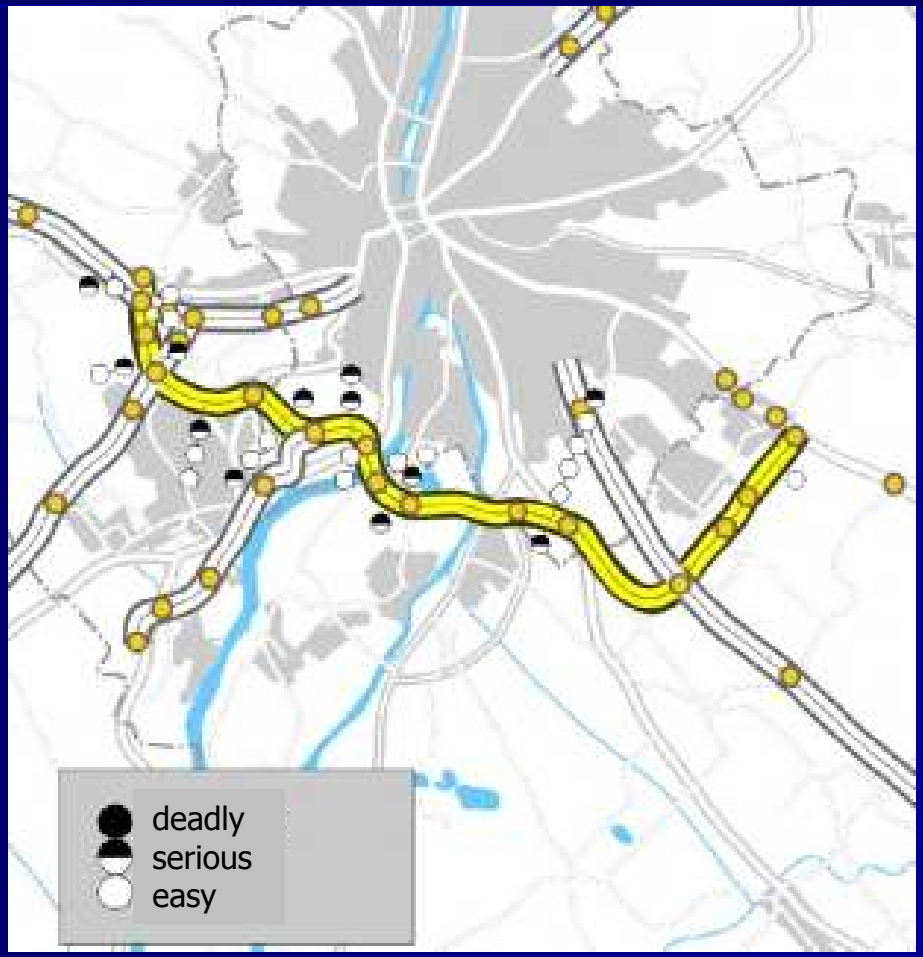
- **Chart diagram** – **c**harts are placed at the different parts of the map, thus demonstrating the magnitude of the index



The density of population and the proportion of the urban population, January 1, 2002
(Source: sdt.sulinet.hu)

Point map

- **Point map** – a kind of stocking rate, the amount of points indicates the size of the index



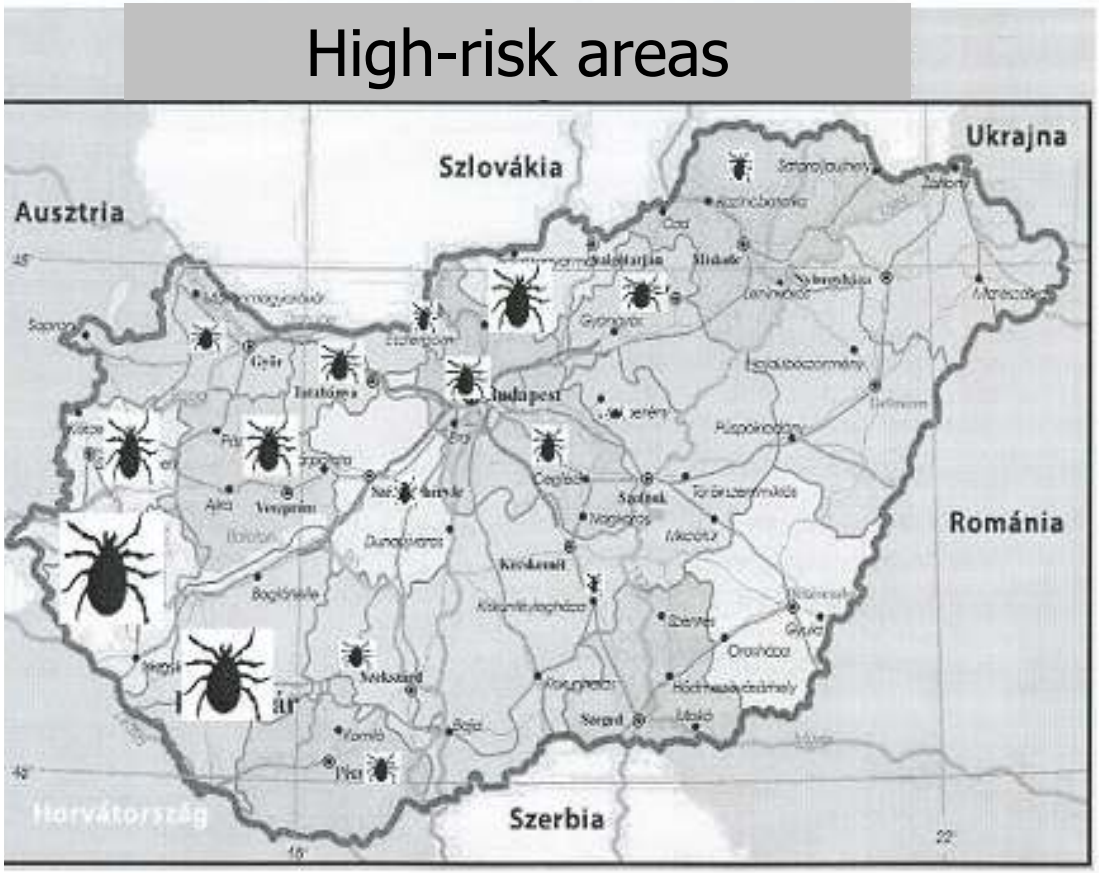
Highway no. M0, car accident point map, January – August, 2007
(Source: www.police.hu)

Pictogram

- This charting type (a signal displaying of an idea) expresses the ratio of the data with the ratio of the area of "pictures", representing the phenomena. For example, road signs; infographic figures, which are used instead of captions); signals used to image writing);

Pictogram

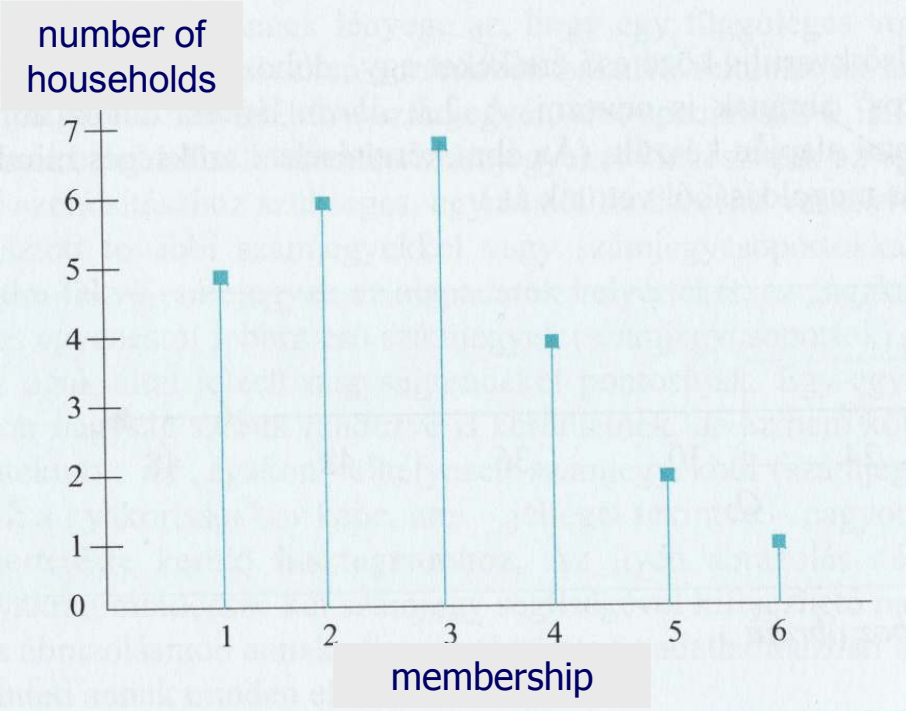
- Piktogram: figural representation, by which magnitude relation is expressed on the different size of figures representing the phenomenon



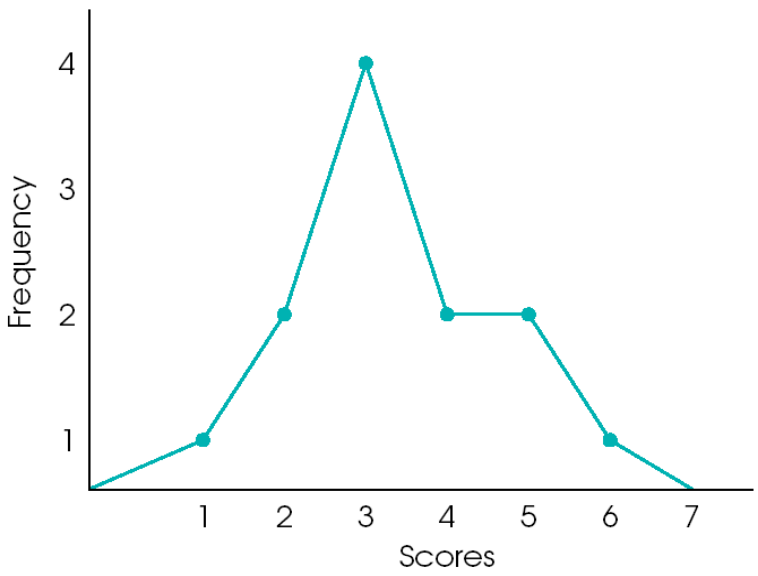
Representation of frequency distributions

Stick diagrams;

- For discrete quantitative variables recording only some values;



Frequency polygon (1)



X	f
6	1
5	2
4	2
3	4
2	2
1	1

A smoothed out histogram

Make a point representing *f* of each value

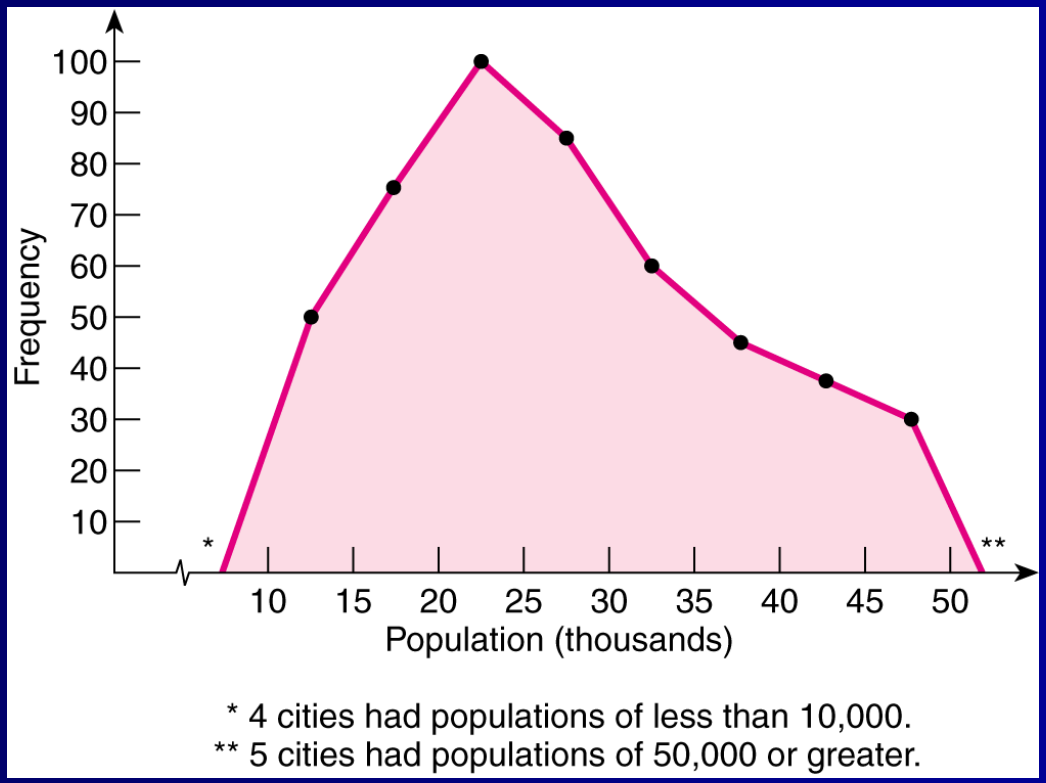
Connect dots

Anchor line on x axis

Useful for comparing distributions in two samples (in this case, plot *p* rather than *f*)

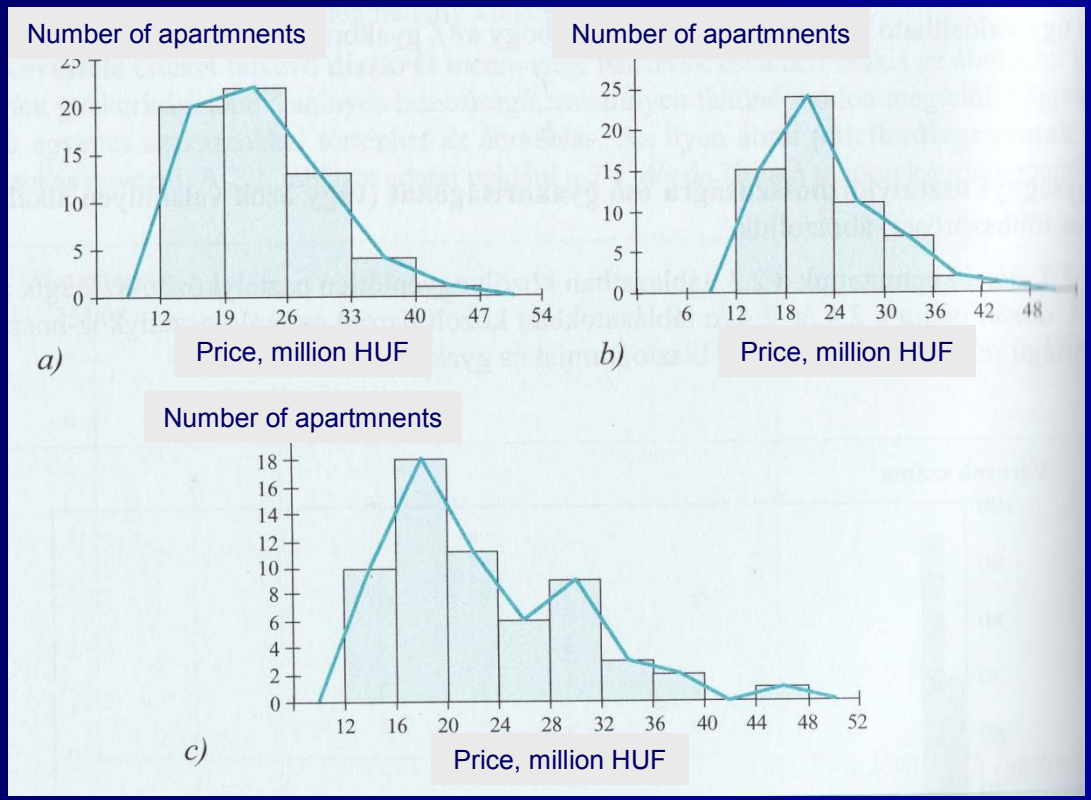
Frequency polygon (2)

- Frequency Polygons
  - Depicts information from a frequency table or a grouped frequency table as a **line graph**

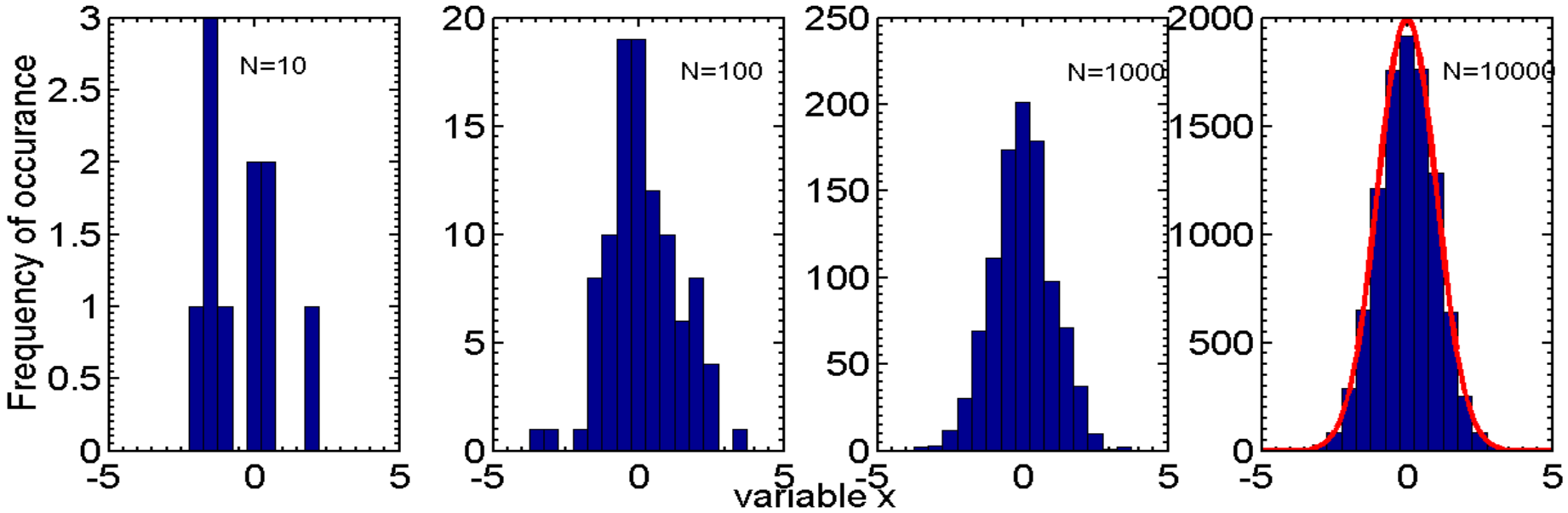


Histogram, polygon

- The area of the columns should be proportionate to the frequency shown.
- In case of different class interval lengths, **f_i/h_i unit length class intervals should be represented** instead of the frequencies f_i.

