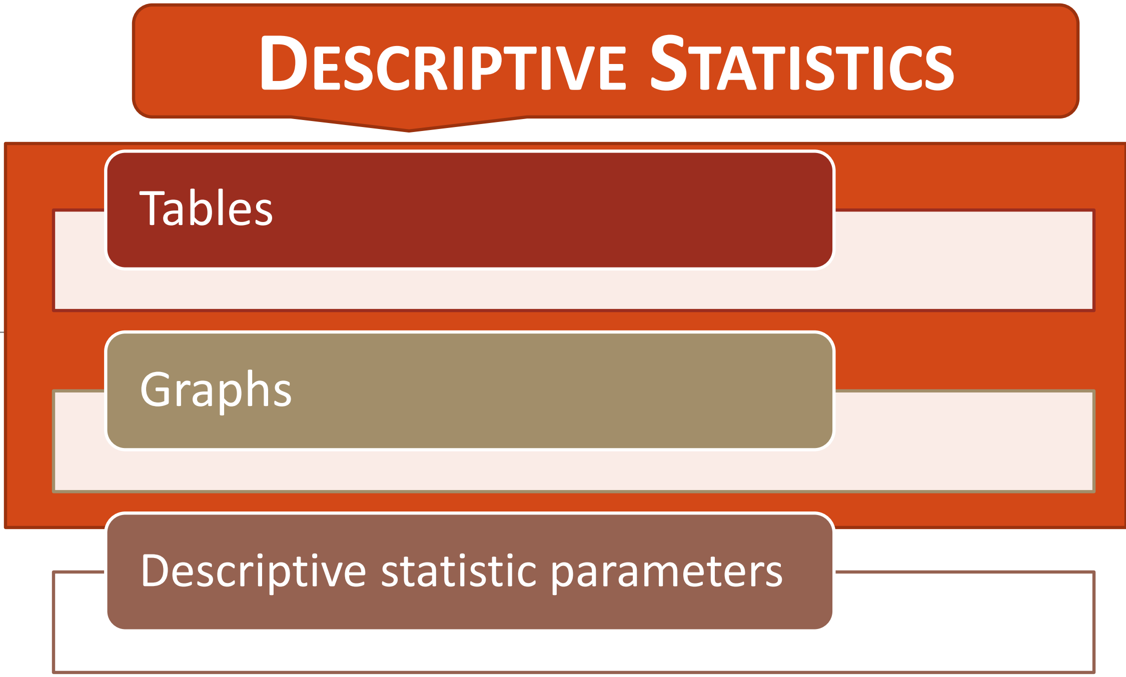
DESCRIPTIVE STATISTICS I

OBJECTIVES



Descriptive Statistics

Tables

Graphs

Descriptive statistic parameters

- Summarizing data with tables
- Graphical representations

- Qualitative or quantitative data

→ table

  - Frequency table (use classes of frequencies if data are quantitative)
  - Contingency tables (applied for qualitative data or quantitative data express as classes of frequencies)

→ Graphical representation

- Quantitative data / <u>Qualitative ordinal data</u>
  - Descriptive statistics parameters

Summarizing medical data

Large amounts of medical data are compressed into more easily assimilated summaries

- Provide the user with a sense of the content

Some ways exist to describe data depending on the type of variables:

- Qualitative variables
- Quantitative variables

Good tables practices

1. Simple: it is preferred to have 2 or 3 small tables instead of one big table
2. Must be information without reading the associated text:
   - Abbreviations and symbols must be explained at the bottom of the table
   - Definitions of rows and columns with units of measurements in headings (if it is applied)
   - Brief descriptive heading: what? when? where?
   - Must not duplicate material in the text or in illustration
   - Synthesis (total) rows and columns
3. If data are taken from another research, the source of data must be referred.

Good graphical practices

Any graphical representation must have:

- Title
- Definitions of axes
- Units of measurements for each ax (if it is applied)
- Legend (if it is applied)

An excellent graphical representation must be as self-explanatory as possible!

Good graphical practices

A graphical representation aims to transmit information

When drawing a graphical representation try to answer the following question: Which is the aim of the graphical representation?

Medical data must be represented graphically in such a way in which to be useful for understanding the clinical phenomena

Notice to:

- The color composition (do not use color background)
- The font size (it is supposed to be readable)

One qualitative variable

One qualitative variable

Raw data

Subject ID	Hypertension class
1	Stage I
2	Normal
3	Prehypertension
4	Stage II
…	…
1000	Stage II

Category	Blood Pressure (mm Hg)
Normal	SBP 90-119 and DBP 60-79
Prehypertension	SBP 120-139 or DBP 80-89
Stage 1 HTN	SBP 140-159 or DBP 90-99
Stage 2 HTN	SBP ≥160 or DBP ≥100

DBP = diastolic blood pressure;
SBP = systolic blood pressure	

FREQUENCY TABLES

Numerical measures …

What information can we extract from these data?

1. What % of subjects fall into each category
2. How are the subjects divided into the hypertension categories?

One qualitative variable: frequency tables

Data are sort ascending

The absolute frequency of each value is

The distinct values and associated frequencies are included in a table :

- Absolute frequency (no): the total amount of occurrences of one variable
- Relative frequency (%) = the absolute frequency divided by the total amount of occurrences

One qualitative variable: frequency tables

Could contain the following types of frequencies:

- Absolute frequency
- Cumulative absolute frequency (ascending (ACAF)/ descending (DCAF))
- Relative frequency
- Cumulative relative frequency (ascending (ACRF)/ descending (DCRF))

Microsoft Excel:

- COUNTIF
- Pivot Table
  - [Data - Pivot Table and Pivot Chart Report ...]

One qualitative variable: frequency tables

Absolute frequency

Relative frequency

Category	No. patients	Percent (%)
Normal	300	= (300/1000)*100 = 30.00
Prehypertension	100	= (100/1000)*100 = 10.00
Stage I	350	= (350/1000)*100 = 35.00
Stage II	250	= (250/1000)*100 = 25.00
Total	n = 1000	100%

Sample size

If you add the percentages, you must have a total of 100%. If the value is higher, then you rounded incorrectly the percentages

One qualitative variable: frequency tables

The sum of relative frequencies of all values in the series that are less than or equal to *x/n*

The sum of absolute frequencies of all values in the series that are less than or equal to *x*

Diagnosis	**No.**	**%**	**No. cumulat ↑**	**% cumulat ↑**
Normal	300	30.00	=300	= 30.00
Prehypertension	100	10.00	=300+100=400	= 30.00+10.00=40.00
Stage I	350	35.00	=400+350=750	=40.00+35.00=75.00
Stage II	250	25.00	=750+250=1000	=75.00+25.00=100
Total	1000	100		

One qualitative variable: frequency tables

Let's have the following incubation time expressed in days for infectious diseases: 5, 6, 7, 7, 8, 8, 5, 7, 8, and 7. Which of the following values correspond to the ascending cumulative relative frequency of 0.7?

A. 8

B. 6

C. 5

D. 7

E. None

One qualitative variable: frequency tables

Let’s have the following incubation time expressed in days for infectious diseases: 5, 6, 7, 7, 8, 8, 5, 7, 8, and 7.

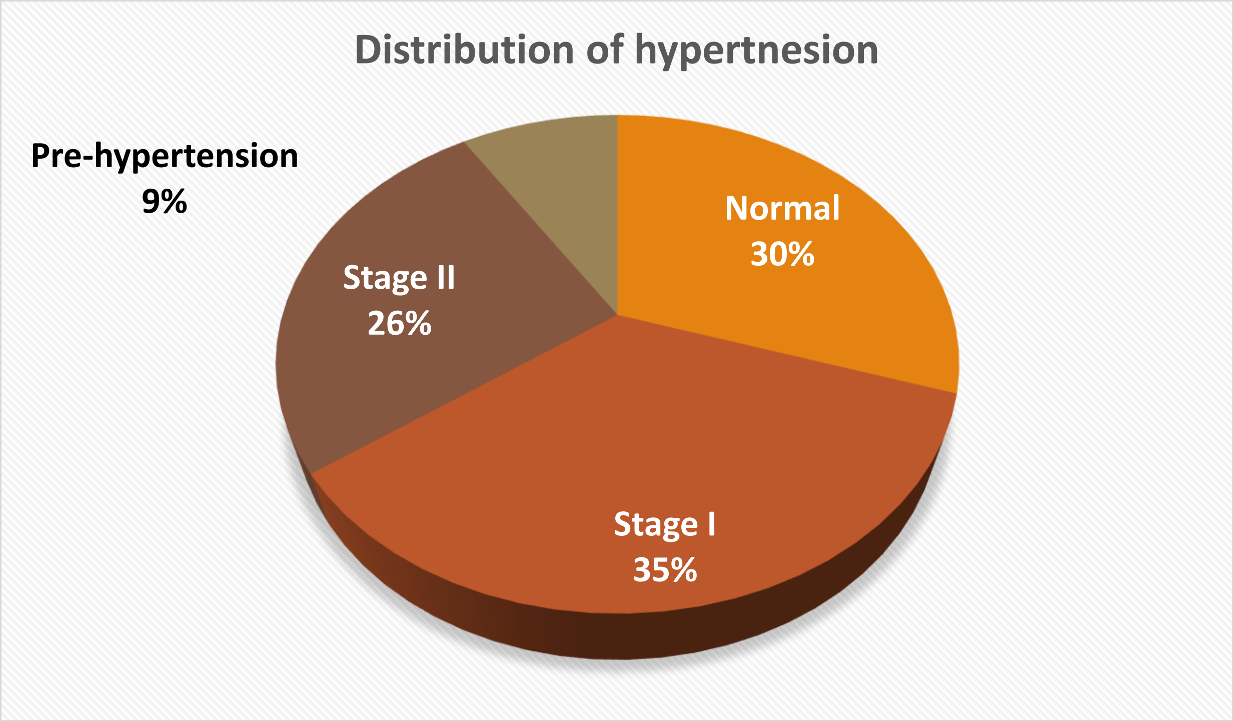
Which of the following values correspond to the ascending cumulative relative frequency of 0.7?

Value	f_a	f_r	f_a ac	f_r rc
5	2	0.20	2	0.20
6	1	0.10	3	0.30
7	4	0.40	7	0.70
8	3	0.30	10	1
Total	10	1		

f_a –absolute freq
f_r –relative freq
f_a ac – ascending cumulative absolute freq (ACAF)
f_r ac –ascending cumulative relative freq (ACRF)

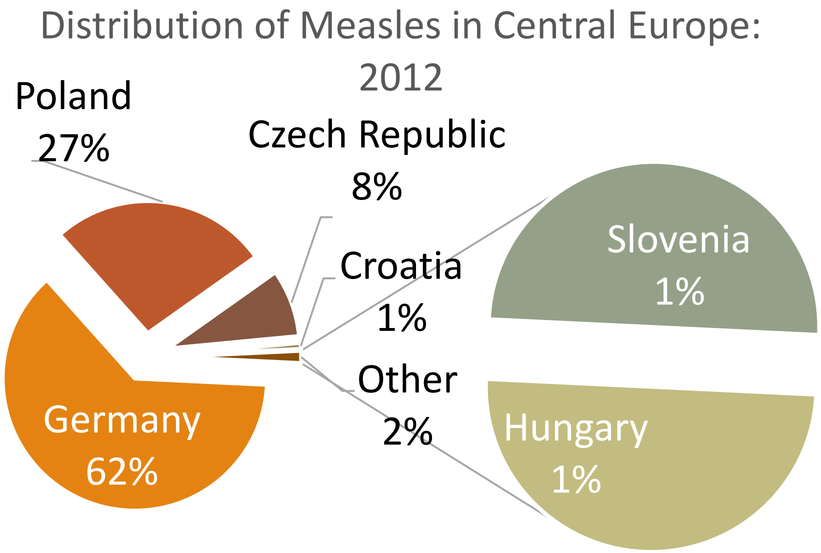
Visual or graphical displays

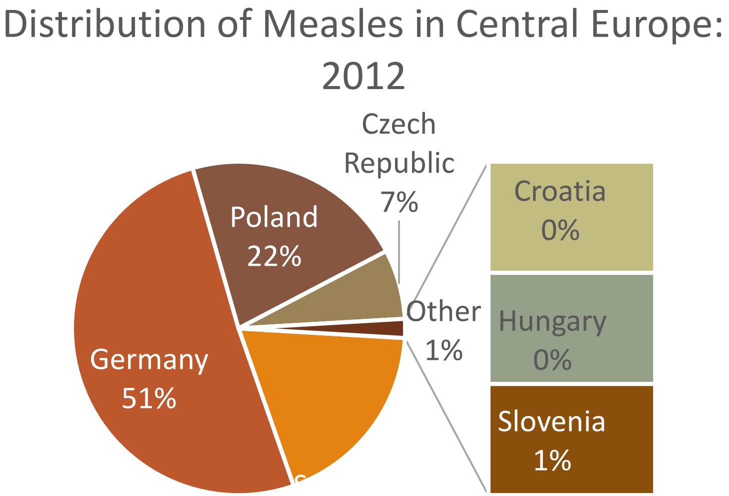
Pie chart: a circular chart used to compare parts of the whole and to look to frequencies of a qualitative variable



Pie of pie / pie of bar

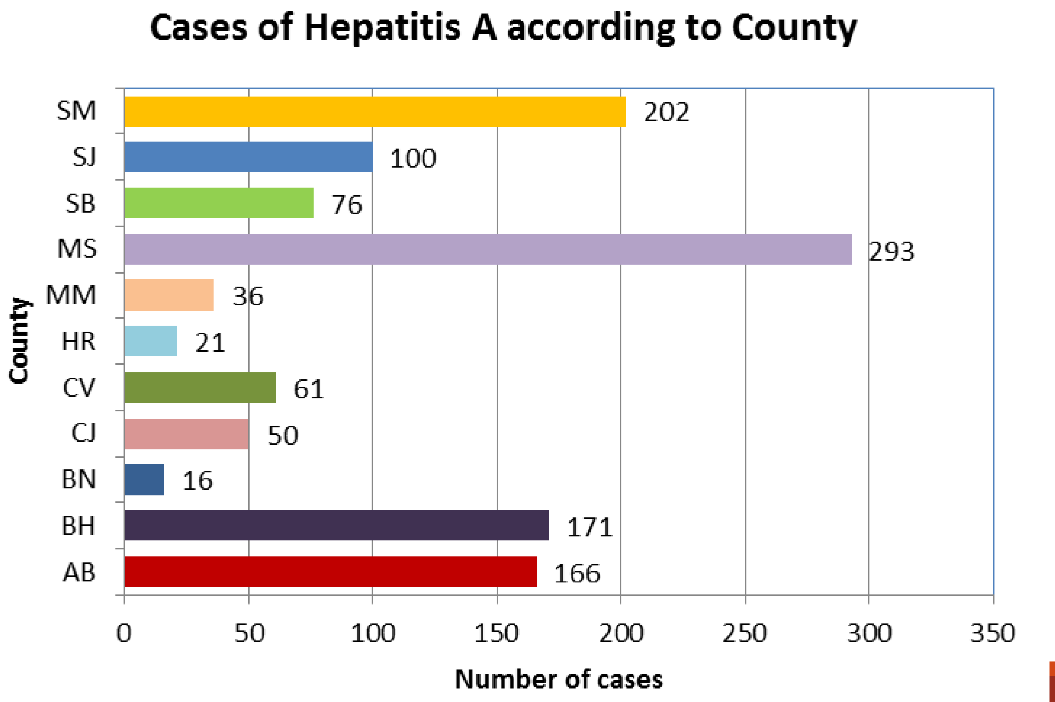
Country	No of cases of measles: 2012
Switzerland	61
Germany	166
Poland	71
Czech Republic	22
Croatia	2
Hungary	2
Slovenia	2





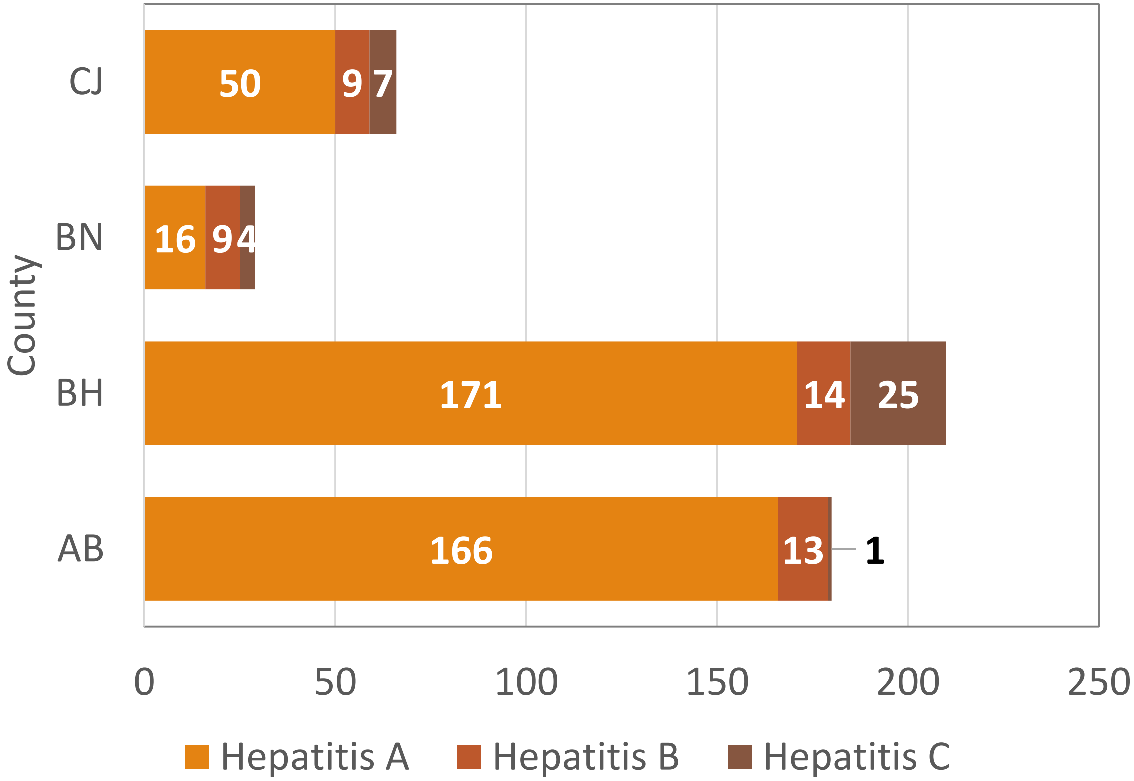
Bar graph

A **bar graph** is composed of discrete bars that represent different categories of data. The length or height of the bar is equal to the quantity within that category of data. Bar graphs are best used to compare values across groups.