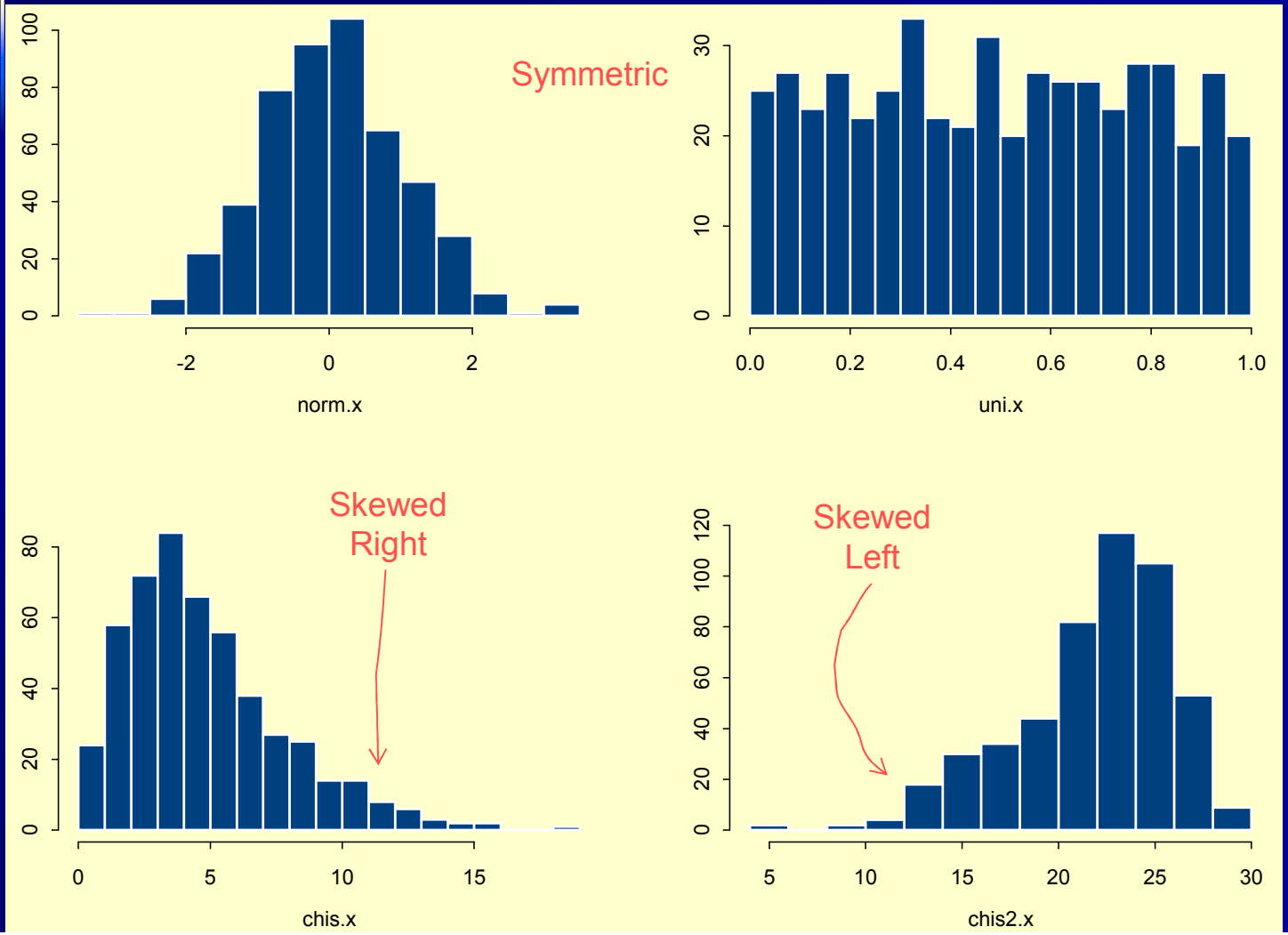


Frequency distribution of random errors – bar chart, histogram



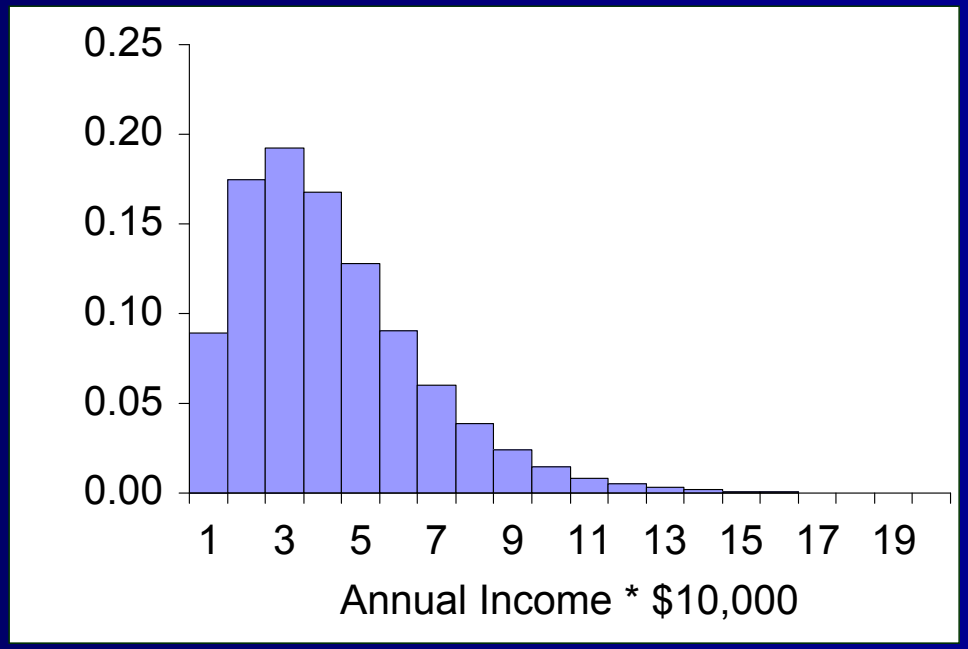
- As number of measurements increases the distribution becomes more stable;
  - **The larger the effect the fewer the data you need to identify it;**
- Many measurements of continuous variables show a bell-shaped curve of values, known as a Gaussian distribution.

Symmetric vs. skew



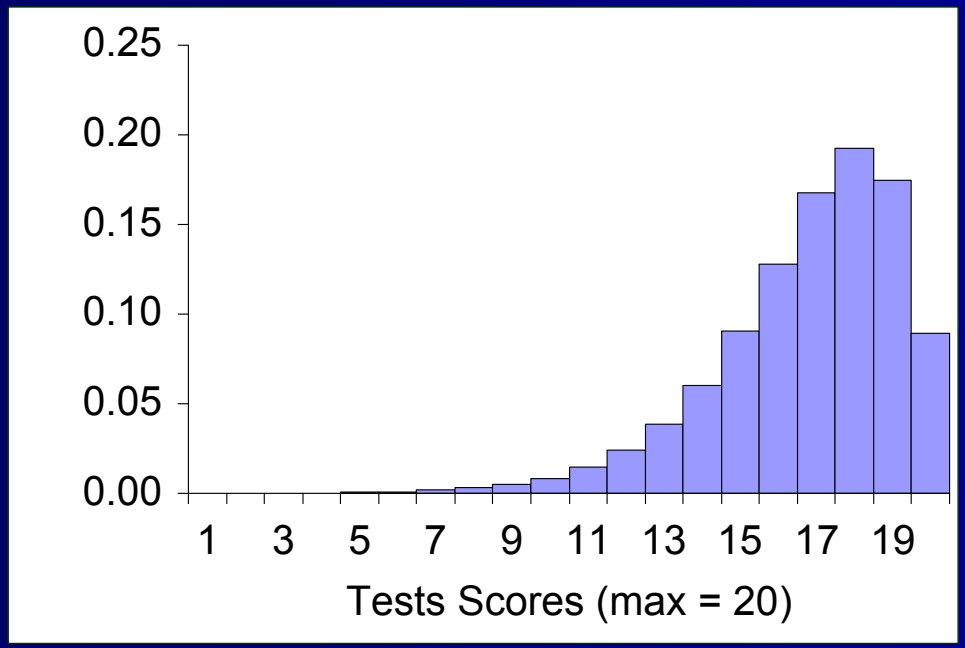
Skew frequency distribution

- Positively skewed;
  - Skewed right;
  - Tail trails to the right;
  - *** ***The skew describes the skinny end*** ***



Skew frequency distribution

- Negatively skewed;
  - Skewed left;
  - Tail trails to the left;

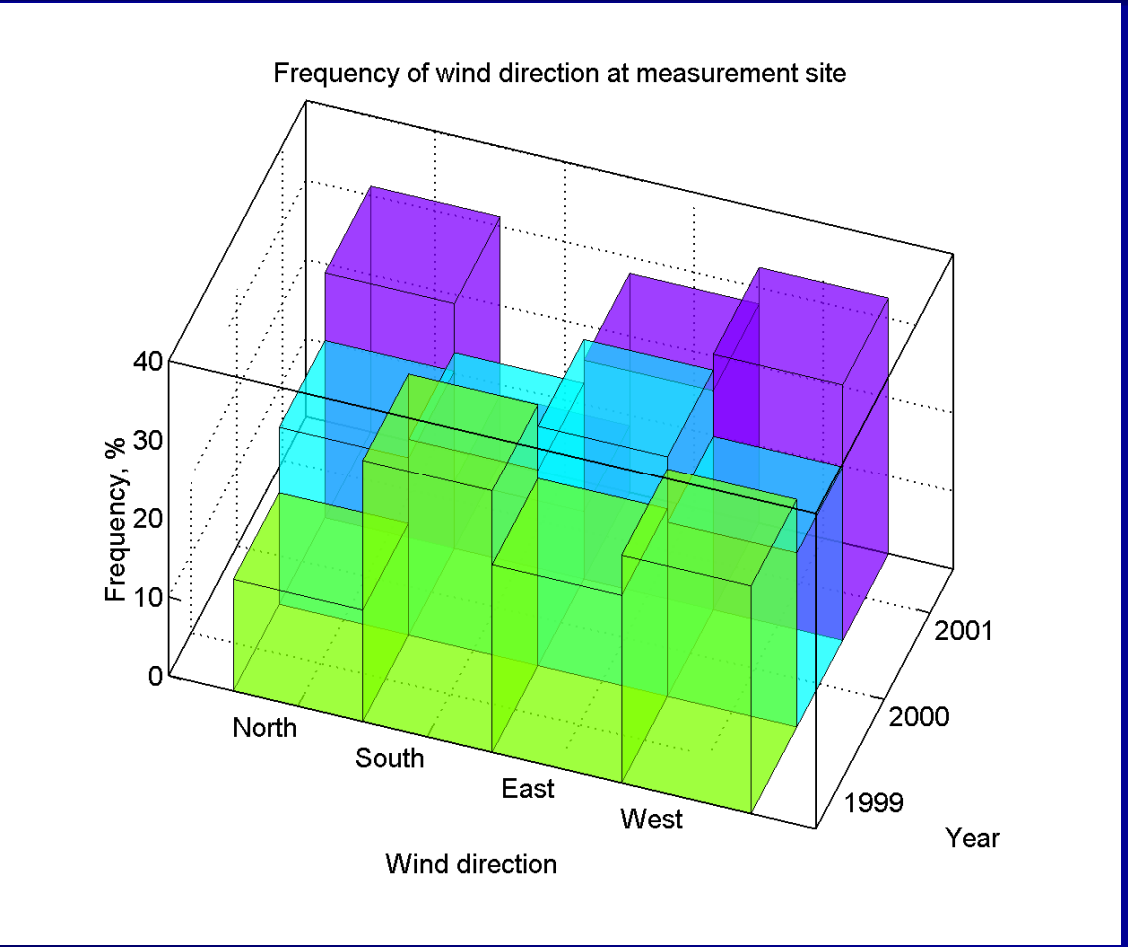


Univariate histogram

X	f
6	1
5	2
4	2
3	4
2	2
1	1

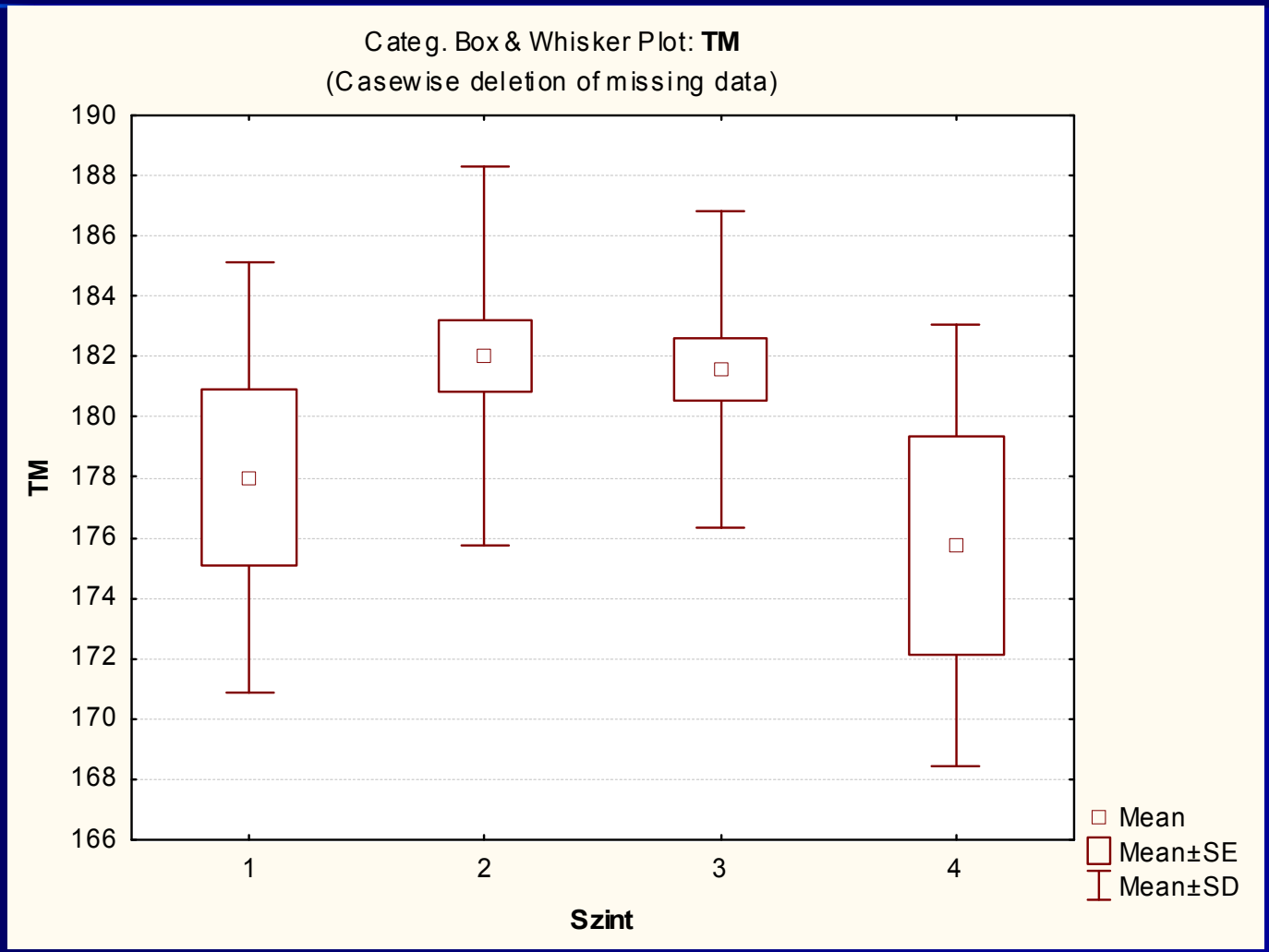
- *f* on *y* axis (could also plot *p* or *%*)
- X values (or midpoints of class intervals) on *x* axis
- Plot each *f* with a bar, equal size, touching
- ***No gaps between bars***

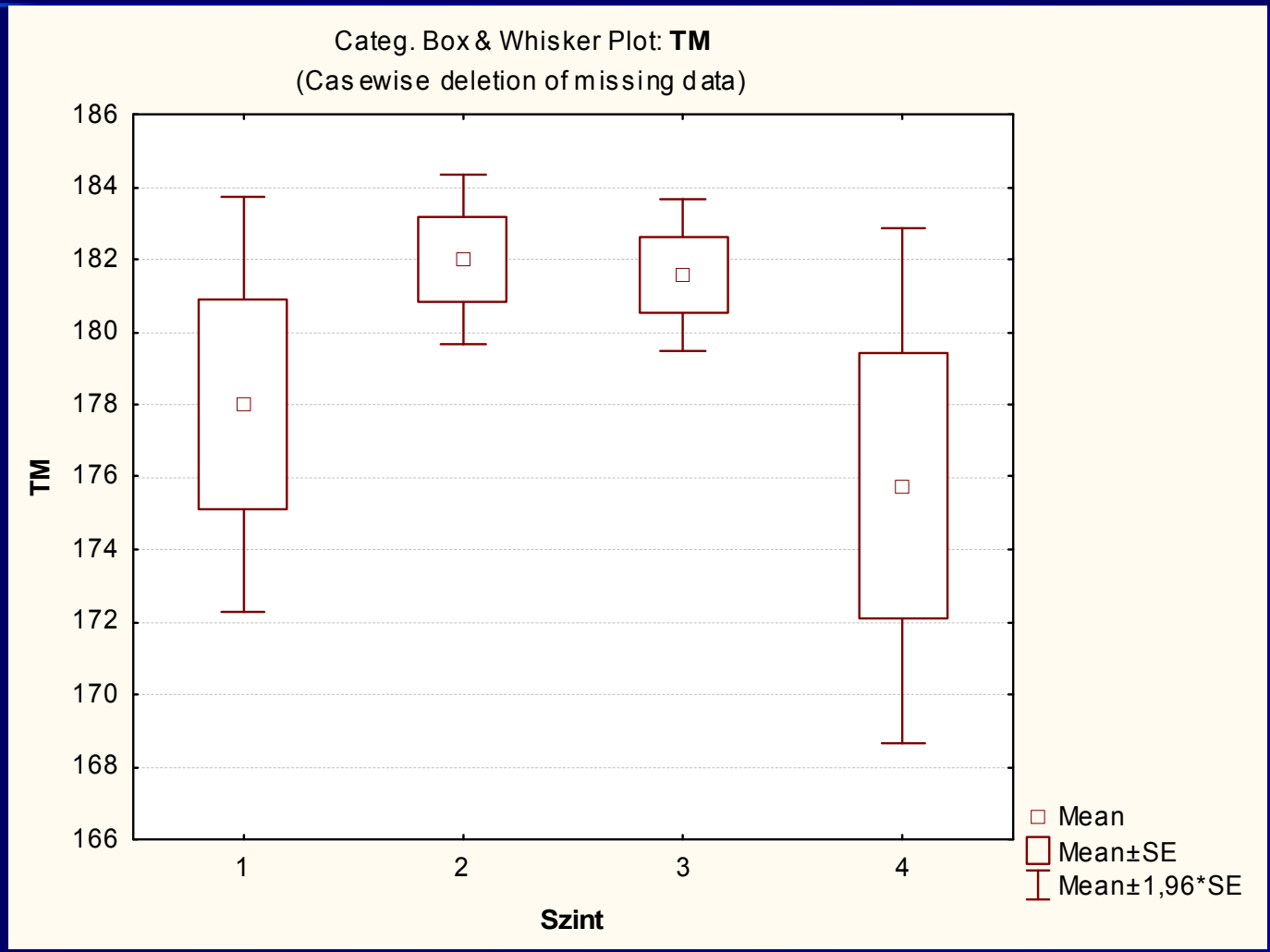
Bivariate histogram

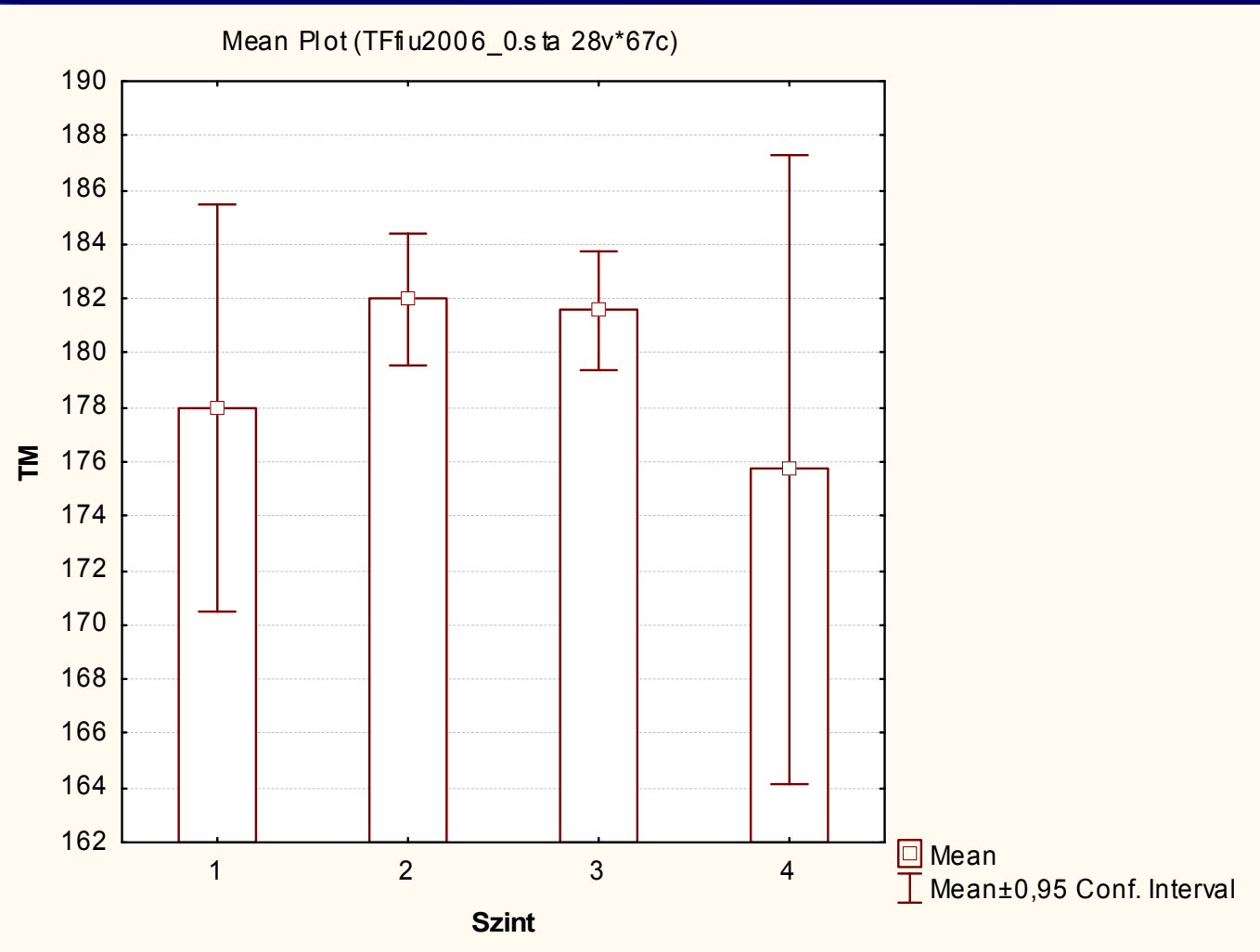


Box plot

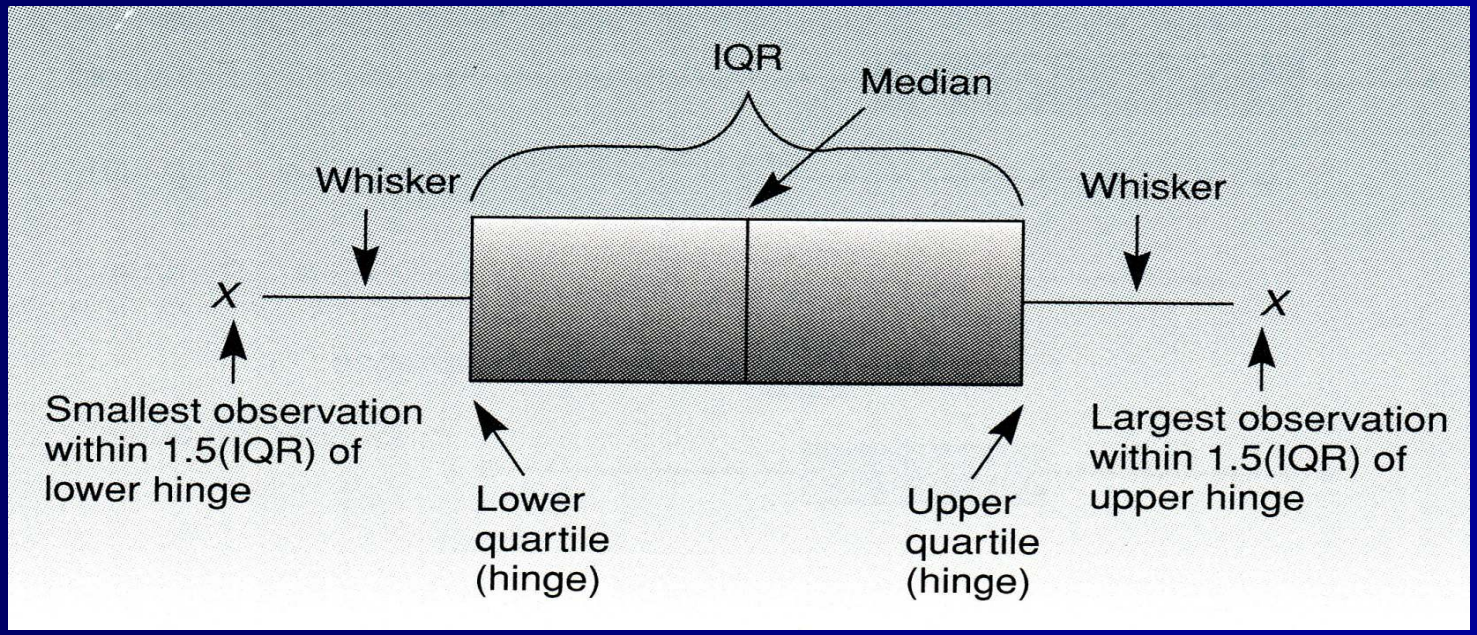
- When representing the results on the charts, the average and the average error or standard deviation is often indicated.
- Claculating all the three parameters mentioned above can be set in statistical programs, and some graph types are able to display all the three values.
- The values are interpreted in ± sense and are generally represented so.


Categ. Box & Whisker Plot: **TM**
(Casewise deletion of missing data)
190
188
186
184
182
180
178
176
174
172
170
168
166
TM
1
2
3
4
Szint
Mean
Mean±SE
Mean±SD

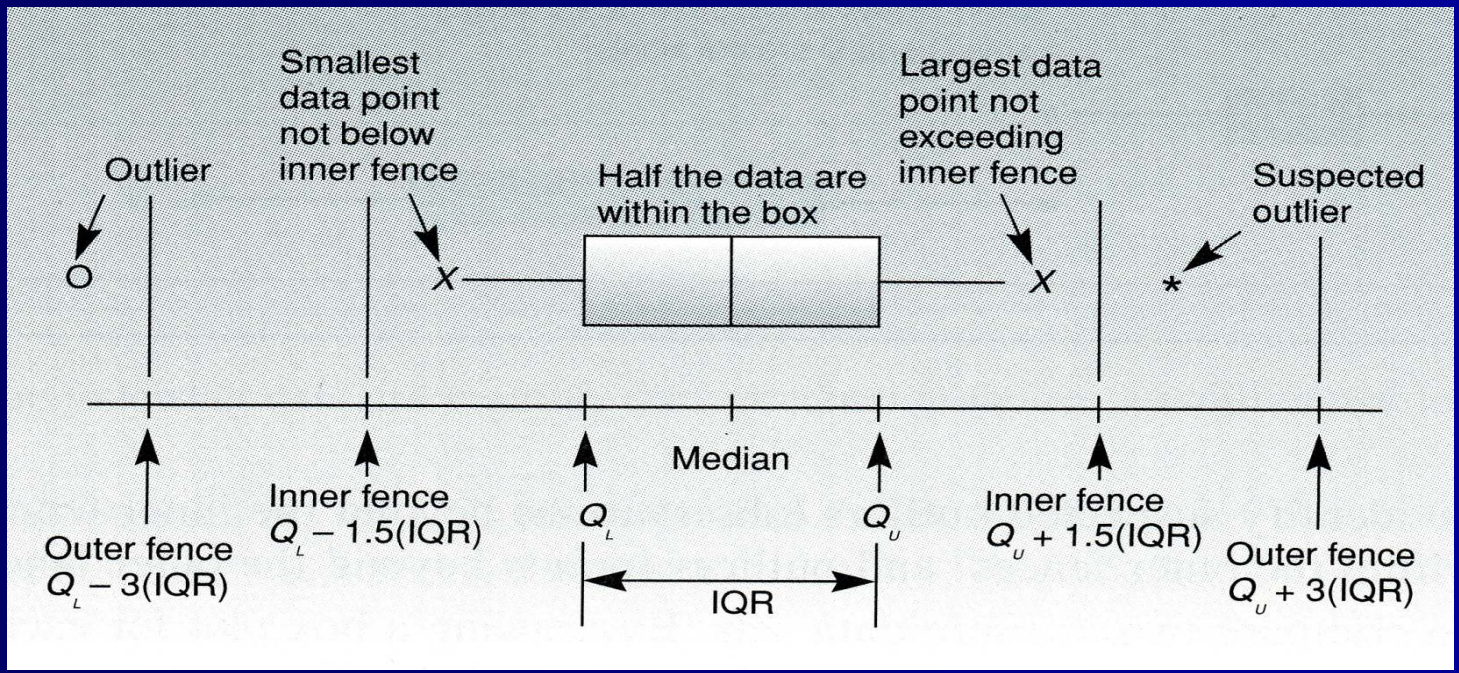

Categ. Box & Whisker Plot: TM
(Casewise deletion of missing data)
186
184
182
180
178
176
174
172
170
168
166
TM
1
2
3
4
Szint
Mean
Mean±SE
Mean±1,96*SE

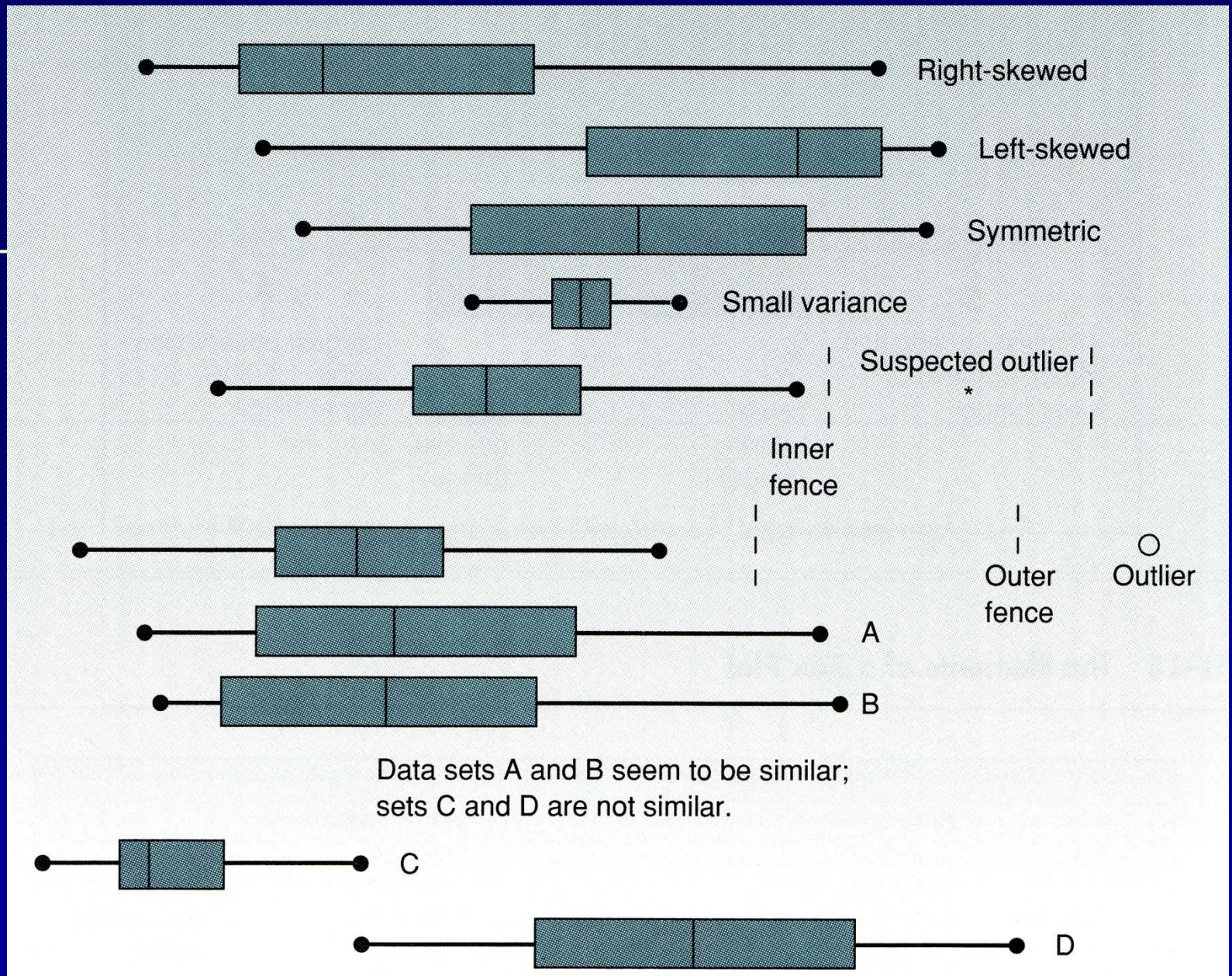

Mean Plot (TFfiu2006_0.sta 28v*67c)
190
188
186
184
182
180
178
176
174
172
170
168
166
164
162
TM
1
2
3
4
Szint
Mean
Mean±0,95 Conf. Interval

Box plot

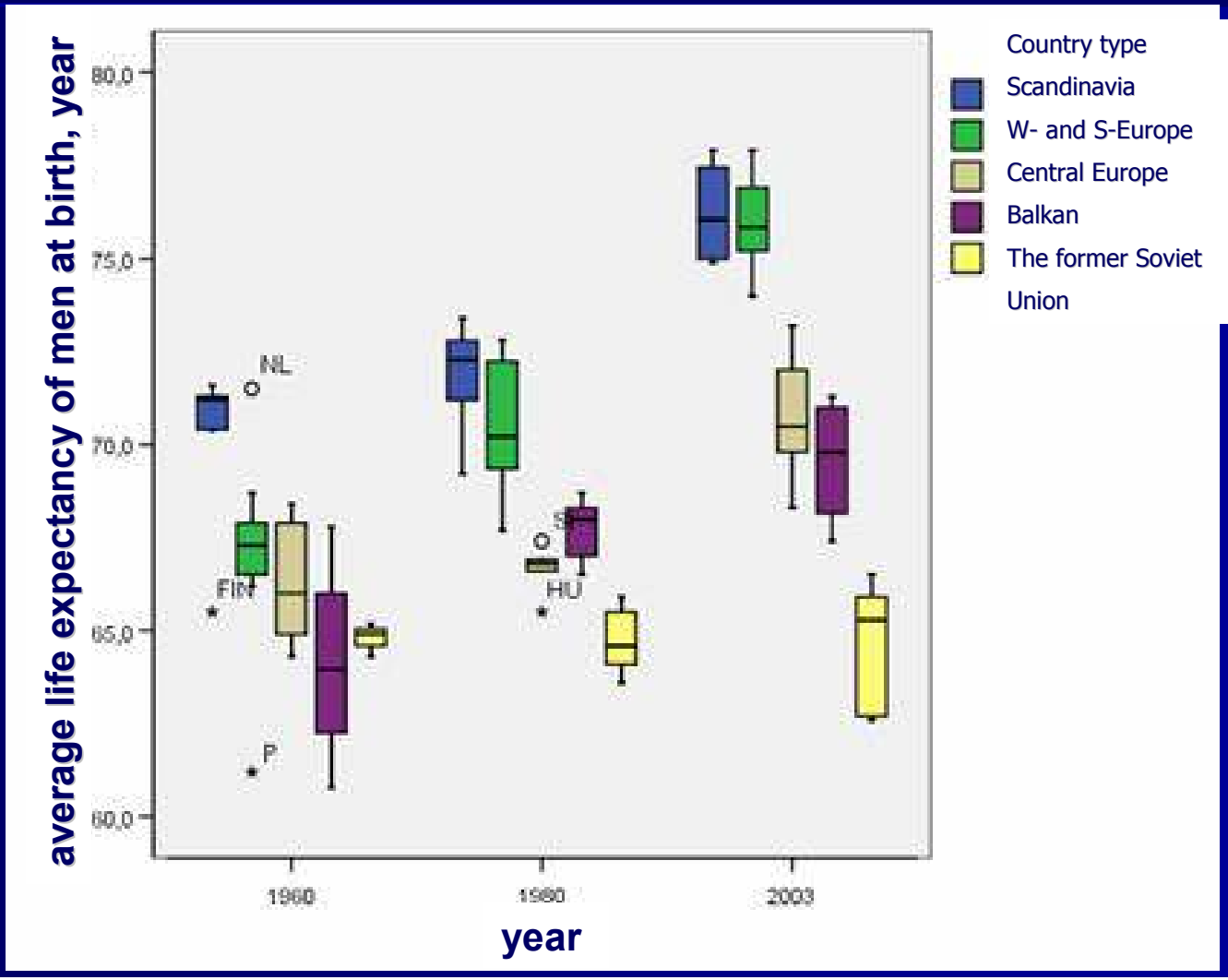

IQR
Median
Whisker
Whisker
X
X
Smallest observation within 1.5(IQR) of lower hinge
Lower quartile (hinge)
Upper quartile (hinge)
Largest observation within 1.5(IQR) of upper hinge

Elemenst of a box plot




Right-skewed
Left-skewed
Symmetric
Small variance
Suspected outlier
*
Inner fence
Outer fence
Outlier
A
B
Data sets A and B seem to be similar;
sets C and D are not similar.
C
D

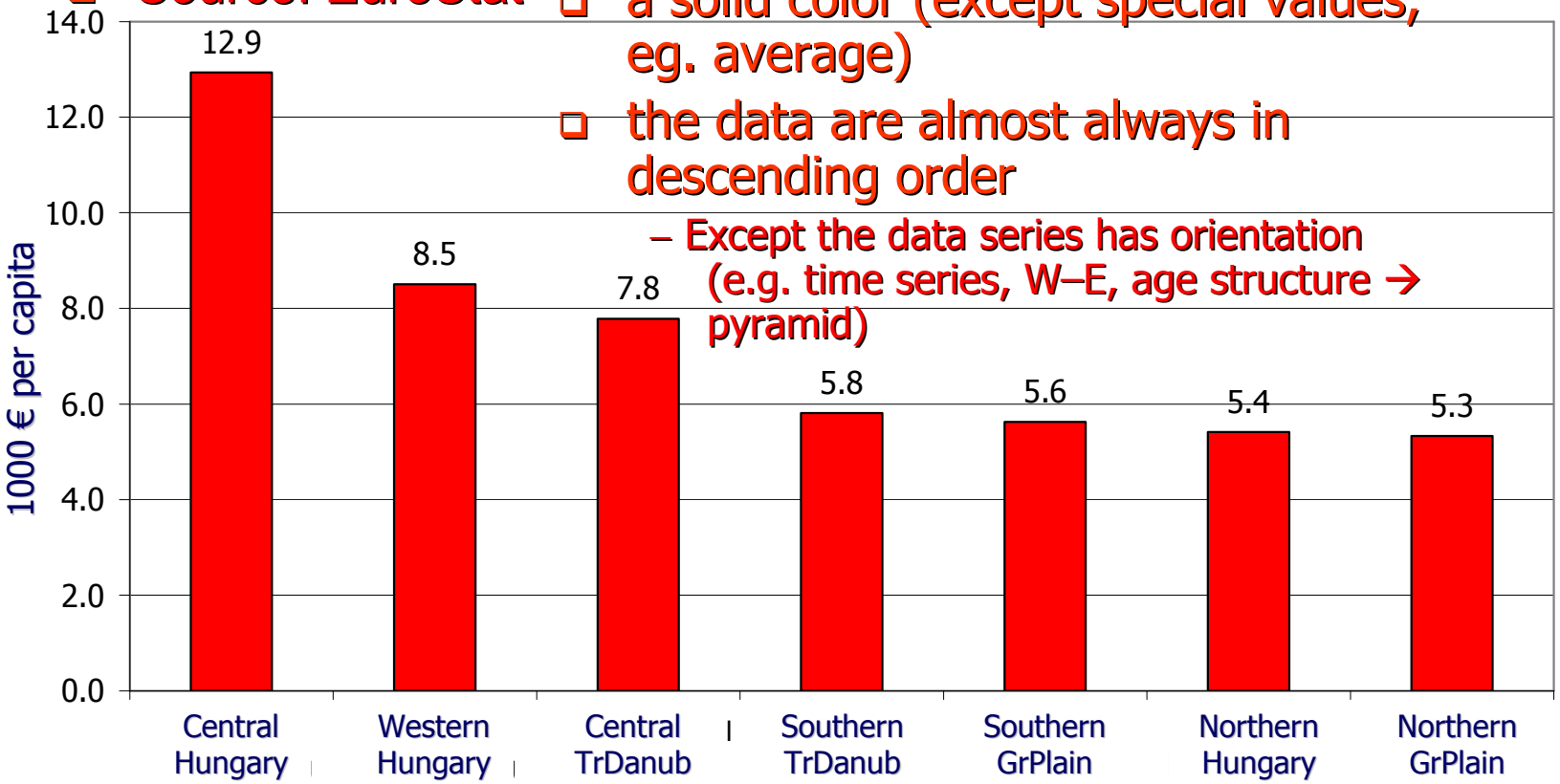
Box plot of average life expectancy of men at birth



Simple bar chart

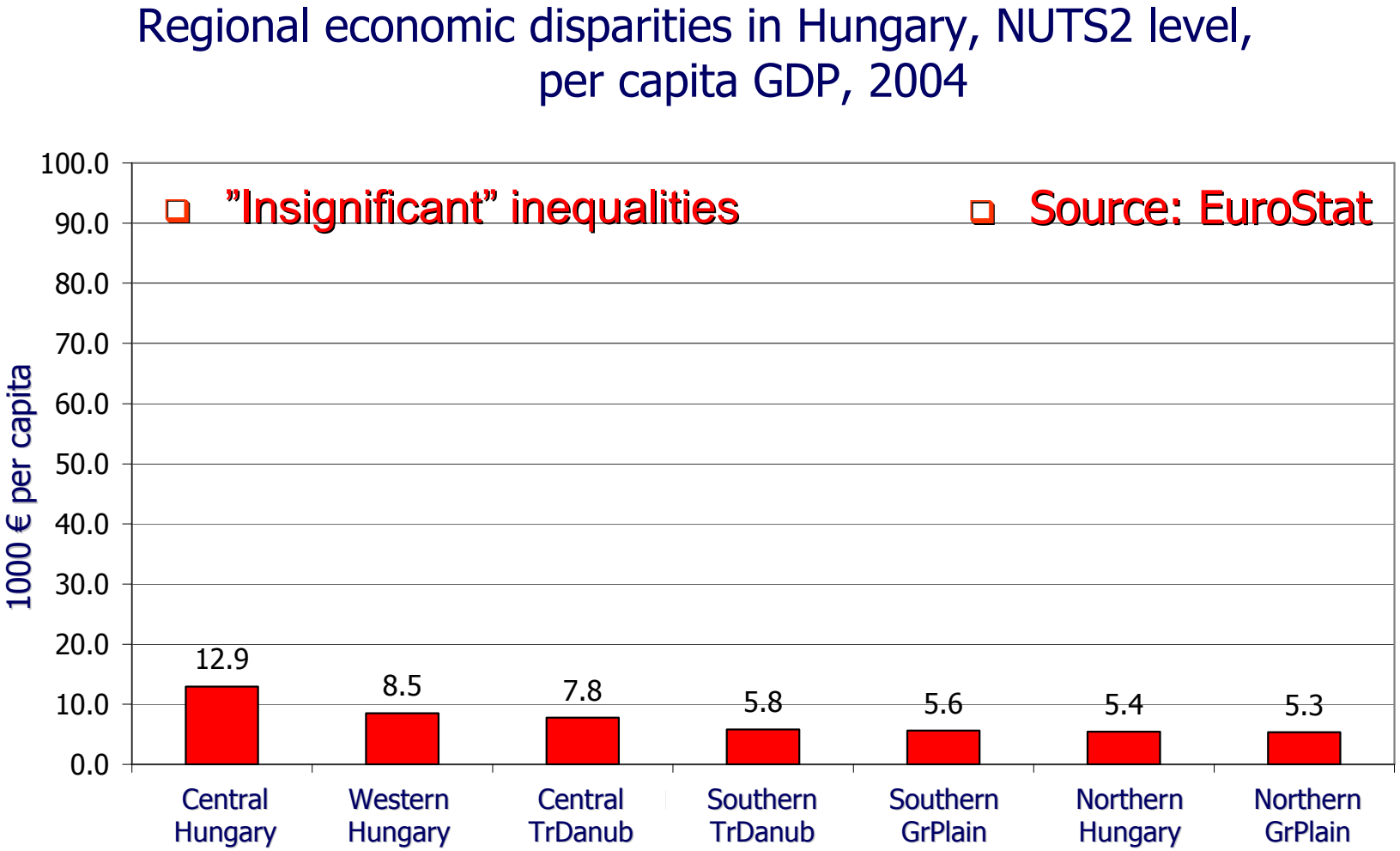
Regional economic disparities in Hungary, NUTS2 level, per capita GDP, 2004