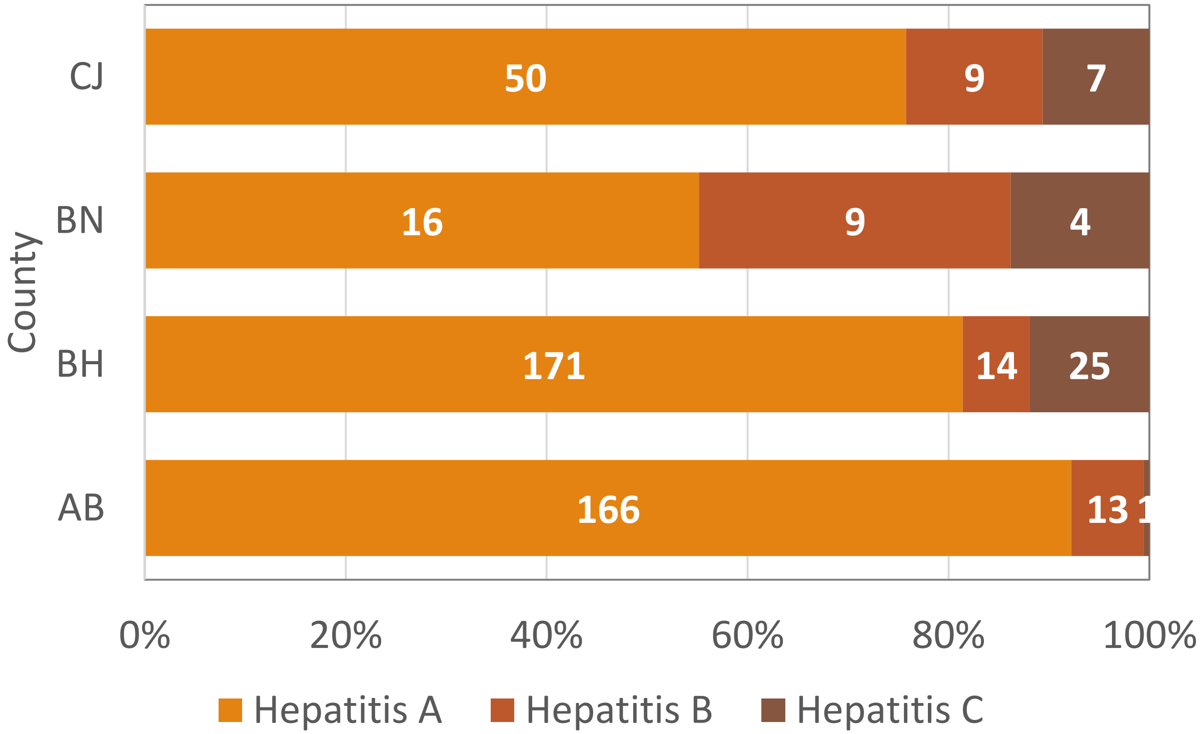
Stacked bar

	AB	BH	BN	CJ
Hepatitis A	166	171	16	50
Hepatitis B	13	14	9	9
Hepatitis C	1	25	4	7



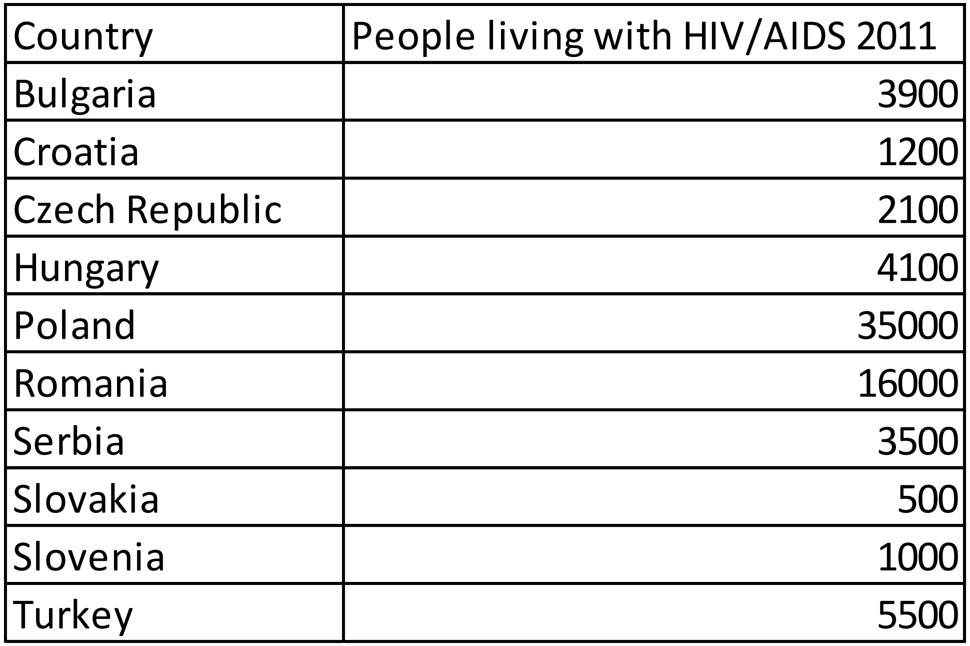
100% stacked bar

	AB	BH	BN	CJ
Hepatitis A	166	171	16	50
Hepatitis B	13	14	9	9
Hepatitis C	1	25	4	7



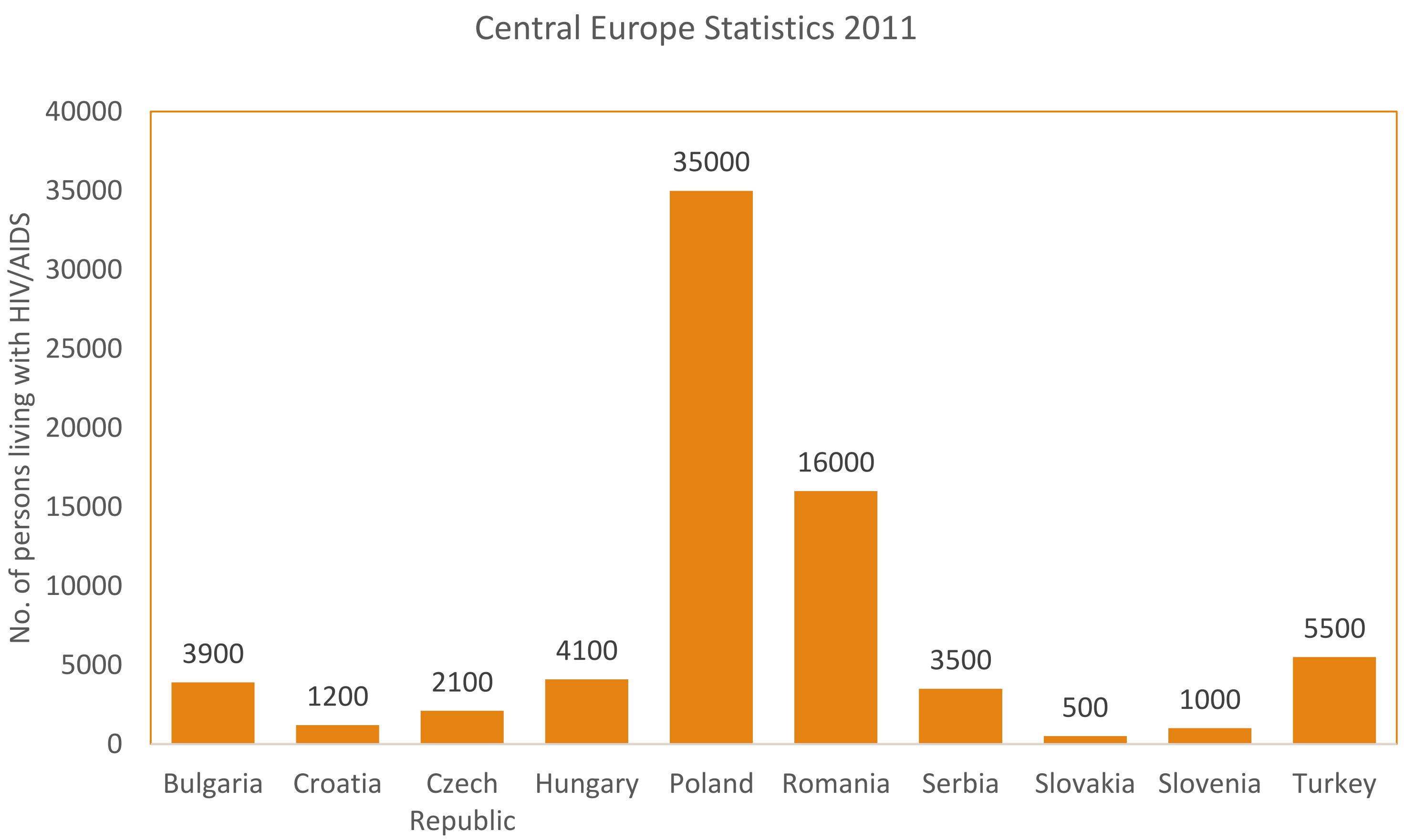
Column graphs

A **column** graph is composed of discrete columns that represent different categories of data. The length or height of the column is equal to the quantity within that category of data. Similar to the Bar graphs, Columns graphs are used to compare values across groups.