- Source: EuroStat
- a solid color (except special values, eg. average)
- the data are almost always in descending order
  - Except the data series has orientation (e.g. time series, W–E, age structure → pyramid)



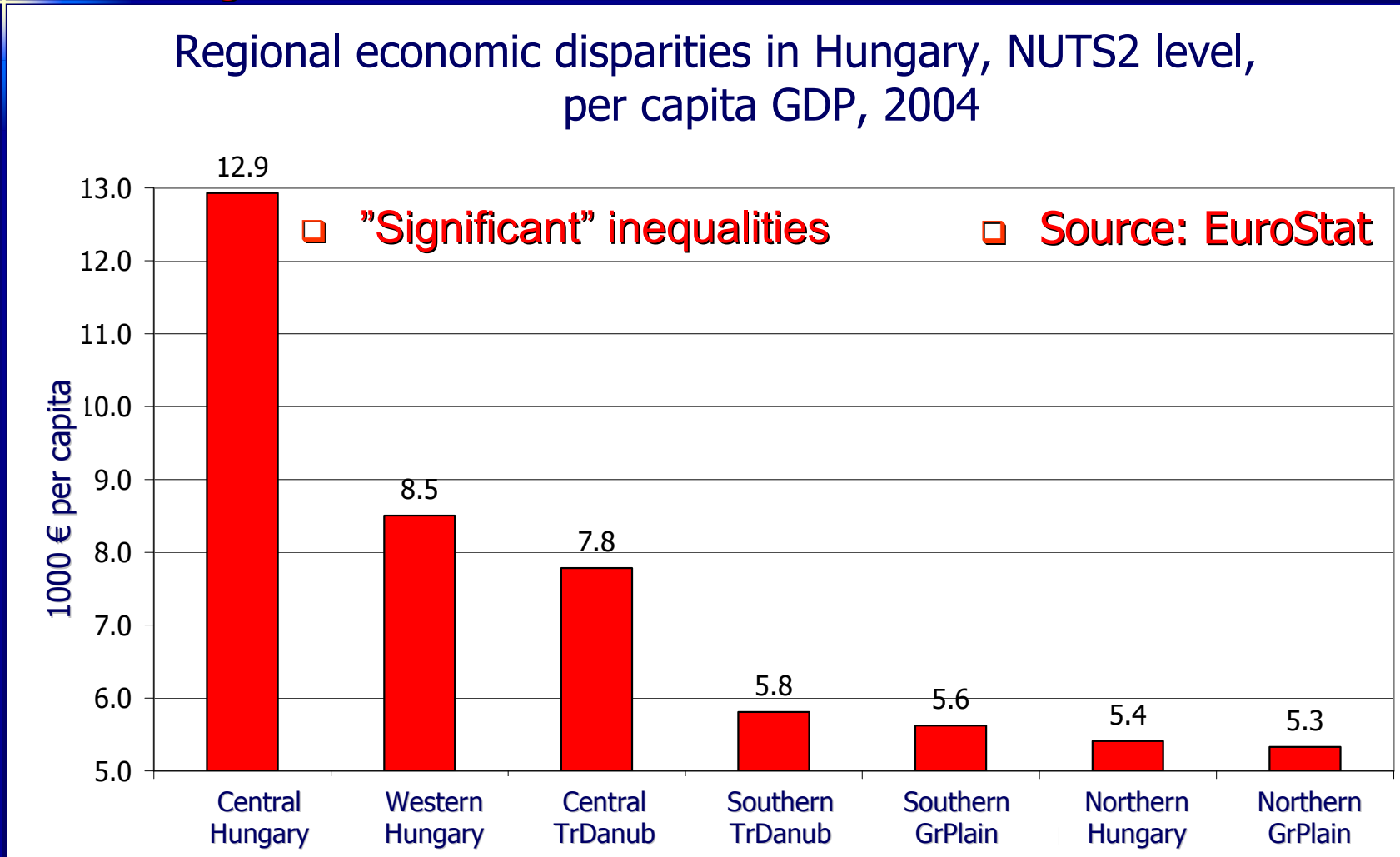
Neutral limits on the category axis

Incorret figure



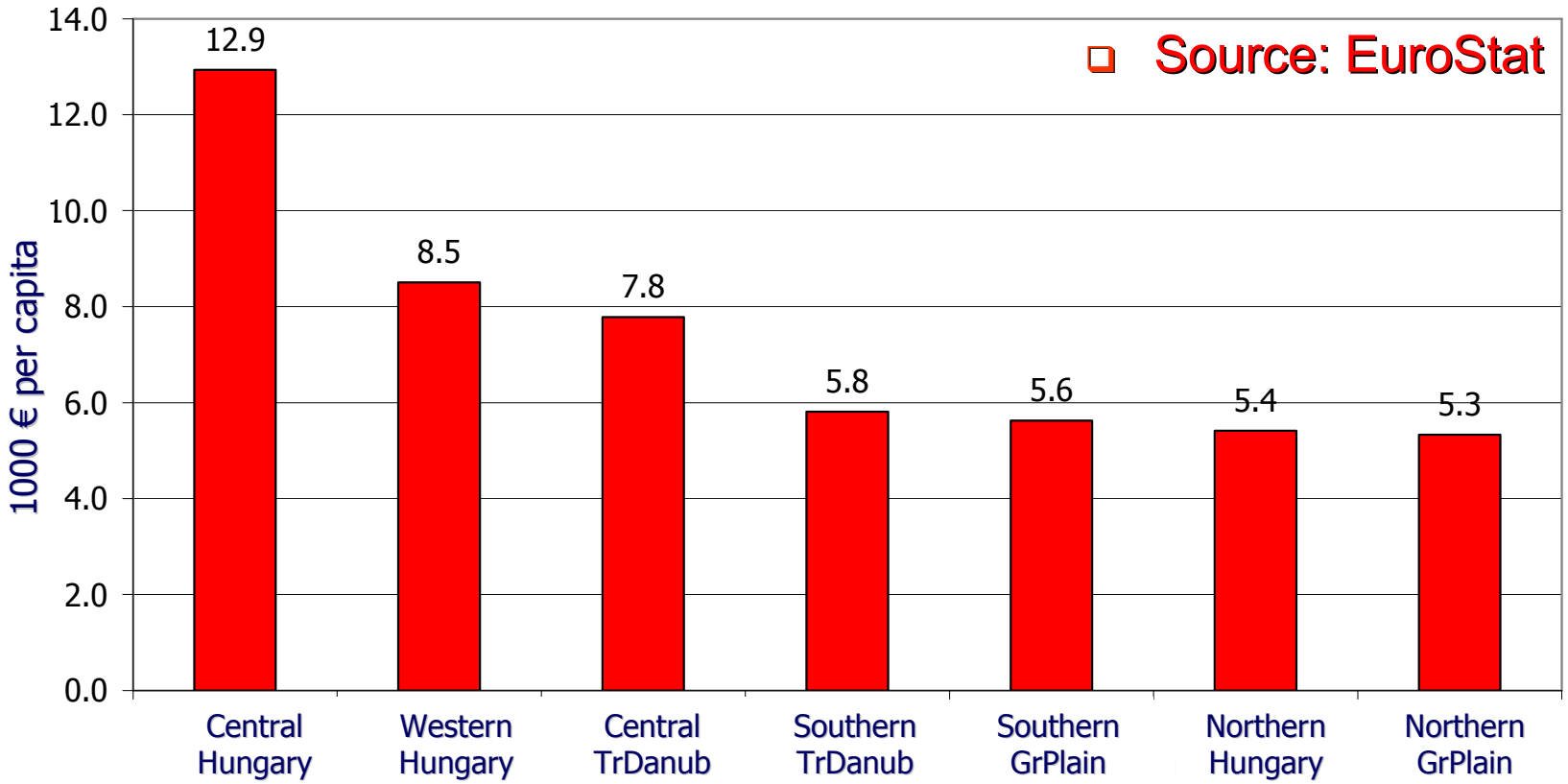
Neutral limits on the category axis

Incorret figure



Criteria of a well-made chart

Regional economic disparities in Hungary, NUTS2 level, per capita GDP, 2004

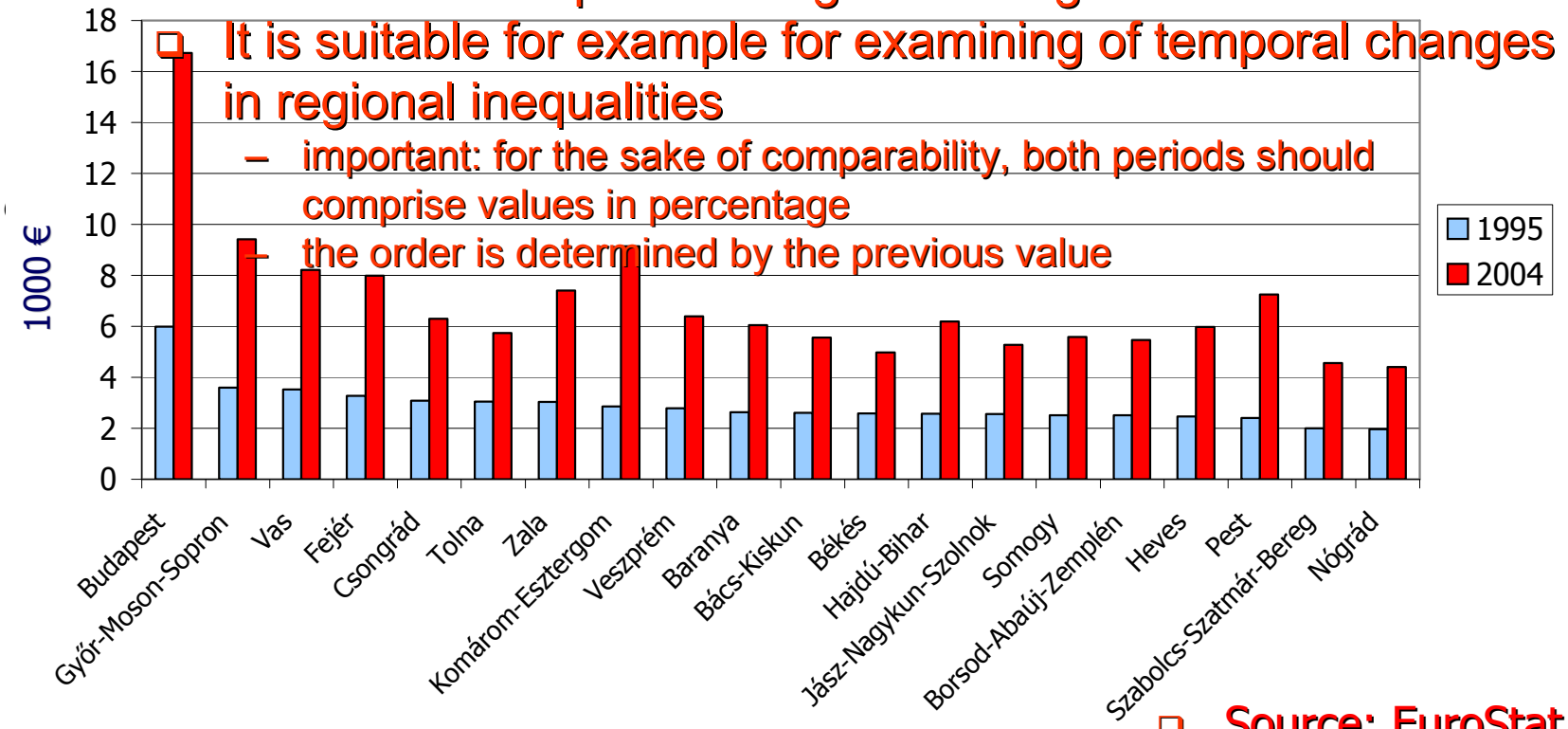


Incorrect clustered bar chart

Incorret figure

Regional economic disparities in Hungary, 1995-2004
(county level, per capita GDP)

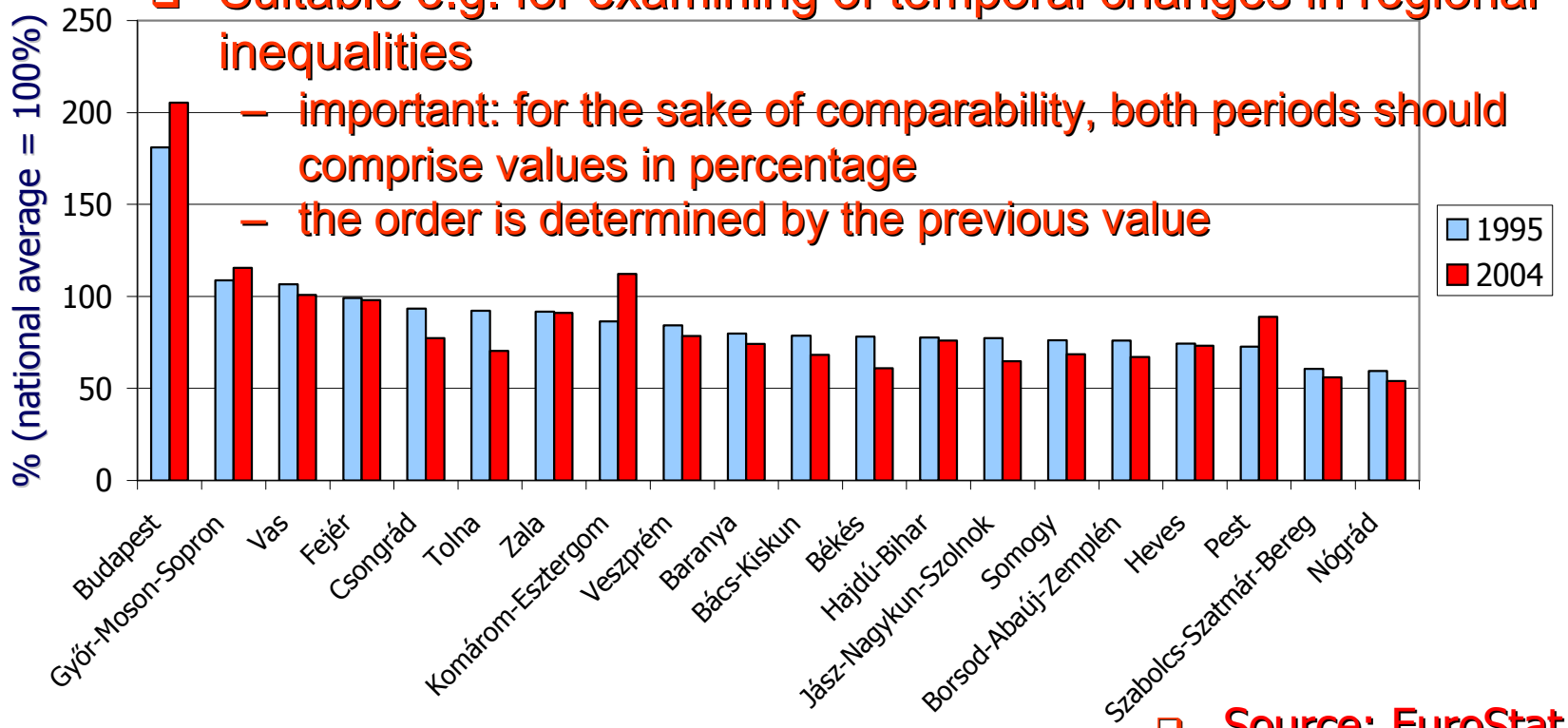
- Multicoloured
- Values are compared along with categories
- It is suitable for example for examining of temporal changes in regional inequalities
  - important: for the sake of comparability, both periods should comprise values in percentage
  - the order is determined by the previous value



- Source: EuroStat

Correct clustered bar chart (divergence in Hungary)

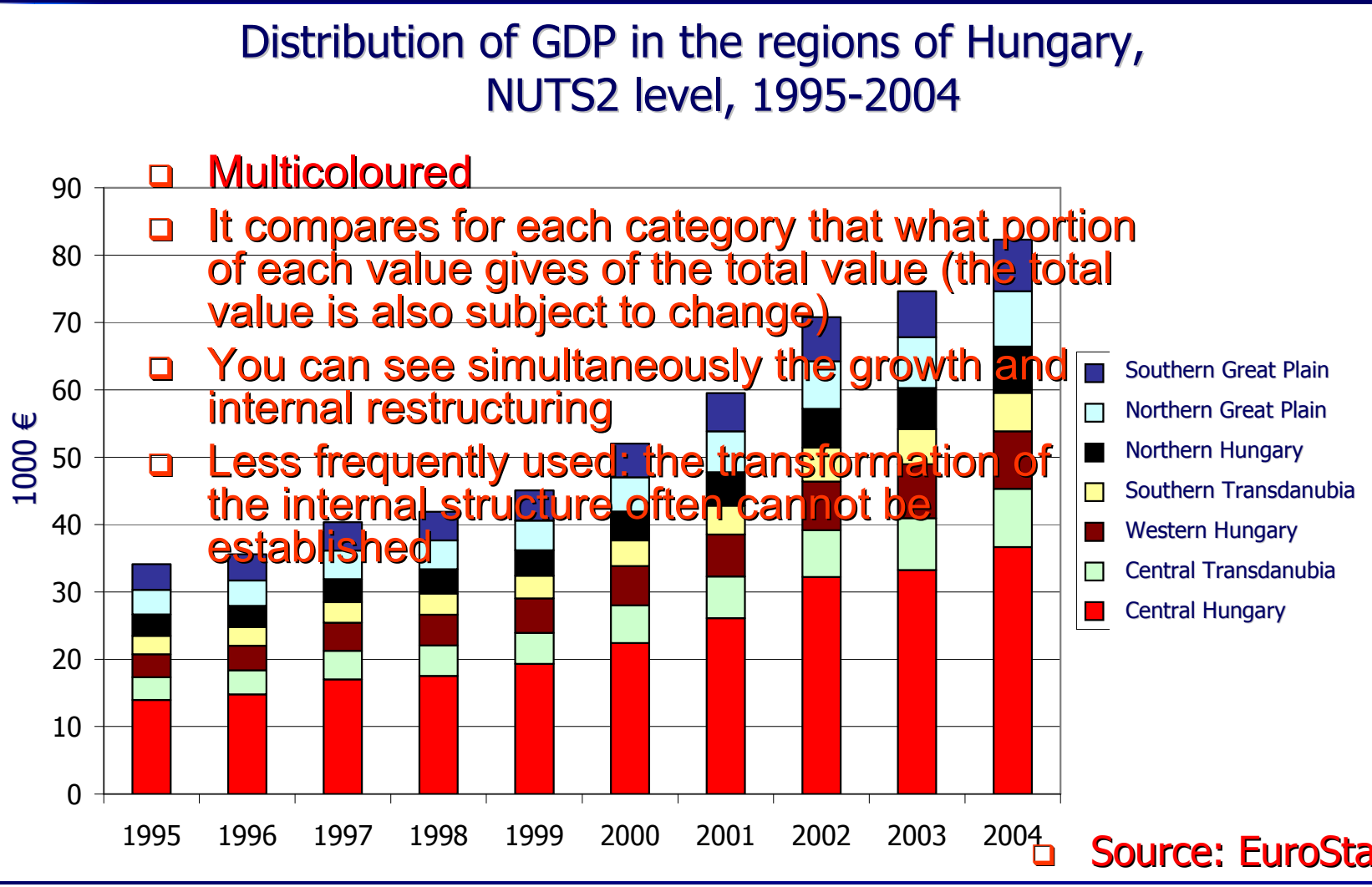
Regional economic disparities in Hungary, 1995-2004
(county level, per capita GDP)



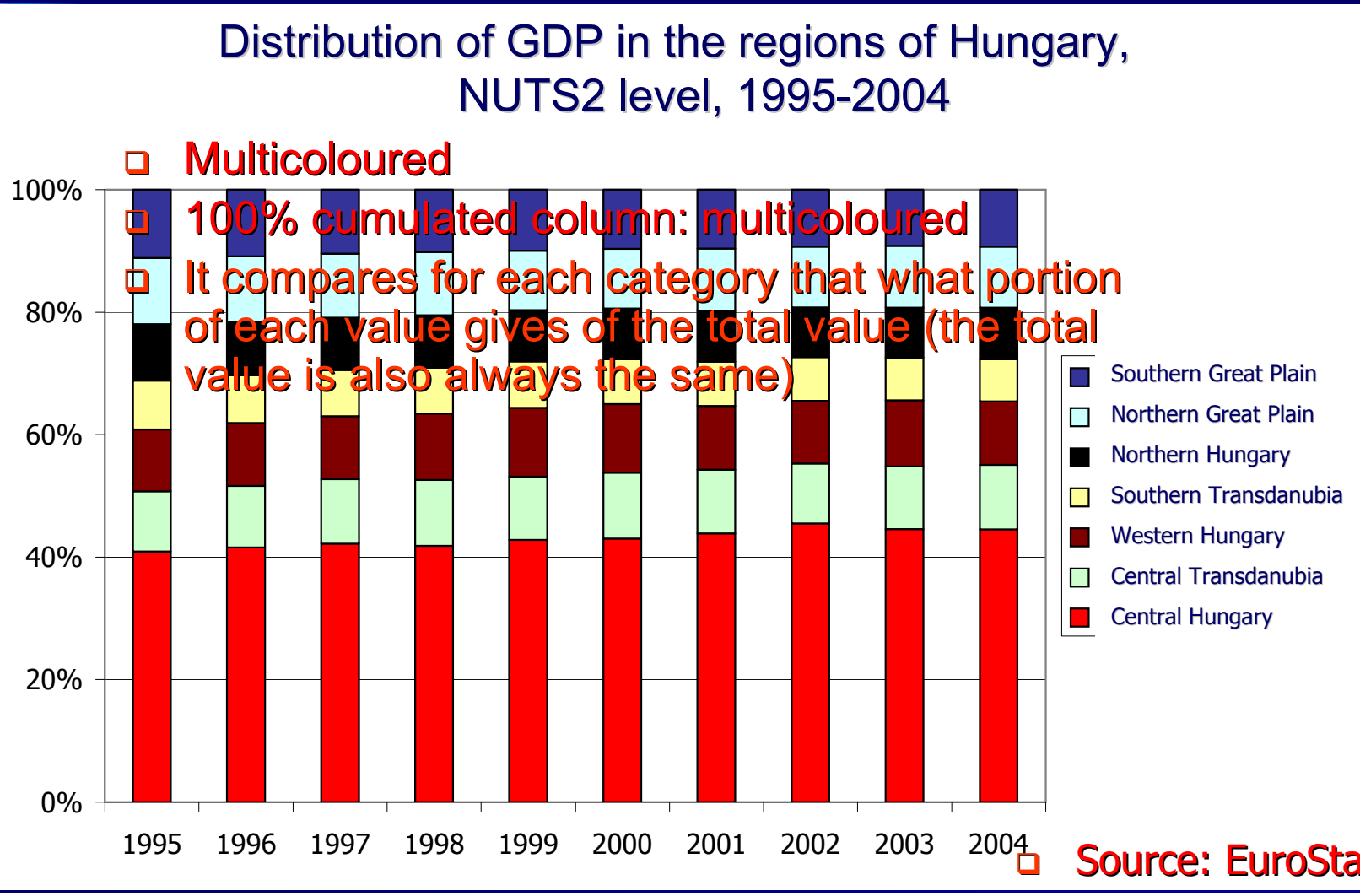
- Suitable e.g. for examining of temporal changes in regional inequalities
  - important: for the sake of comparability, both periods should comprise values in percentage
  - the order is determined by the previous value

- Source: EuroStat

Cumulative bar chart: only for absolute metrics!



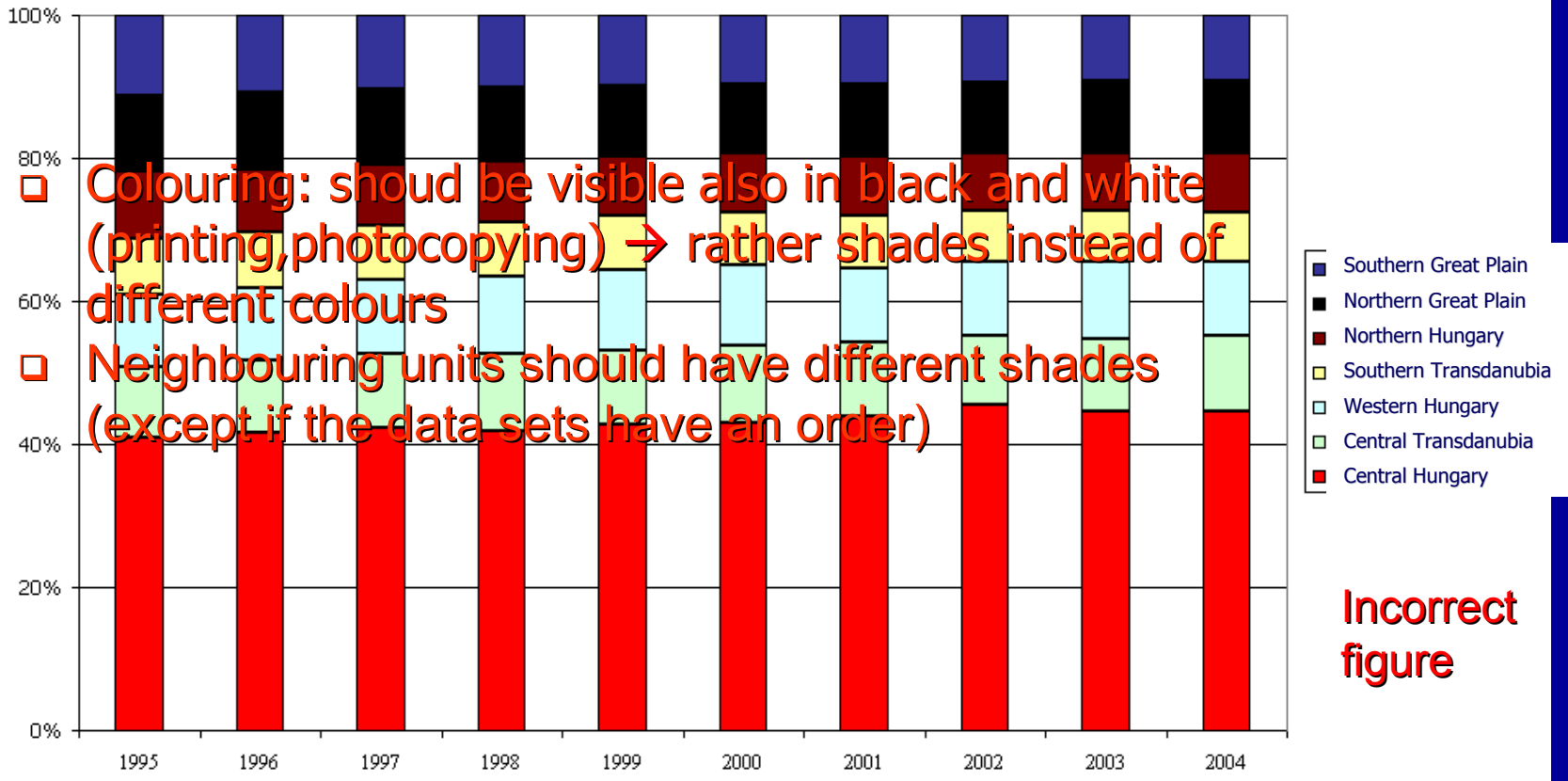
100% cumulated column chart: only for absolute metrics!



Correct colouring is necessary (adjacent units of similar colour do not coincide in black and white)

Distribution of GDP in the regions of Hungary, NUTS2 level, 1995-2004

- Source: EuroStat



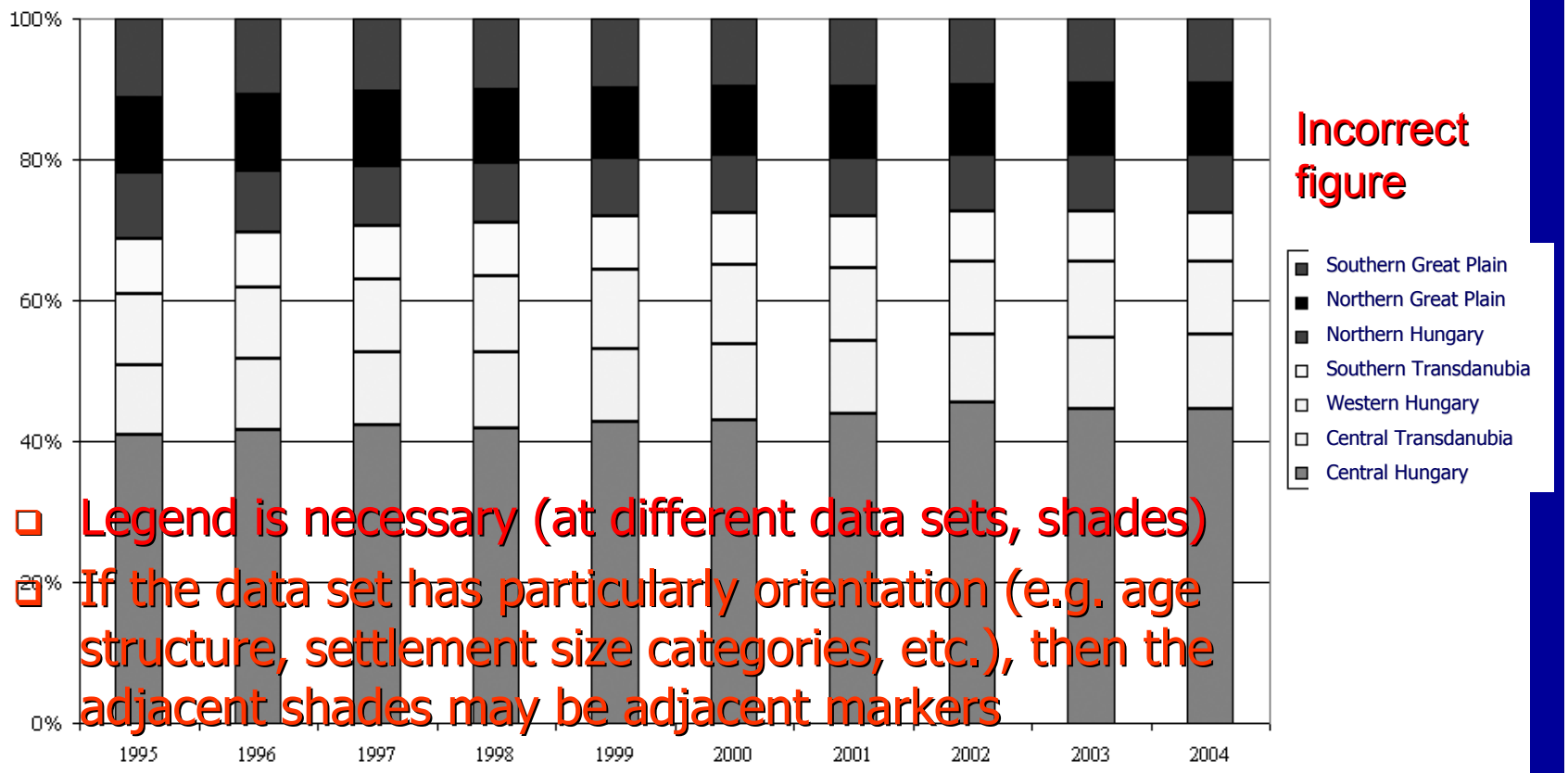
- Colouring: shoud be visible also in black and white (printing,photocopying) → rather shades instead of different colours
- Neighbouring units should have different shades (except if the data sets have an order)

Incorrect figure

Correct colouring is necessary (adjacent units of similar colour do not coincide in black and white)

Distribution of GDP in the regions of Hungary, NUTS2 level, 1995-2004

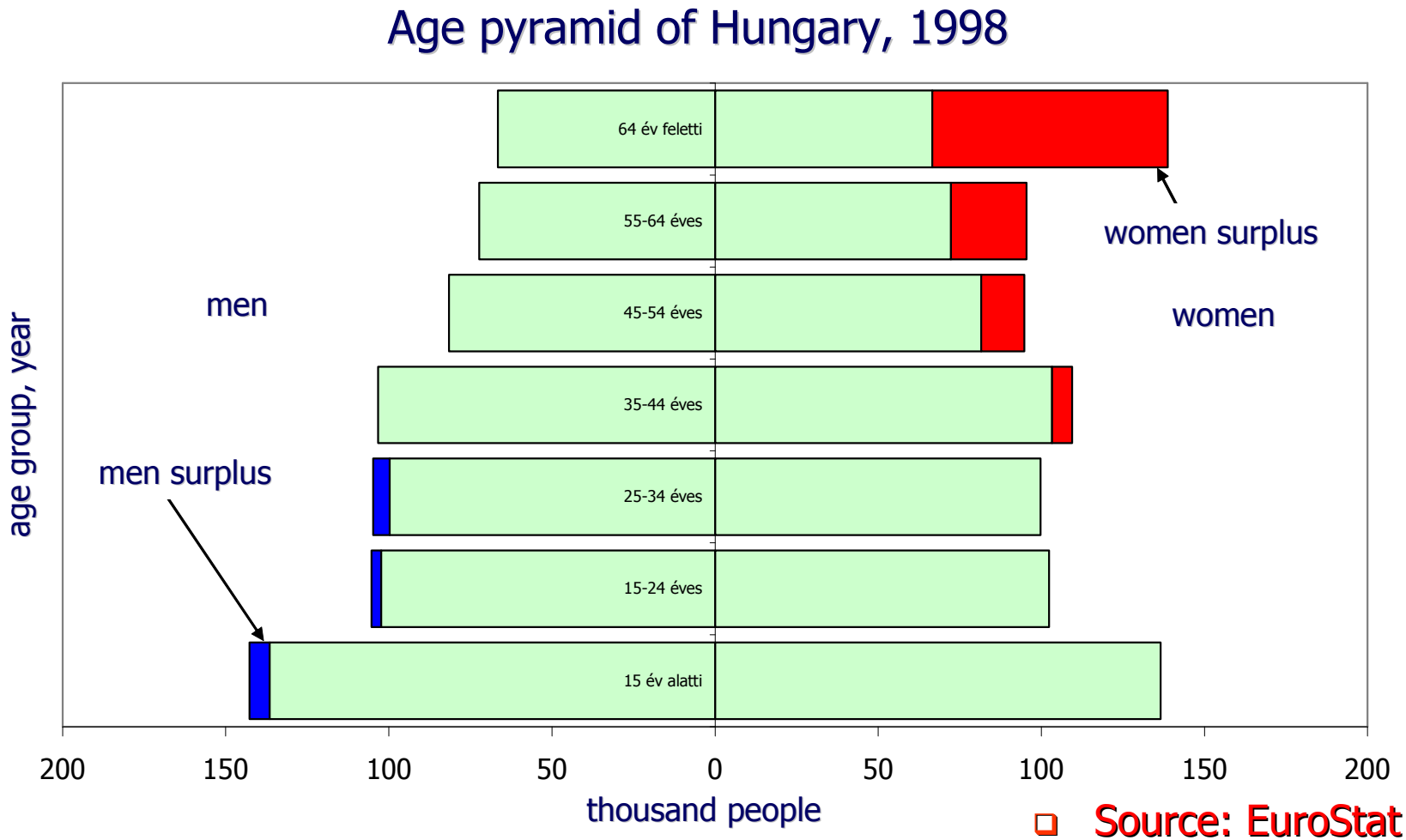
- Source: EuroStat



Incorrect figure

- Legend is necessary (at different data sets, shades)
- If the data set has particularly orientation (e.g. age structure, settlement size categories, etc.), then the adjacent shades may be adjacent markers

Cumulative bar chart; its special type: age pyramid

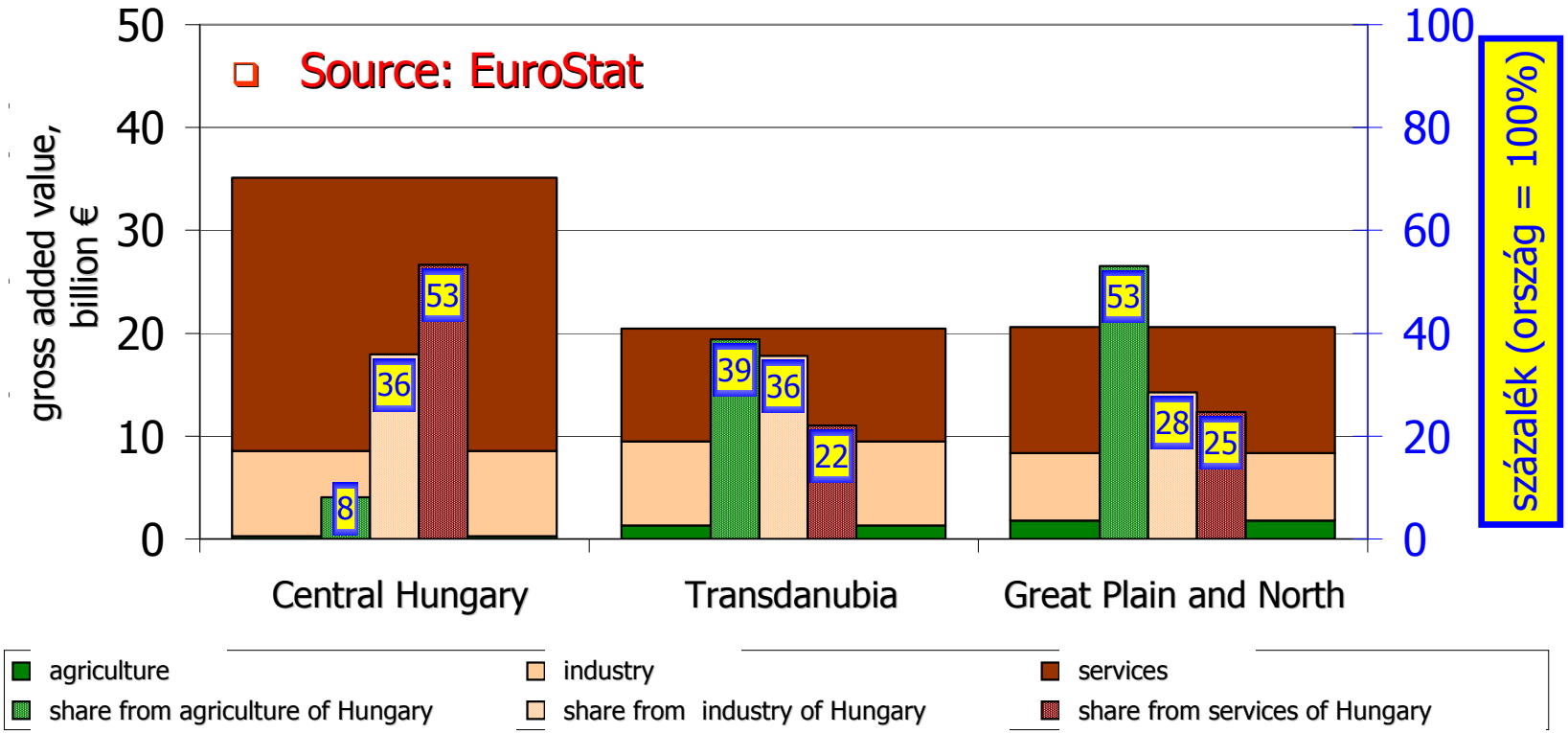


- Source: EuroStat

It should be target-oriented, clear, not too complex (represent rather on separate diagrams)!