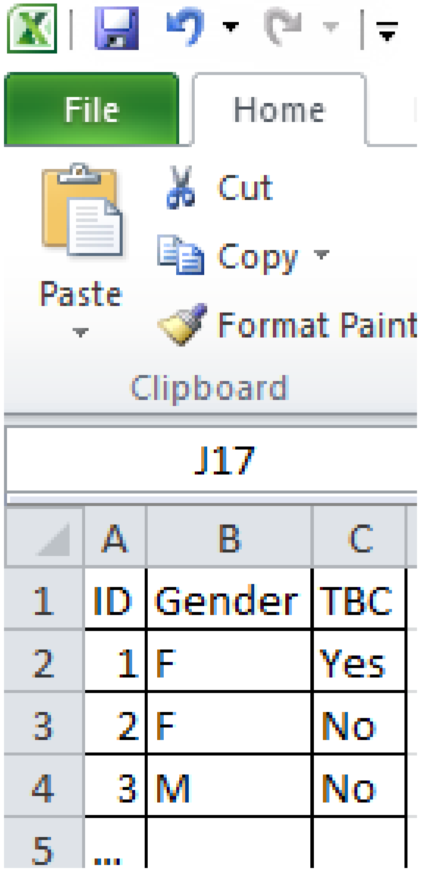
Country	People living with HIV/AIDS 2011
Bulgaria	3900
Croatia	1200
Czech Republic	2100
Hungary	4100
Poland	35000
Romania	16000
Serbia	3500
Slovakia	500
Slovenia	1000
Turkey	5500

Column graph



MORE THAN ONE QUALITATIVE VARIABLE

Two qualitative variables:
2 by 2 contingency table



	A	B	C
1	ID	Gender	TBC
2	1	F	Yes
3	2	F	No
4	3	M	No
5	...		

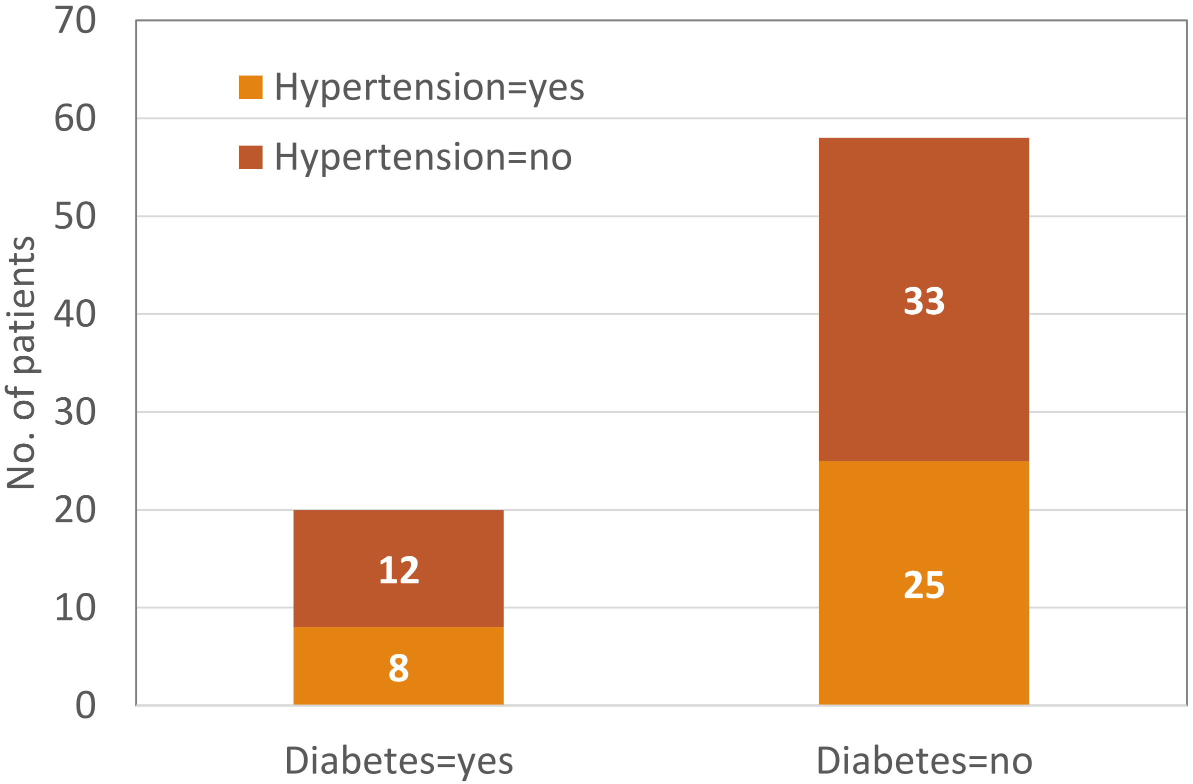
Two dichotomial variables:

- Variable 1 = Gender
- Variable 2 = Tuberculosis

	TBC=yes	TBC=no	**Total**
Gender=F	2	10	**12**
Gender=M	24	54	**78**
Total	**26**	**64**	**90**

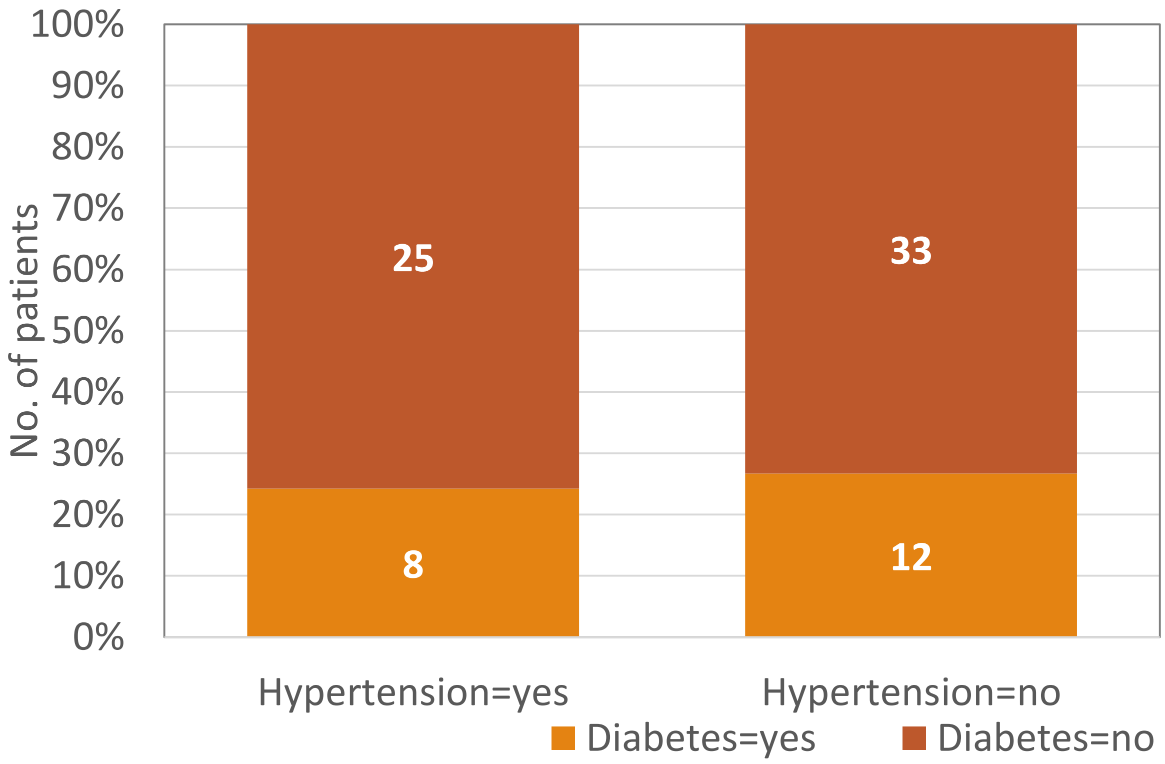
Stacked column

	Diabetes=yes	Diabetes=no
Hypertension = yes	8	25
Hypertension = no	12	33



100% stacked column

	Diabetes=yes	Diabetes=no
Hypertension = yes	8	25
Hypertension = no	12	33



x qualitative variables: frequency tables

Table 1. Distribution of pulmonary pathologies associated with silicosis

Grade of silicosis	BrC	BPOC	Emphysema	CPC	TBC	**Total**
I	12	20	0	0	14	**46**
I/II	1	5	1	1	1	**9**
II	3	7	1	1	7	**19**
II/III	0	1	0	0	0	**1**
III	0	3	0	0	4	**7**
Total	**16**	**36**	**2**	**2**	**26**	**82**