Economy of the regions of Hungary by sectors and the proportion of the regions from the sectors on the basis of the gross added value, NUTS1 level, 2004

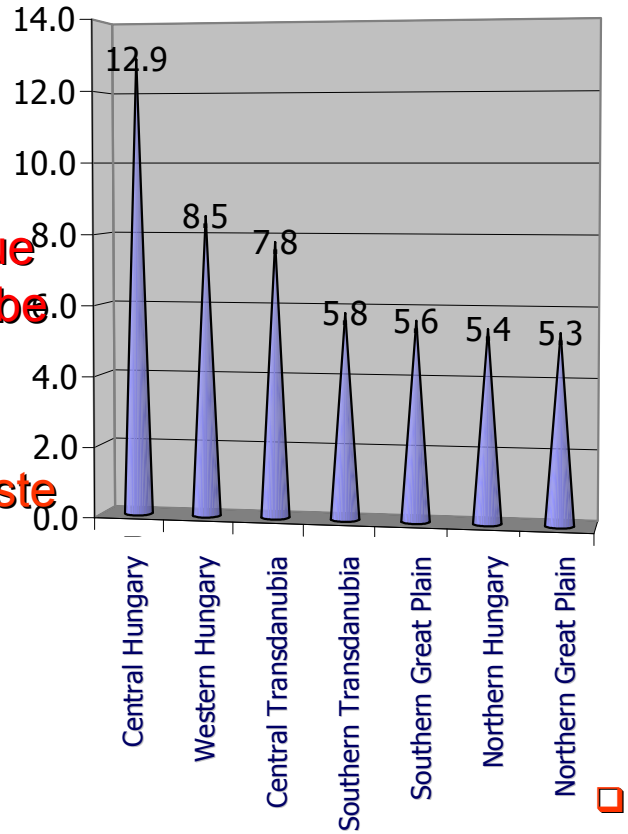
Incorrect figure



Let it be simple, not overly designed!

Regional economic disparities in Hungary, NUTS2 level, per capita GDP, 2004

- Avoid coloured or patterned background
  - white background is more elegant
  - Old Excel: greyish-blue indicators can hardly be seen in grey background
  - It is unnecessary "waste toner" when printing
- Columns, instead of cones
- 2D instead of 3D



Incorrect figure

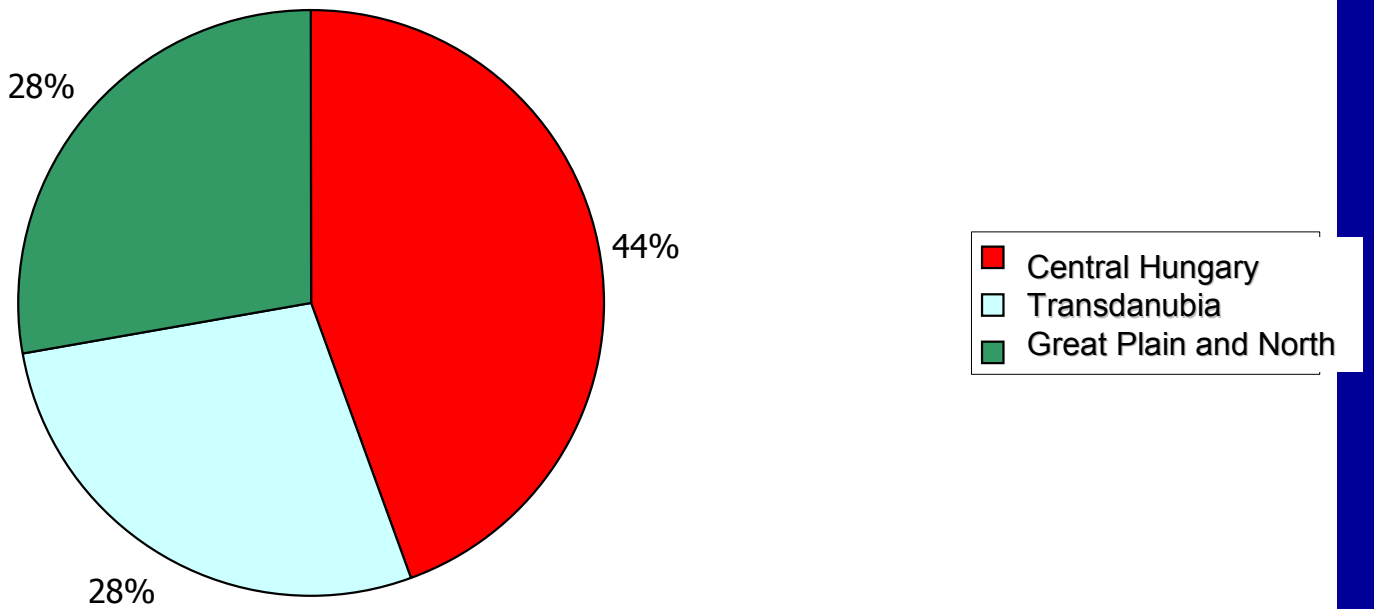
- Source: EuroStat

Pie chart: only for absolute metrics!

- Not recommended (instead vertical bars);
- It is difficult to measure the changes of the composition (pretzel diagram);
- It should not comprise more than 3-4 sectors (smaller values should be merged → other category);
- Utterly wrong :
  - 3D, exploded Circle;

A better pie chart

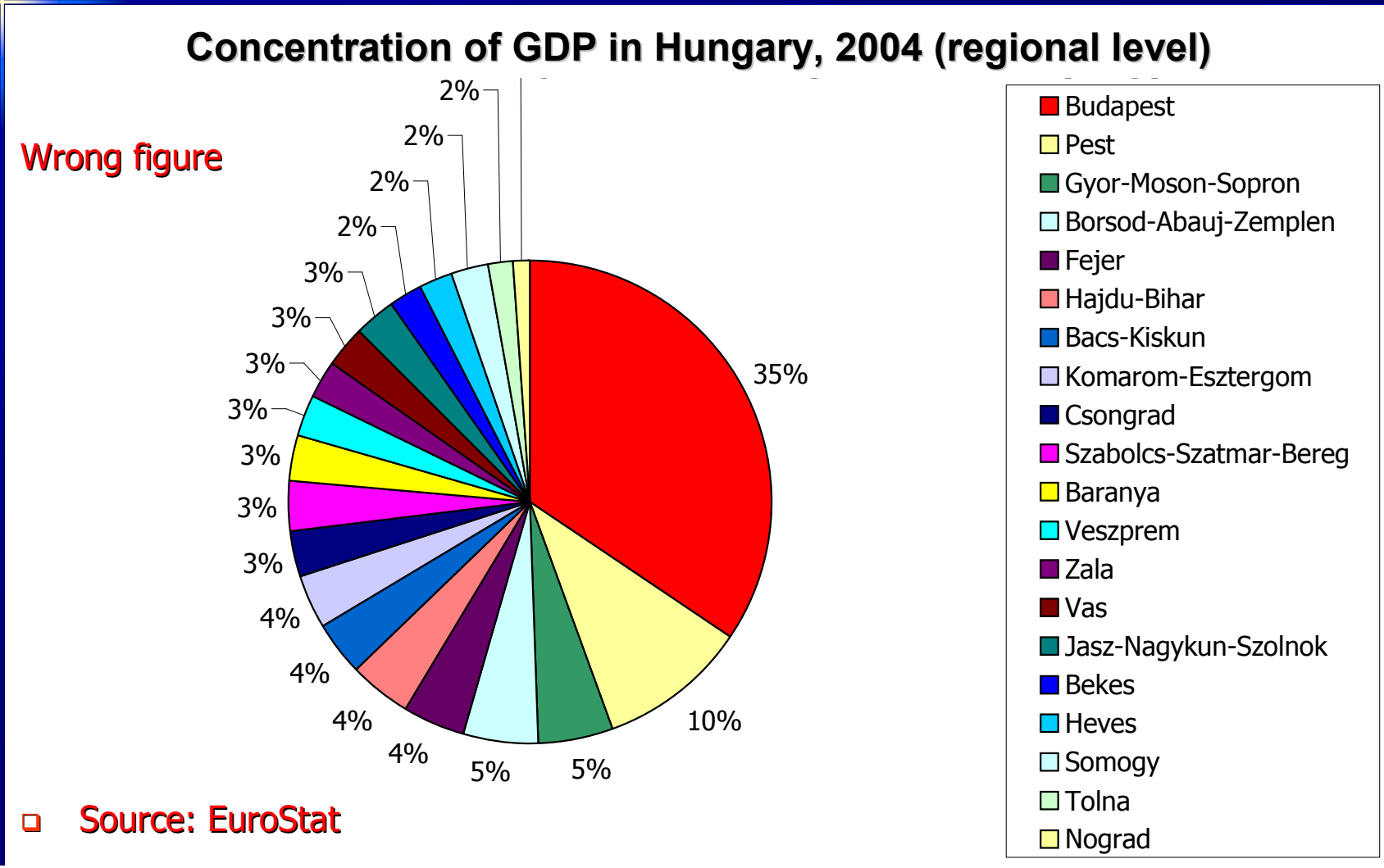
GDP in regions of Hungary, 2004 (NUTS1 level)



- Source: EuroStat

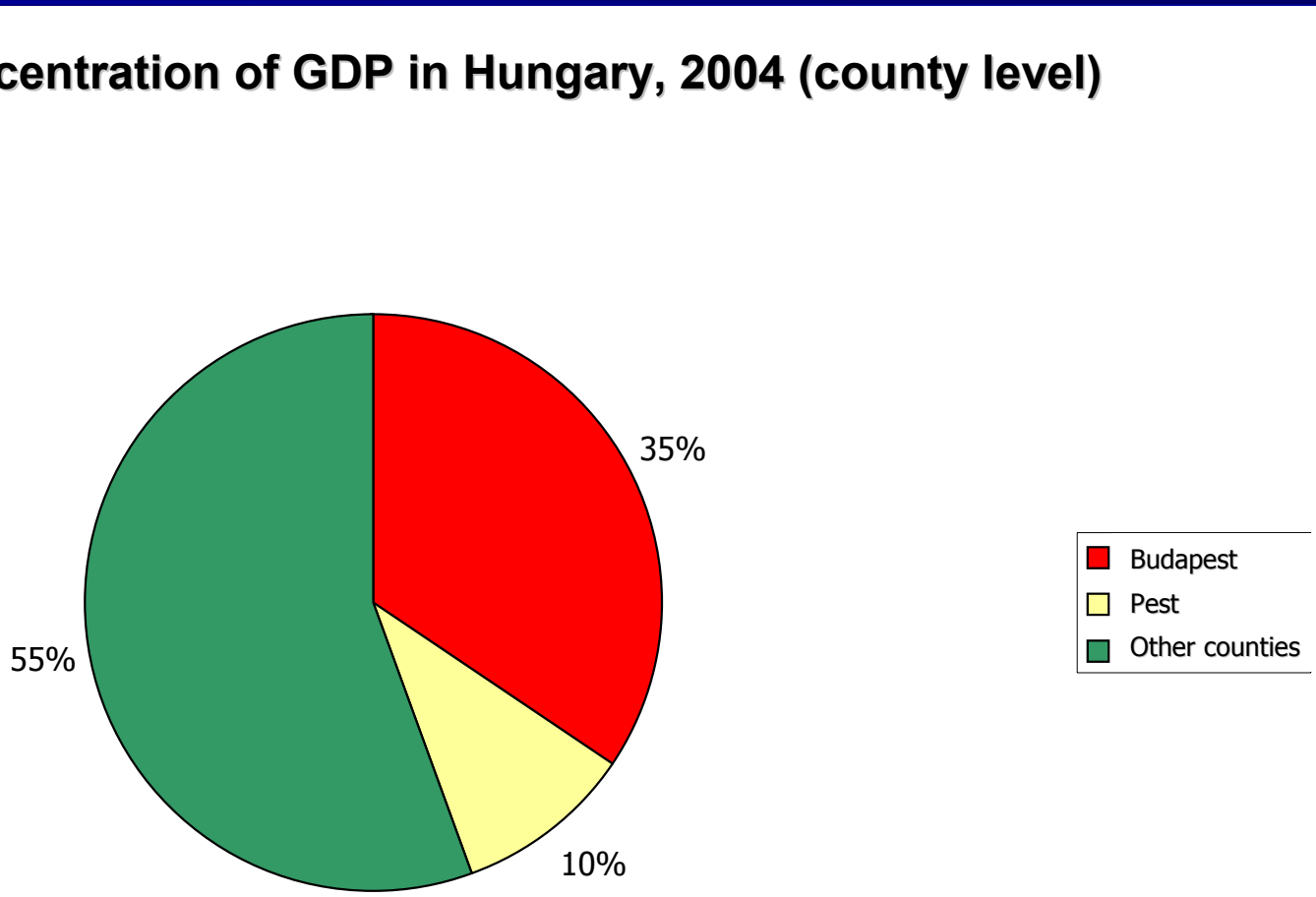
Wrong pie chart (let it be no more than 3-4 quadrants)

Wrong figure



- Source: EuroStat

Better pie chart: at a higher element number it should be aggregated



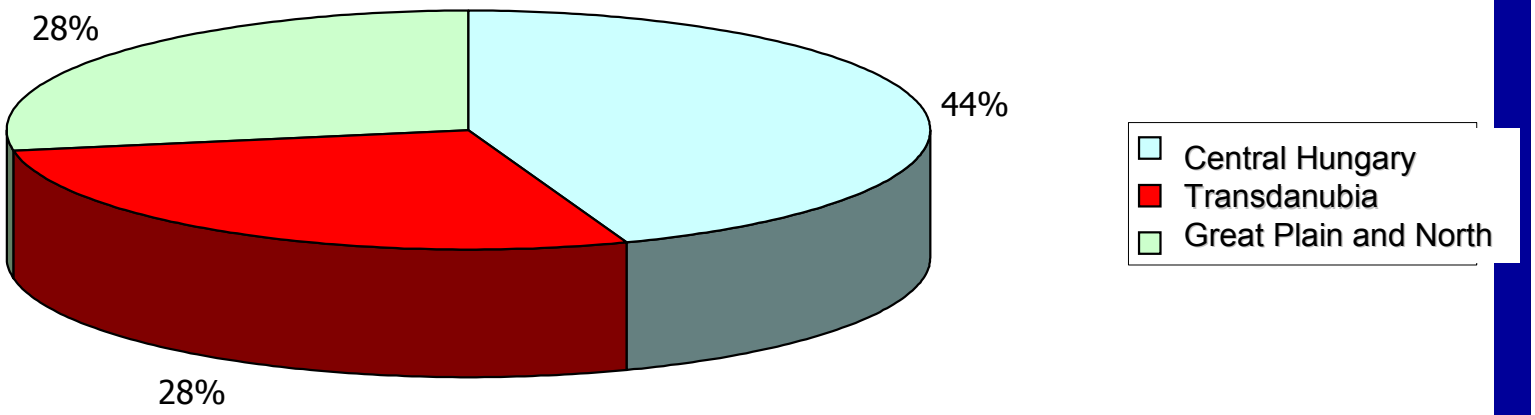
- Source: EuroStat

Let it be optically neutral, avoiding spatial impact (3D-t)!

Distribution of GDP in Hungary, 2004 NUTS1 level)

Wrong figure

- Source: EuroStat

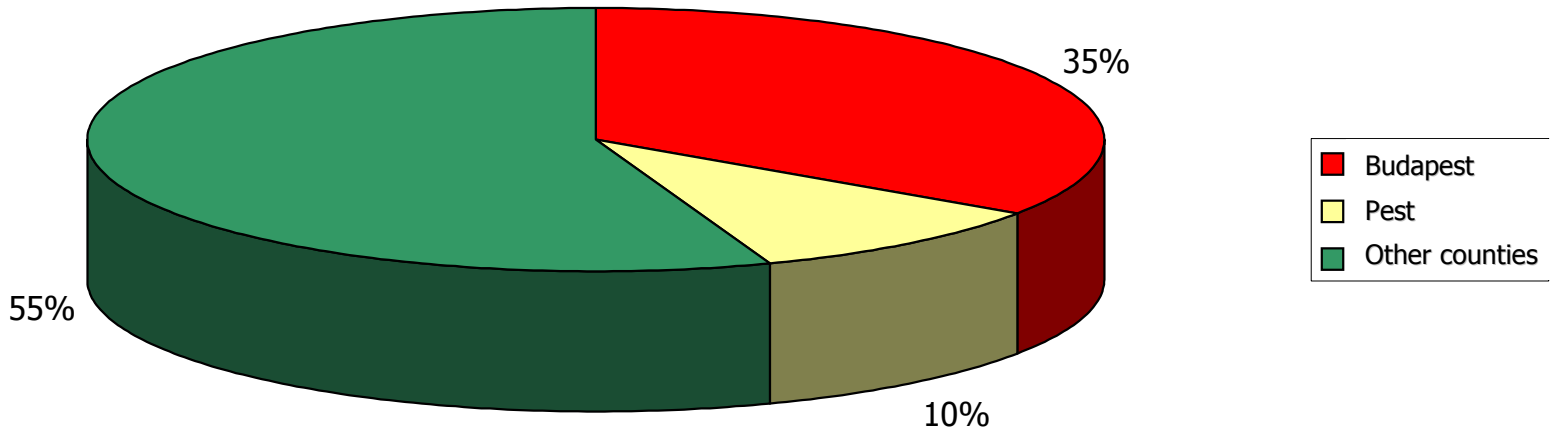


- Dimensional pie chart is not good;
  - Sections in foreground appear larger;
  - Stereoscopy, frivolous design (also for bar charts);

Wrong pie chart (not to be stereoscopic)

Concentration of GDP in Hungary, 2004 (county level)

Wrong figure

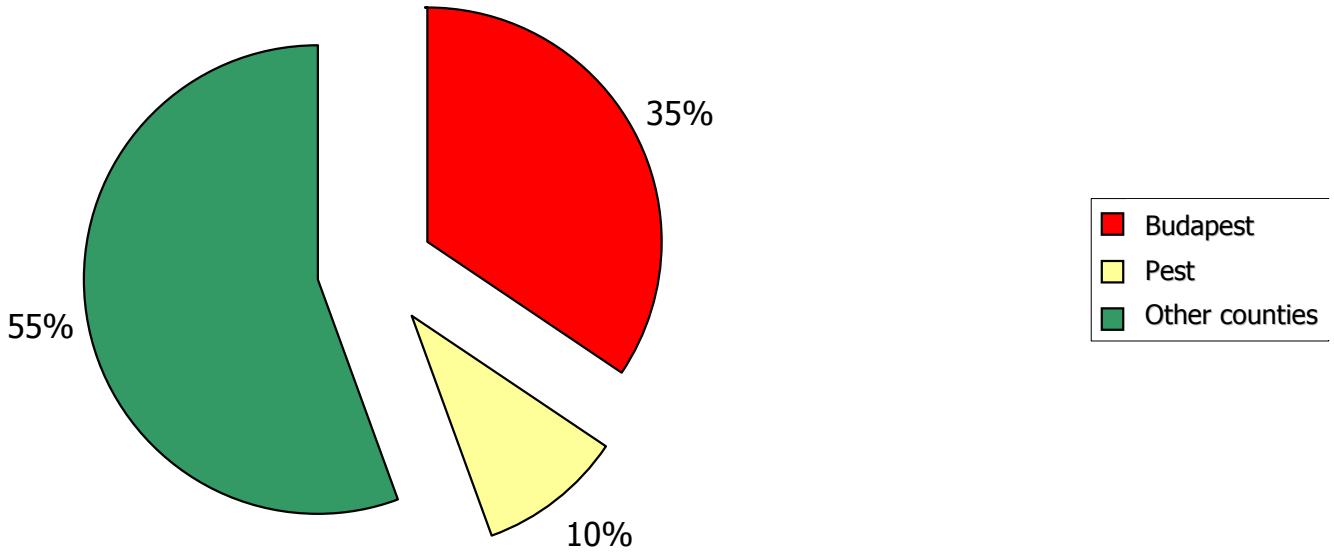


- Source: EuroStat

Wrong pie chart (not to be exploded)