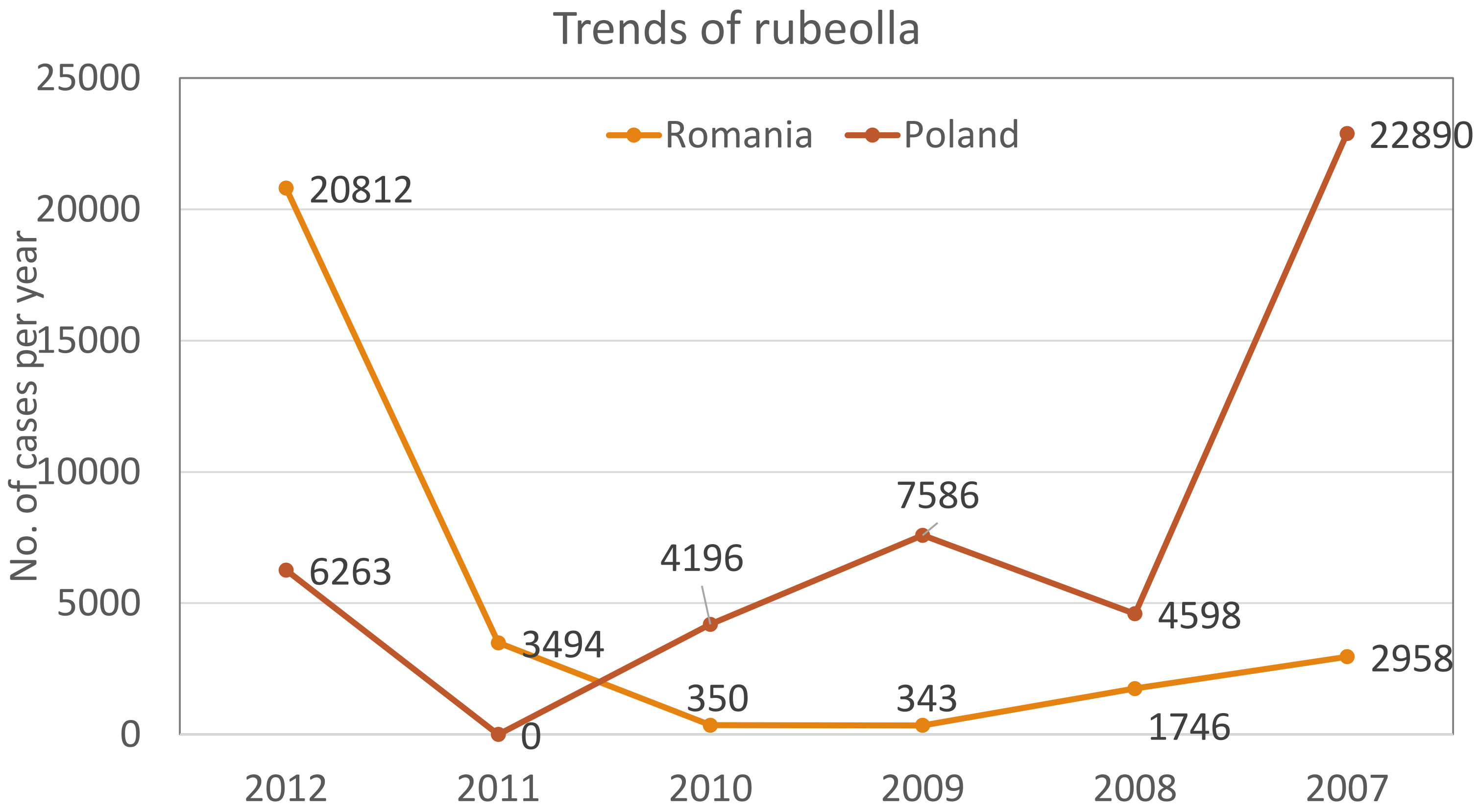
BrC = chronic bronchitis; BPOC = broncho-pneumonitis chronic obstructive;
CPC = chronic pulmonary heart; TBC = pulmonary tuberculosis

Line graph

- The primary use: to convey information similar to a bar chart but for intervals that form a sequence of time or order of events from left to right.
- Relationship of a Line Chart to a Probability Distribution: as the sample size increases and the width of the intervals decreases, the line chart of a sample distribution approaches the picture of its probability distribution.
- Line graphs provide information on the relation between two variables and are used to illustrate trends over time.

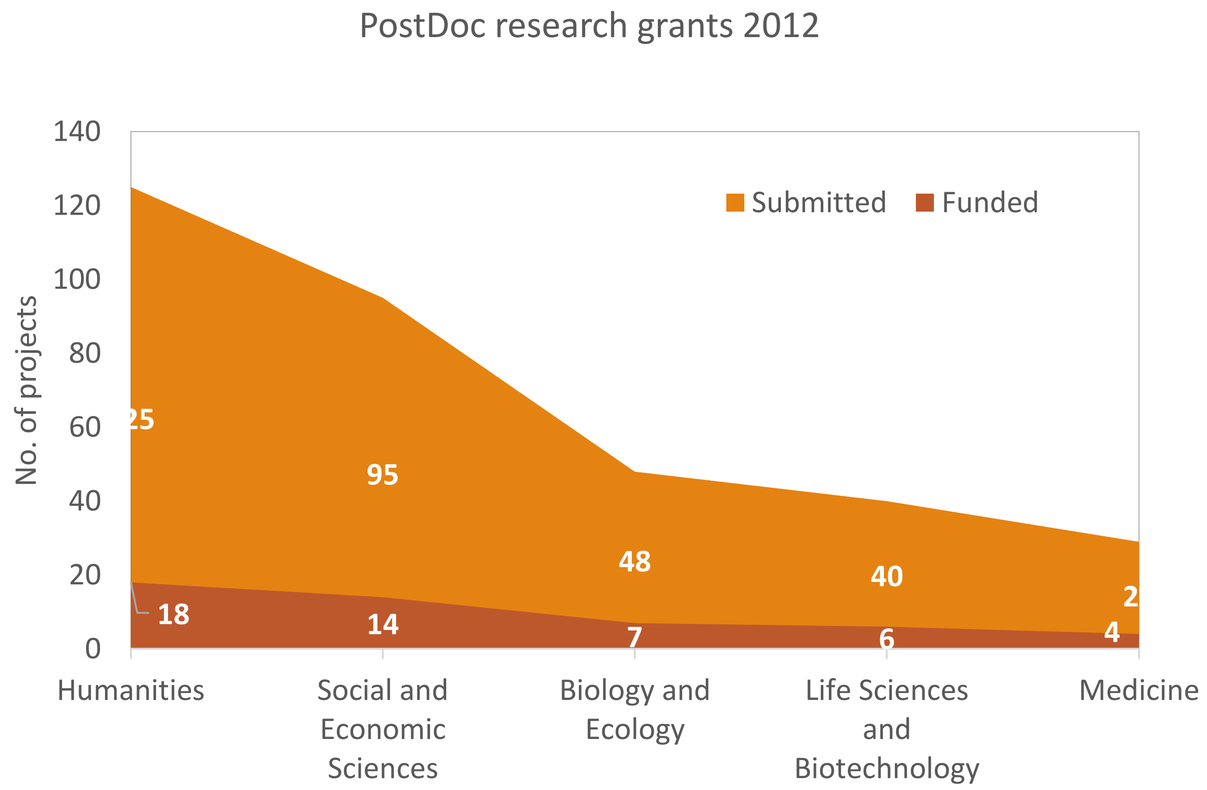
	2012	2011	2010	2009	2008	2007
Romania	20812	3494	350	343	1746	2958
Poland	6263	0	4196	7586	4598	22890

Line graph



Area graph

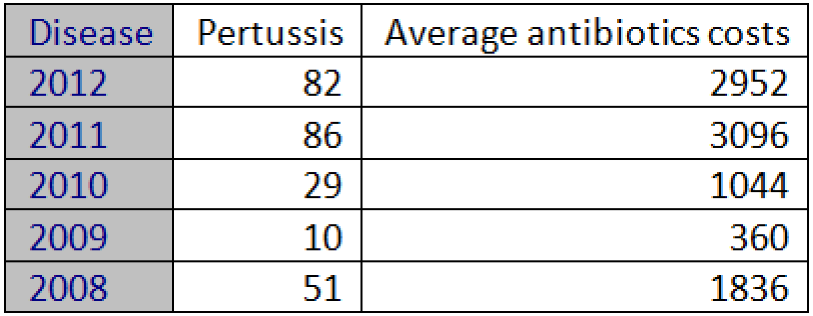
Research domain	Submitted	Funded
Humanities	125	18
Social and Economic Sciences	95	14
Biology and Ecology	48	7
Life Sciences and Biotechnology	40	6
Medicine	29	4