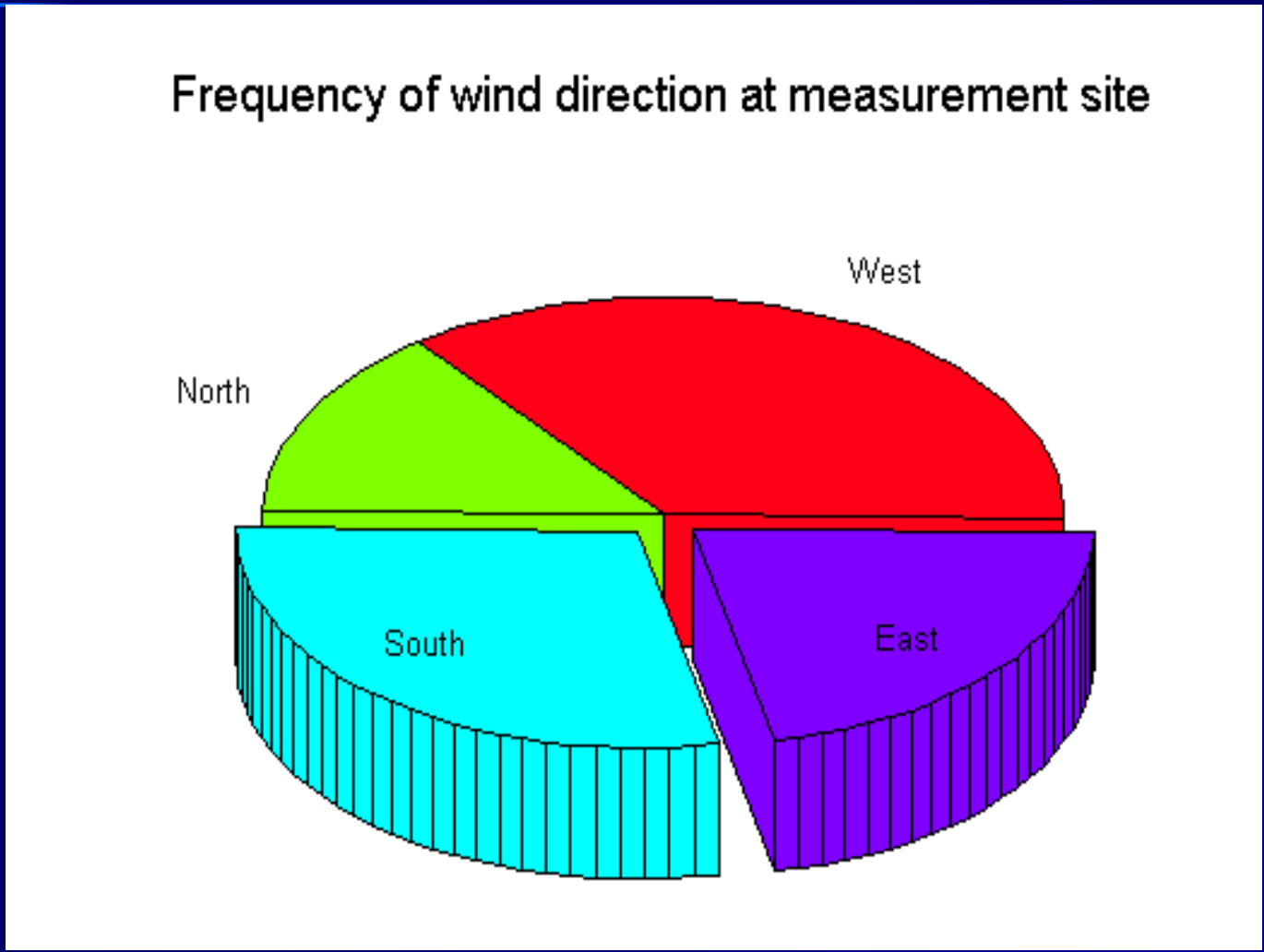
Concentration of GDP in Hungary, 2004 (county level)

Wrong figure

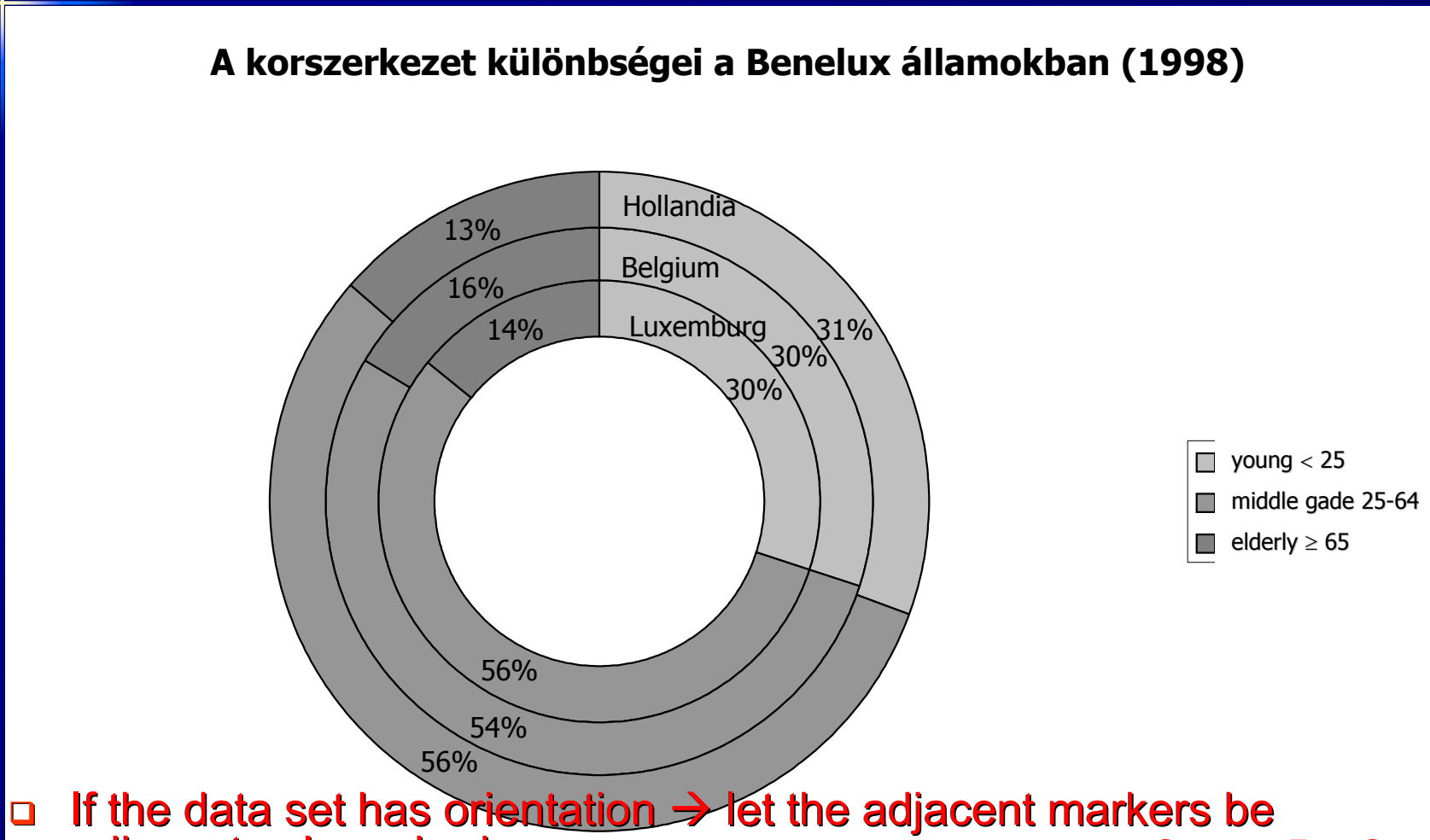


- Source: EuroStat

Wrong pie chart (an inadequate way of representation was chosen to the given criterion)



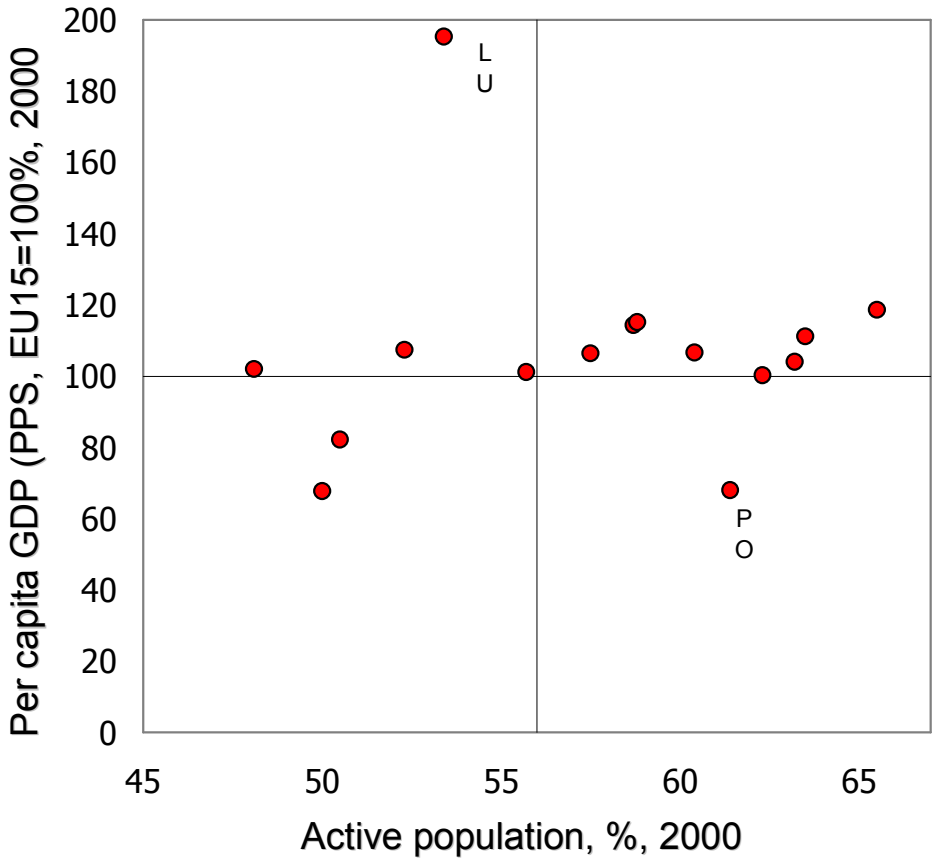
A better pie chart: colouring



- If the data set has orientation → let the adjacent markers be adjacent colour shades
- Source: EuroStat