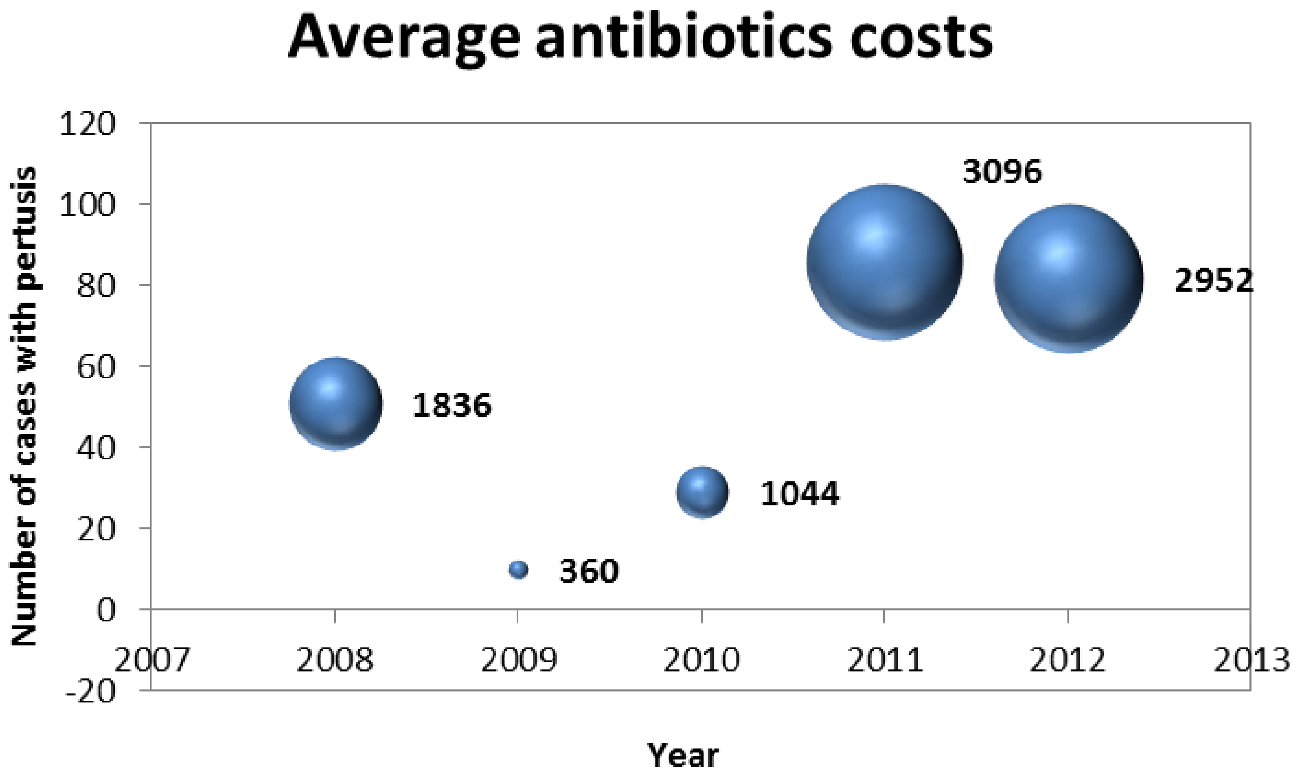
Bubble graph

A bubble chart is a type of graph that displays three dimensions of data. Each entity with its triplet (v_1, v_2, v_3) of associated data is plotted as a disk that expresses two of the v_i values through the disk's xy location and the third through its size. The size of the bubble (data marker) indicates the value of the third selected variable.

Bubble graph



Disease	Pertussis	Average antibiotics costs
2012	82	2952
2011	86	3096
2010	29	1044
2009	10	360
2008	51	1836



Remember!

Qualitative variables are summarized using:

- Visual display: pie, bar, or column charts
- Numerical measures: frequency table (counts and %)

Pie, Bar or Column charts can be used to visualize the distribution of qualitative variables.

Remember!

Pie Chart (qualitative variable on frequency tables):

- Represents proportions rather than amounts.
- Its primary use is to visualize the relative prevalence of the phenomena.
- Has the advantage of avoiding the illustration of sequence that sometimes is implied by the bars charts.
- Pie chart emphasizes how the different groups relate to the whole.

The bar/column chart emphasizes how the different categories compare with each other.

ONE QUANTITATIVE VARIABLE

One quantitative variable

Histogram:

- The choice of the intervals is essential (an unfortunate choice of intervals can change the apparent pattern of the distribution).
  - Enough intervals should be used so that the pattern will be minimally altering the beginning and ending positions.
- The choice of number, width, and starting points of intervals arise from the user's judgment (they should be considered carefully before forming the chart).

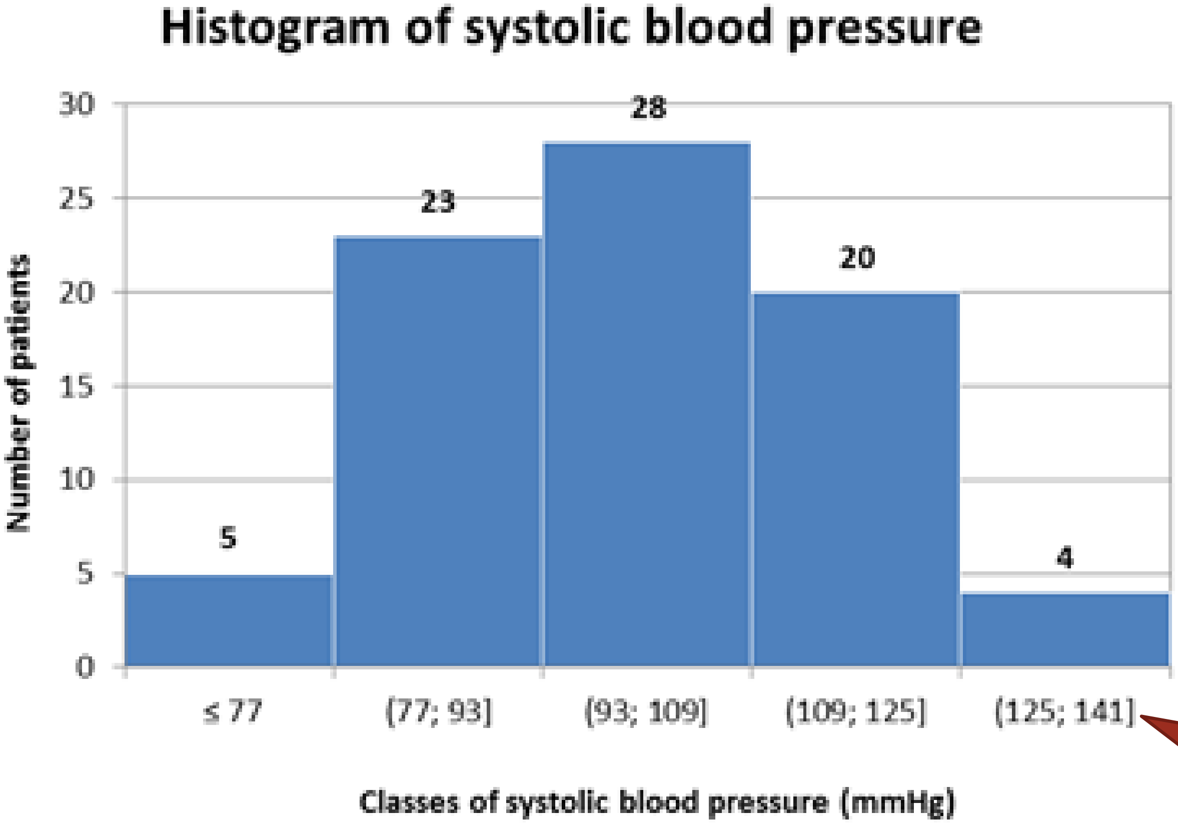
One quantitative variable

Histogram :

- Appears like the bar chart but differs in that the number of observations lying in an interval is represented by the area of a rectangular (or bar) rather than its height.
- If all intervals are of equal width, the histogram is no different from the bar chart except cosmetically (no blank space between bars).
- It is crucial that each observation is counted only in one interval.

Histogram

Classes of frequency	*Frequency*
≤ 77	5
(77; 93]	23
(93; 109]	28
(109; 125]	20
(125; 141]	4

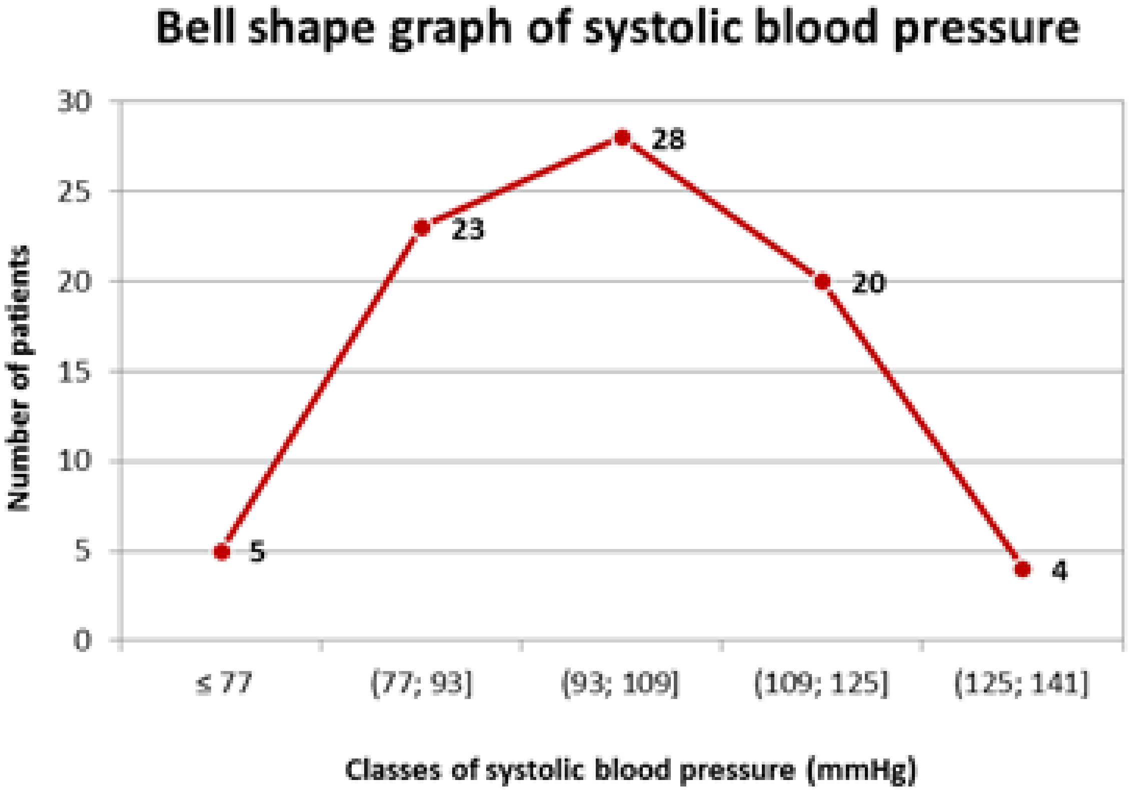


(125; 141] → all subjects with SBP higher than 125 mmHg and smaller or equal with 141 mmHg