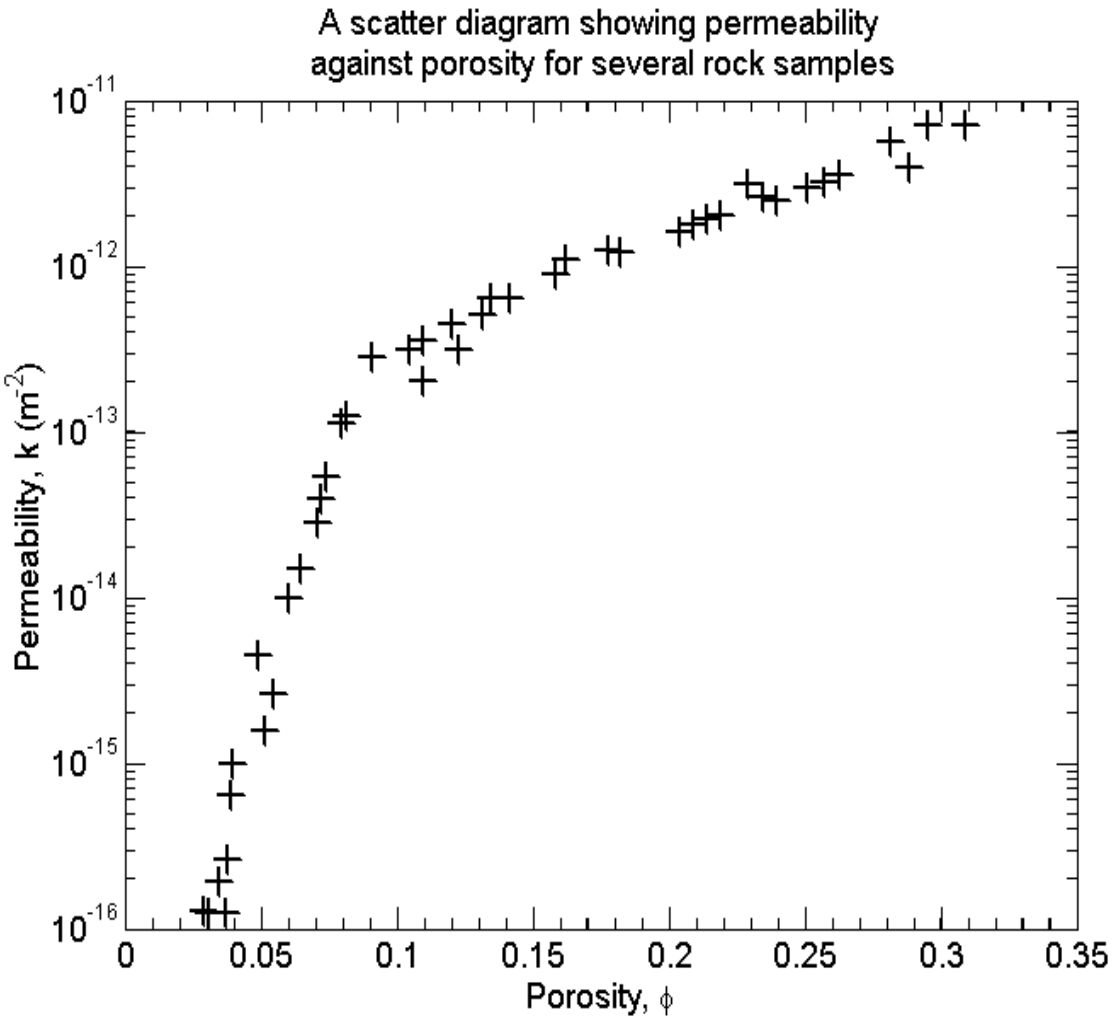
Score chart: two-domensioanl comparison

Classification of the EU Member States according to two indices

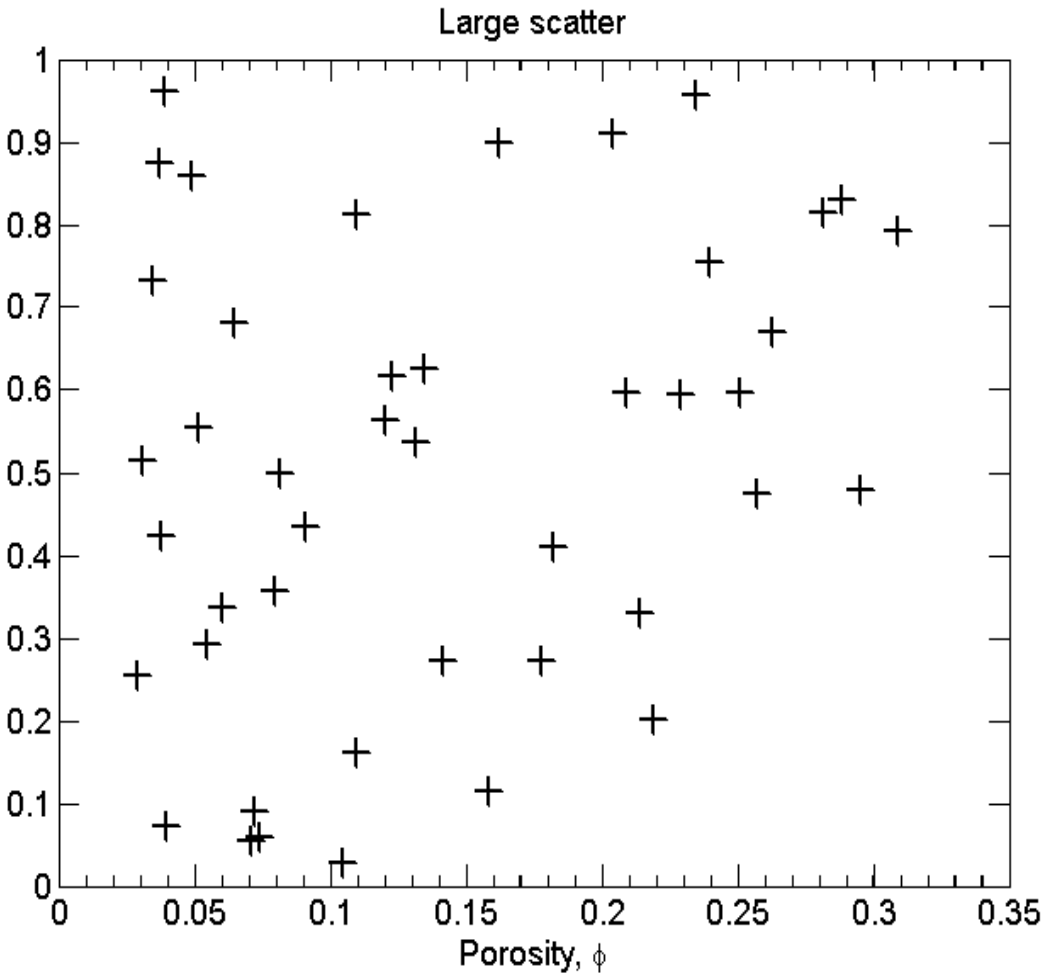


- Source: EuroStat

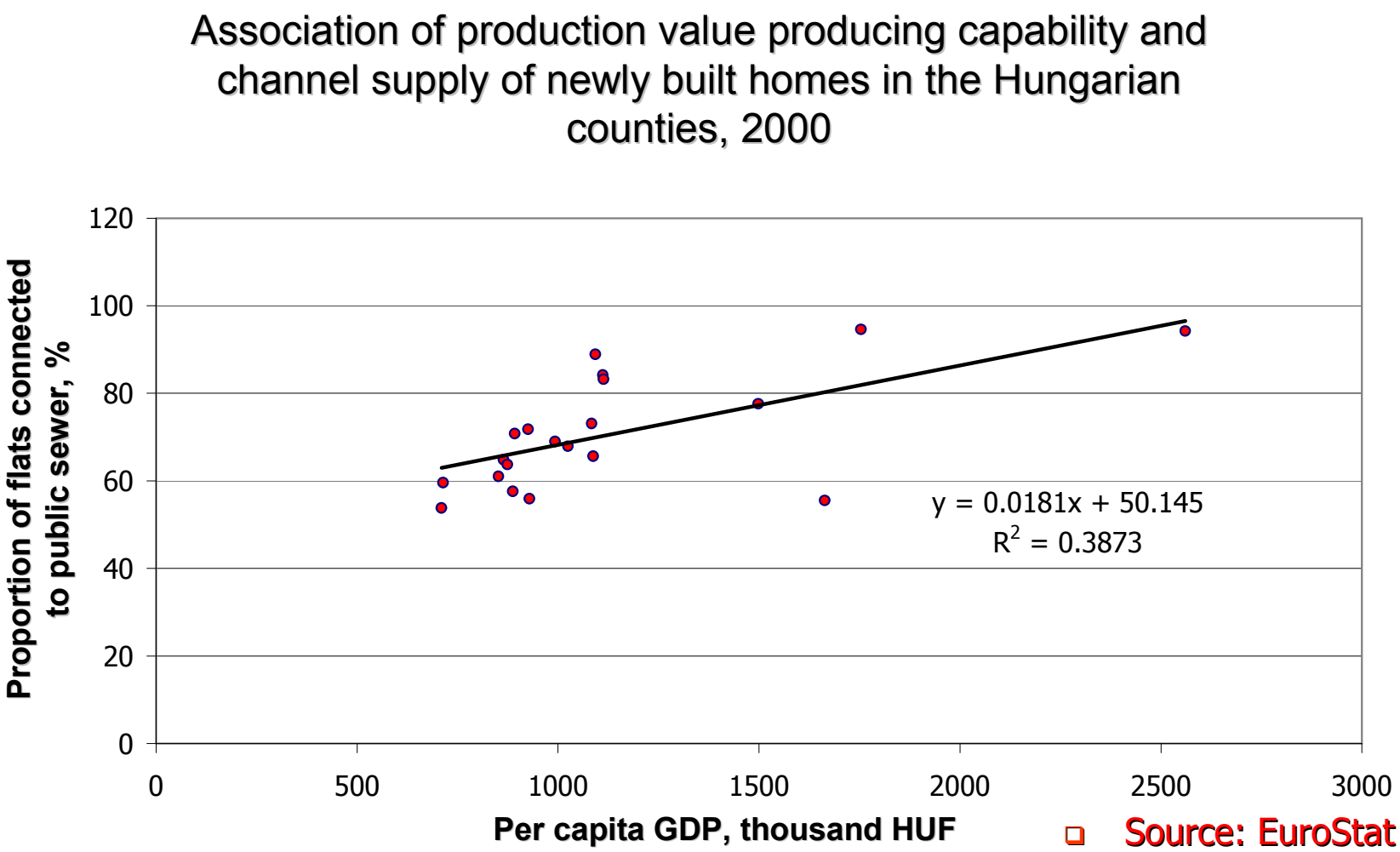
Scatter diagram



Scatter diagram

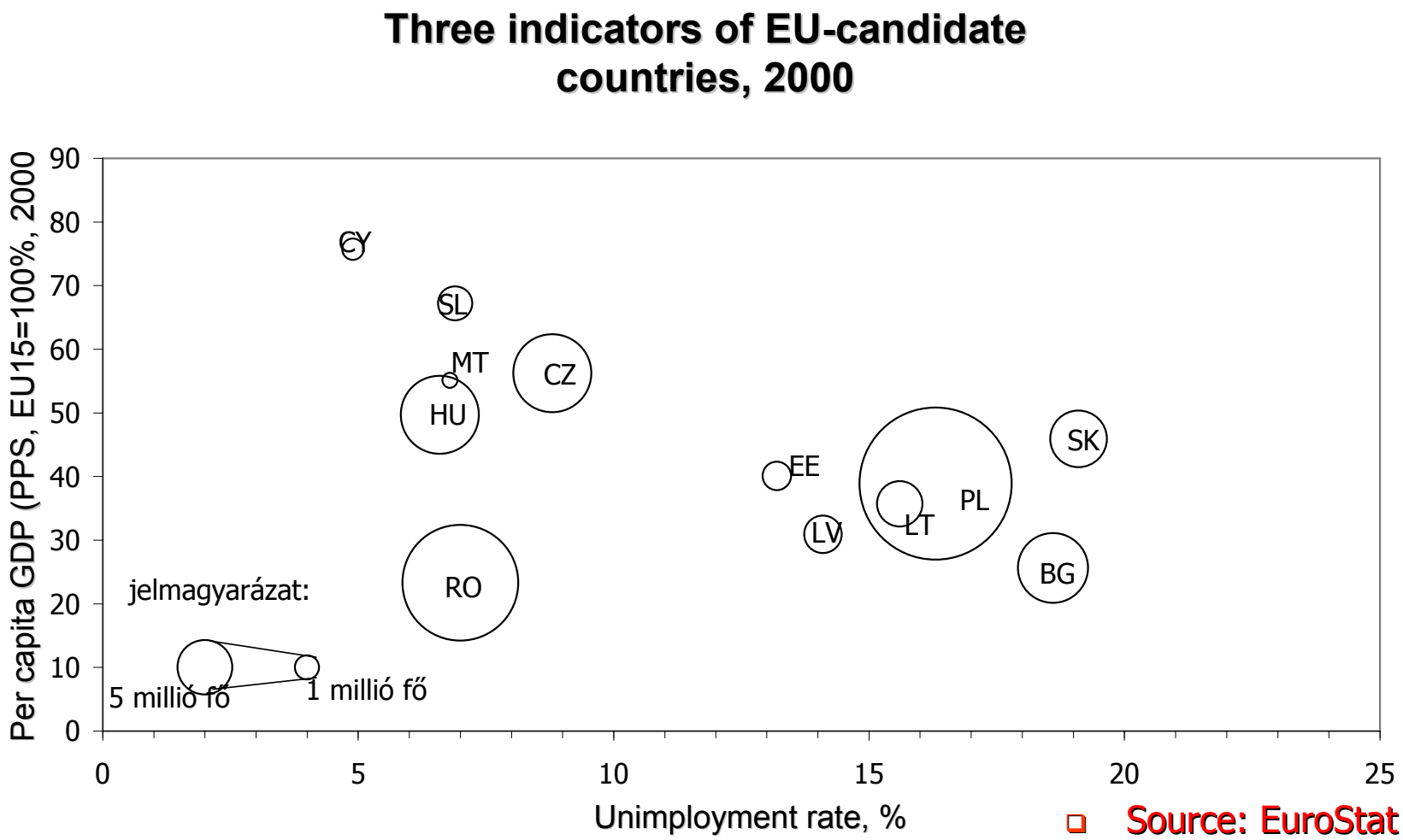


Scatter diagram – its special type is regression diagram



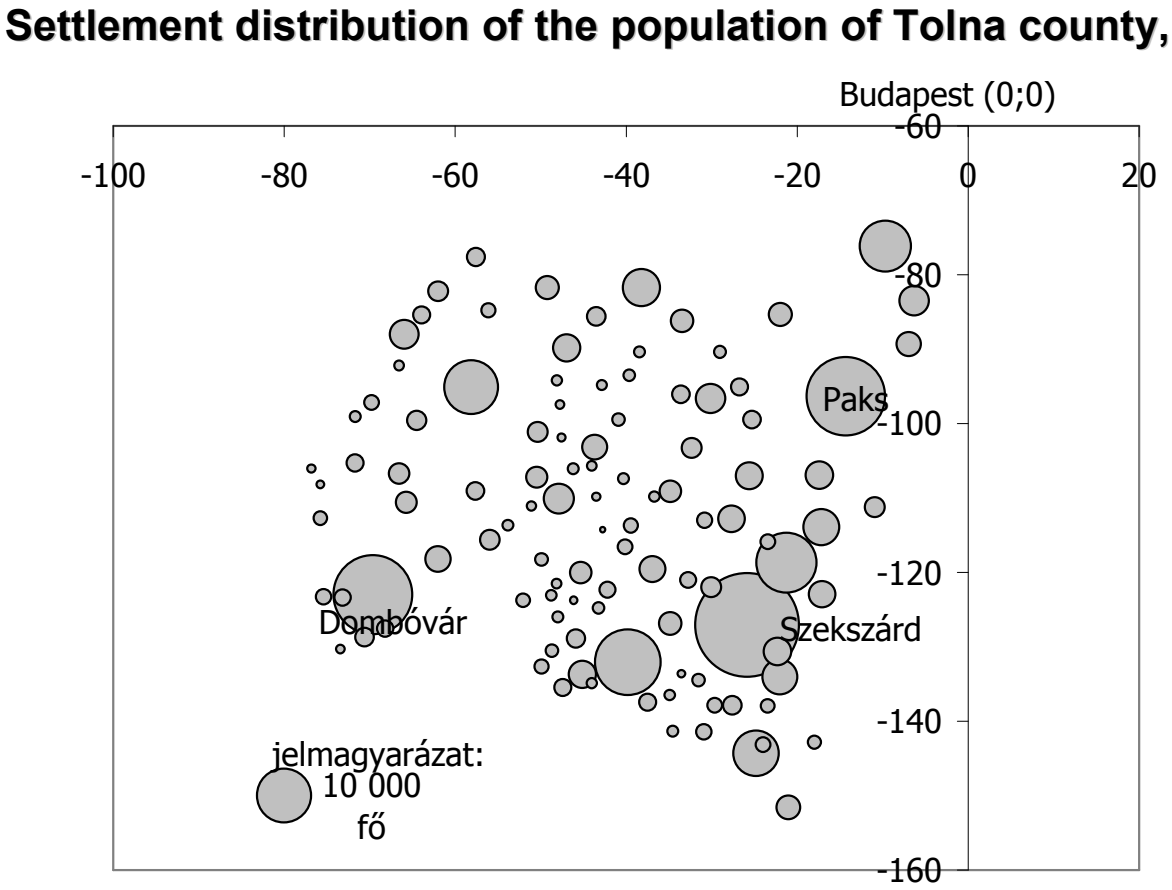
- Source: EuroStat

Bubble diagram: three-dimensional comparison



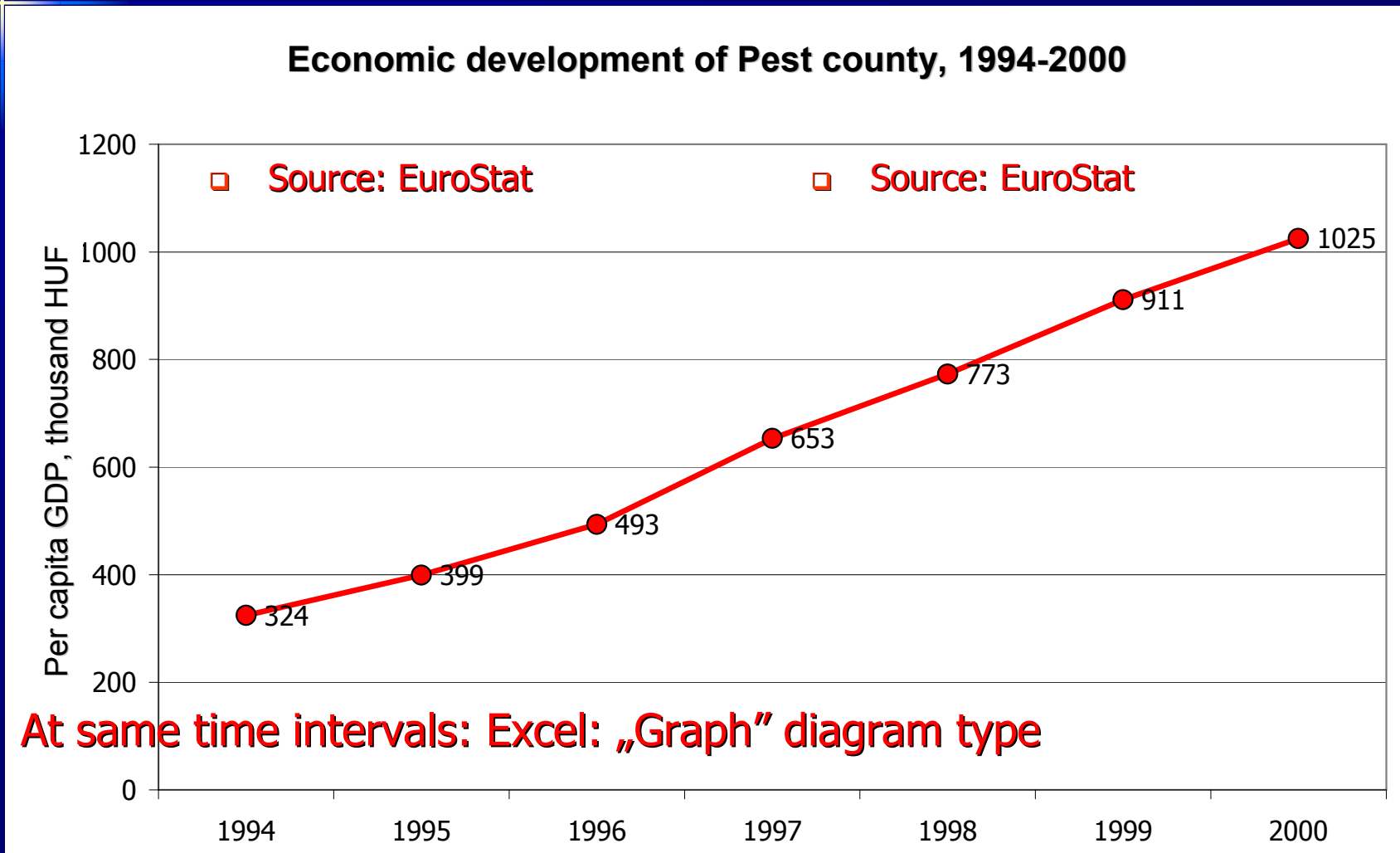
- Source: EuroStat

Bubble diagram: its special case is pictogram map



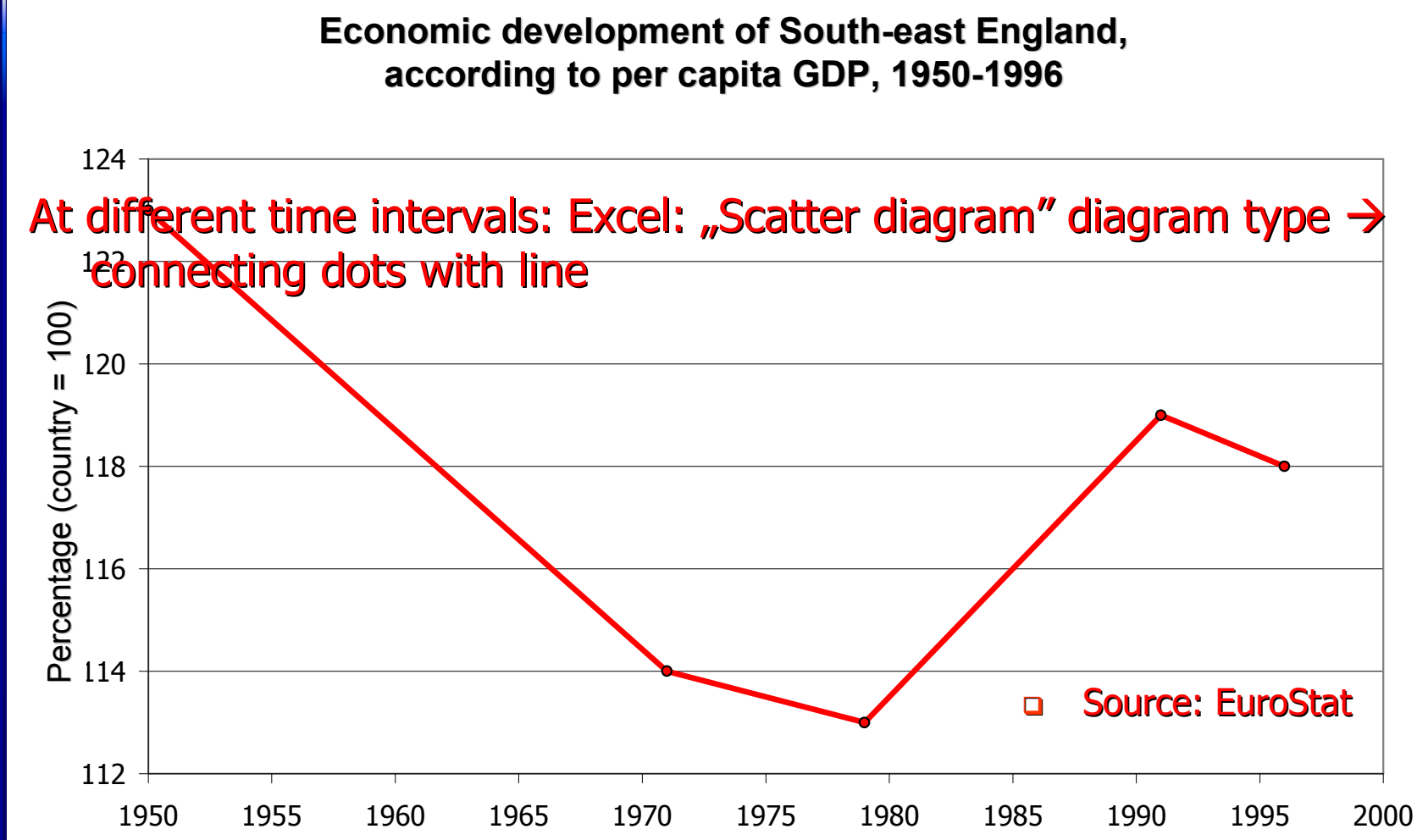
- Source: KSH T-Star

Line chart (graph): two-dimensional comparison (one dimension is time)

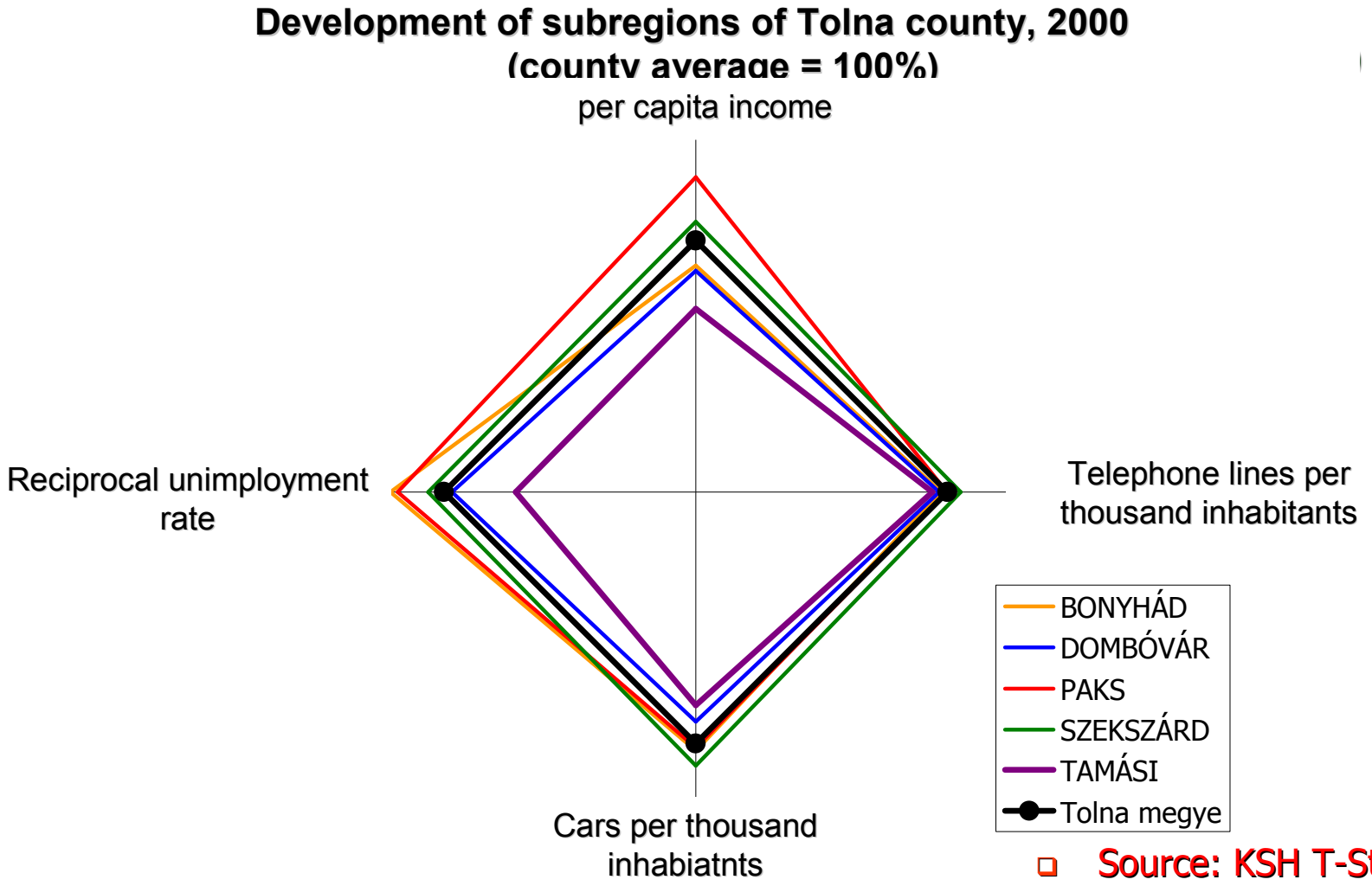


At same time intervals: Excel: „Graph" diagram type

Line chart (graph): two-dimensional comparison (one dimension is time)



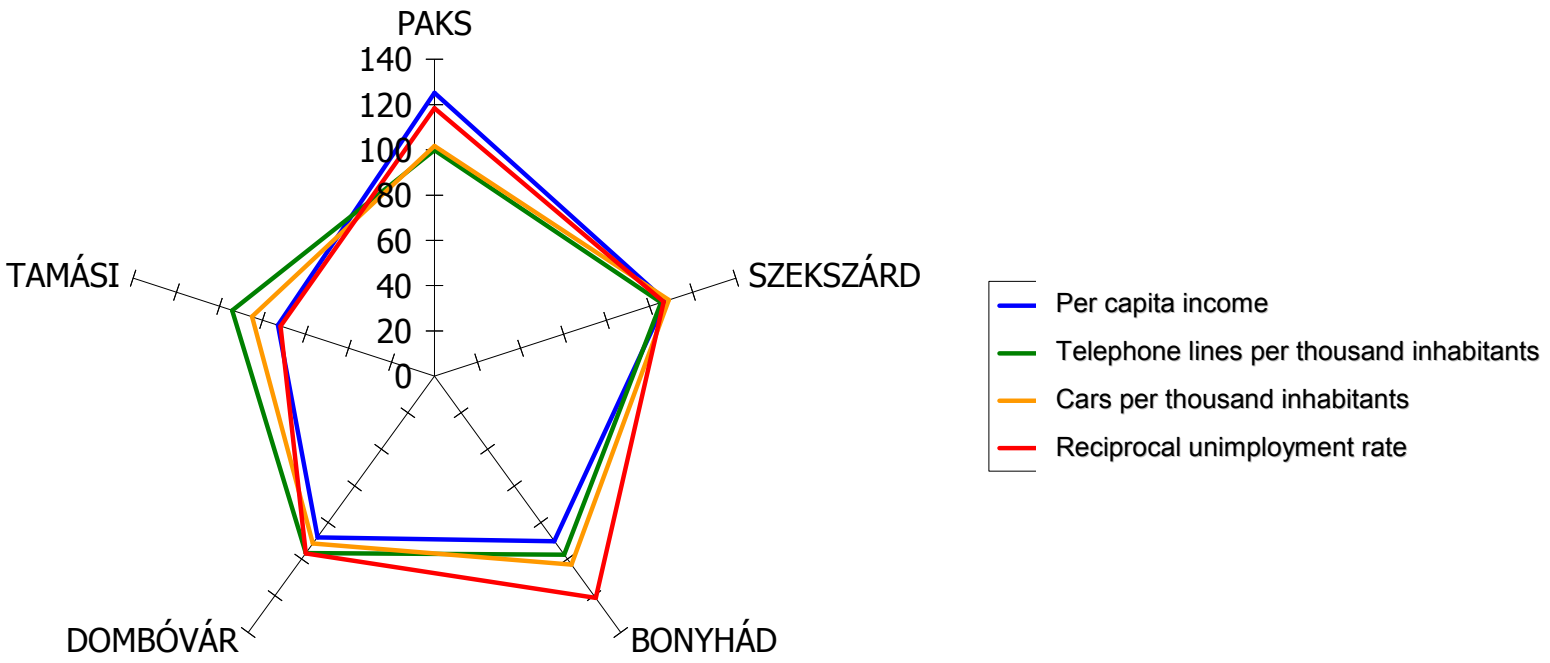
Radar diagram: multi-dimensional comparison



- Source: KSH T-Star

Radar diagram: multi-dimensional comparison

The development of regional disparities in Tolna County, 2000
(county average = 100%)



- Source: KSH T-Star

Always look on the bright side of things!

We finished for today, goodbye!

ямарваа нэг зүйлийн гэгээлэг талыг нь үргэлж олж харцгаая

өнөөдөртөө ингээд дуусгацгаая, баяртай

让我们总是从光明的一面来看待事物吧！

今天的课程到此结束，谢谢！

دعونا ننظر دائما إلى الجانب المشرق من الأشياء!

انتهينا لهذا اليوم، وداعا!