'(' = the value is not included in the range
']' = the value is included in the interval

Histogram

Classes of frequency	*Frequency*
≤ 77	5
(77; 93]	23
(93; 109]	28
(109; 125]	20
(125; 141]	4



Histogram

When data are displayed in a histogram, some information is lost. Using histogram we

- can answer: “How many subjects has SBP > 125 mmHg?” (4)
- cannot answer: “What was the lowest value of SBP?”
  - The lowest value of SBP is less than 77 mmHg

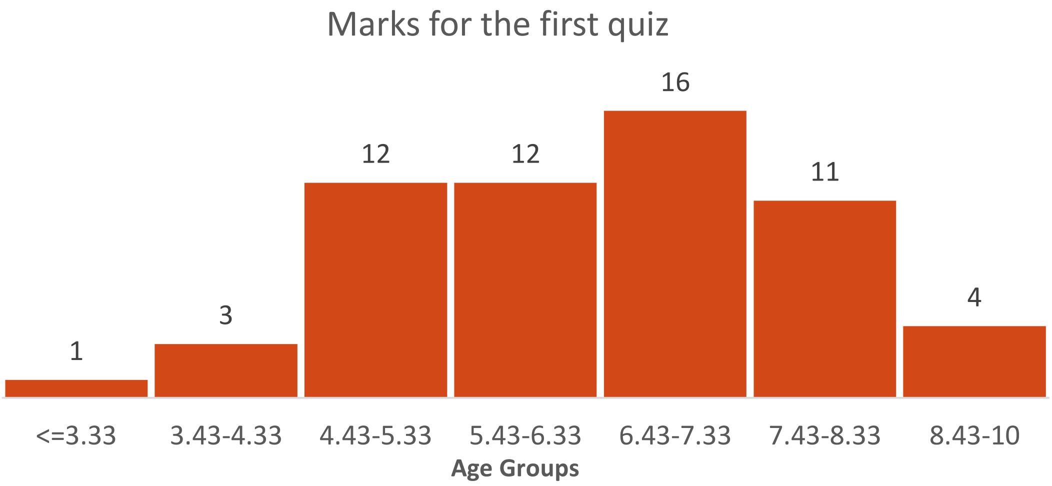
Different width of the intervals provides a different graphical representation of a variable.

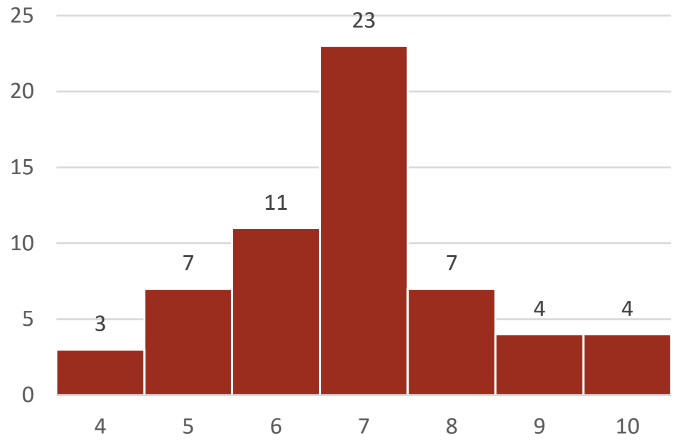
Histogram

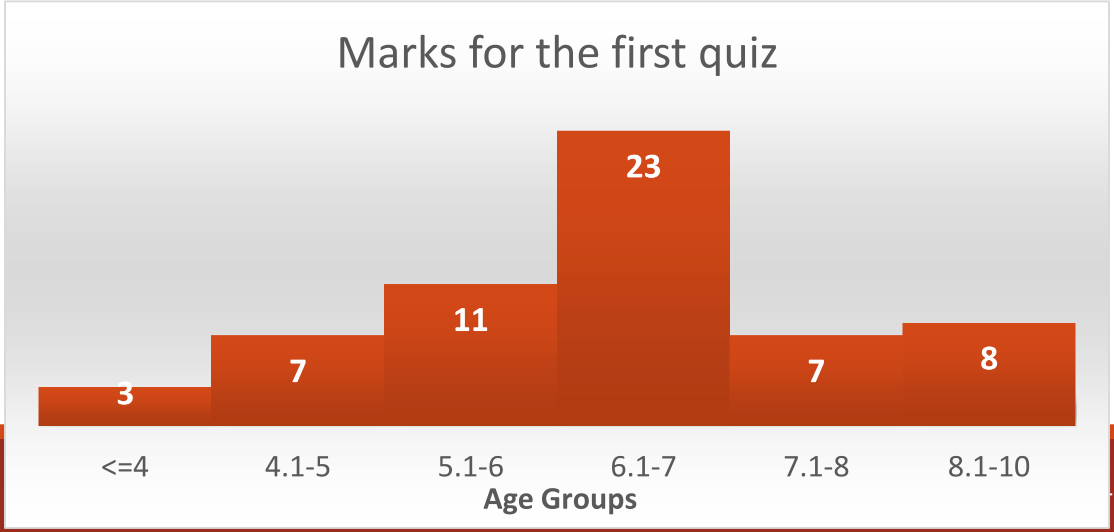
How do I know what interval width to choose?

- Different approaches are used to determine the width of the interval, and different statistical software uses different rules of thumbs to find the optimal value.
- However, in this course we will rely on the software, or we will create the histograms for given intervals.

Histogram & the width of the interval by example







TWO QUANTITATIVE VARIABLE: SCATTE

Scatter

Each point represents an individual

The explanatory (independent) variable on the horizontal X-axis, and the response variable (dependent) on the vertical Y axis.

It describes the overall pattern (direction & form & strength) of the relationship and any deviations from that pattern (see regression analysis).

It is important to have labels of the axis and the unit of measurement for each variable

Scatter

Explanatory (independent) variable

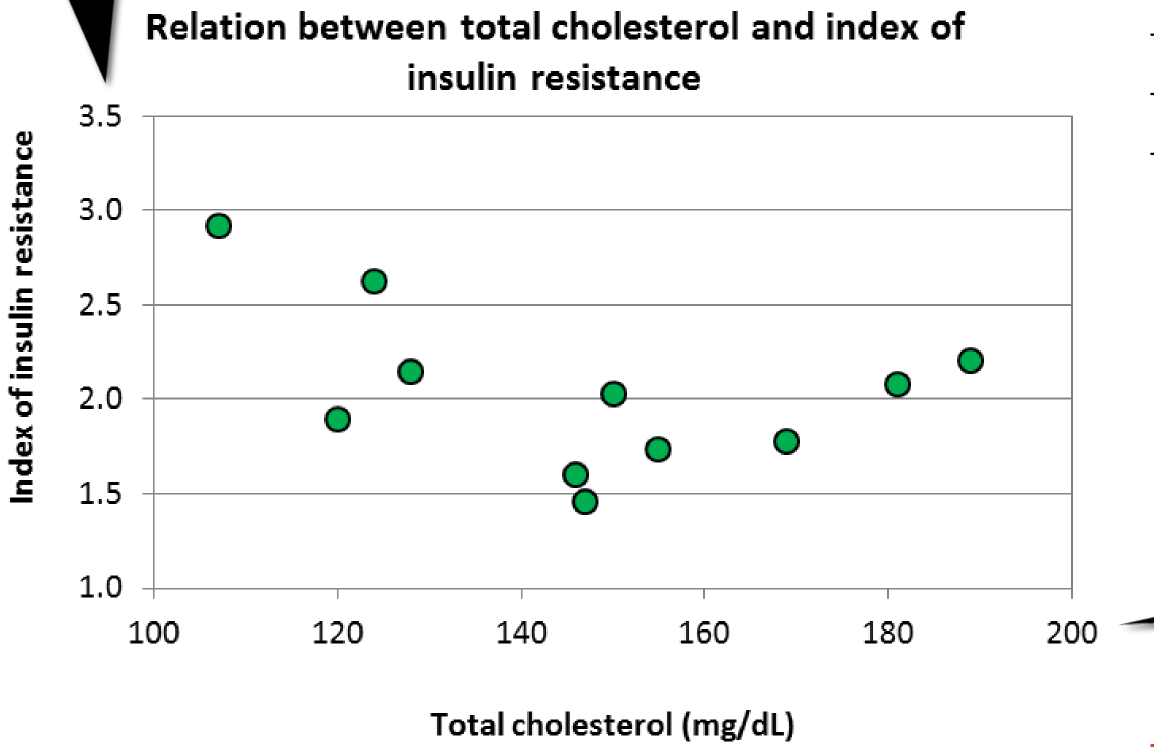
Response (dependent) variable

Total Cholesterol (mg/dL)	Index of insulin resistance
181	2.08
146	1.60
155	1.73
107	2.92
128	2.14
120	1.90
150	2.03
169	1.77
147	1.46
189	2.21
124	2.62

Scatter

Total Cholesterol (mg/dL)	Index of insulin resistance
181	2.08
146	1.60
155	1.73
107	2.92
128	2.14
120	1.90
150	2.03
169	1.77
147	1.46
189	2.21
124	2.62

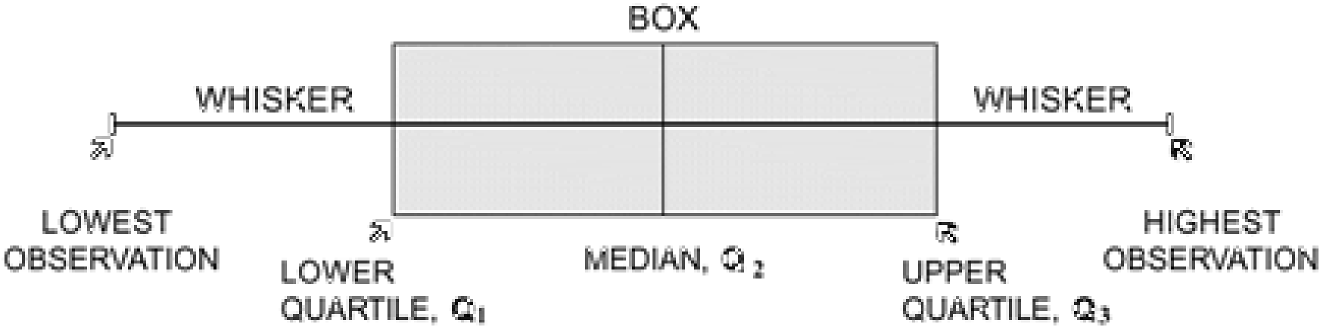
Y axis

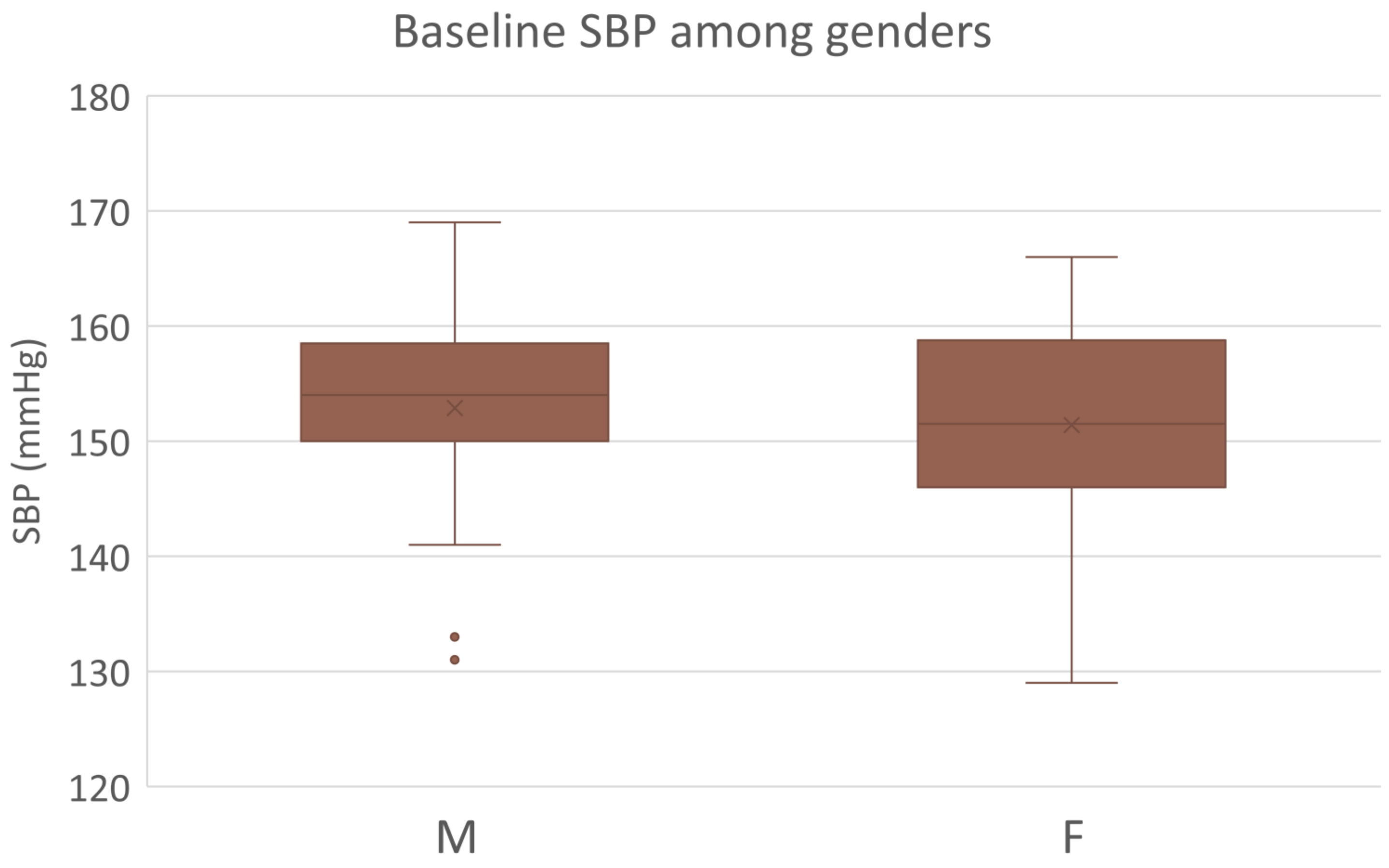


X axis

Box and wishers

A box and whisker plot, also called a box plot, displays the five-number summary of a set of data.




Baseline SBP among genders
SBP (mmHg)
180
170
160
150
140
130
120
M
F

Good tables practices: summary!

Tables:

- Capture: information concisely and display it efficiently.
- Provide information at any desired level of detail and precision.
- Number tables consecutively in the order of their first citation in the text and supply a brief title for each.
- Give each column a short or an abbreviated heading. Authors should place the explanatory matter in footnotes, not in the heading.
- Explain all nonstandard abbreviations in footnotes.
- Identify statistical measures of variations.
- If you use data from another published or unpublished source, obtain permission and acknowledge that source fully.

Good graphic practices: summary!

Figures should be made as self-explanatory as possible.

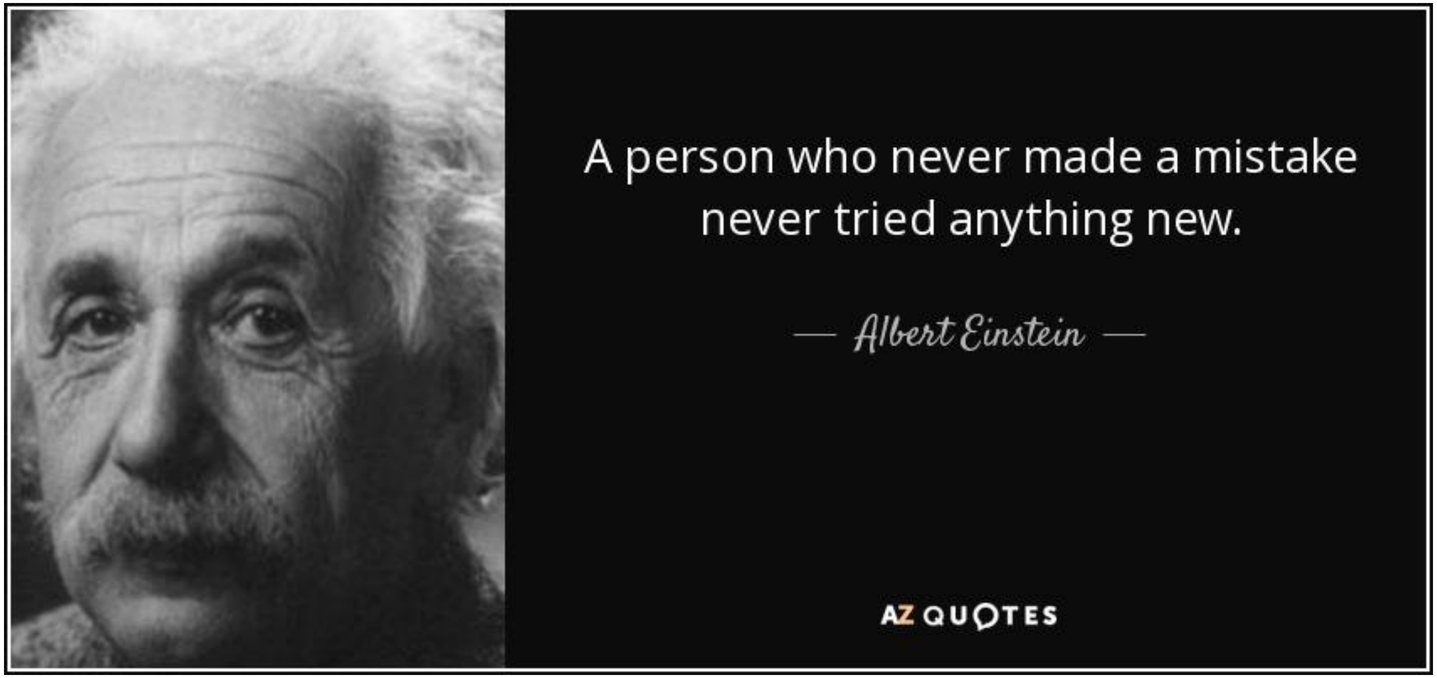
Titles and detailed explanations belong in the legends-not on the illustrations themselves.

Figures should be numbered consecutively according to the order in which they have been cited in the text.

If a figure has been published previously, acknowledge the original source and obtain written permission from the copyright holder to reproduce the figure.

Explain clearly in the legend each symbol, arrow, number, or letter used in the figure.

Avoid 3D graphical representations!



A person who never made a mistake never tried anything new.

— Albert Einstein —