VŠB - Technical University of Ostrava
Faculty of Electrical Engineering and Computer Science
Department of Applied Mathematics

PROBABILITY AND STATISTCS

prof. Ing. Radim Briš, CSc.
www.am.vsb.cz/bris/

Teaching

Prob & Stat.pdf

EXPLORATORY DATA ANALYSIS

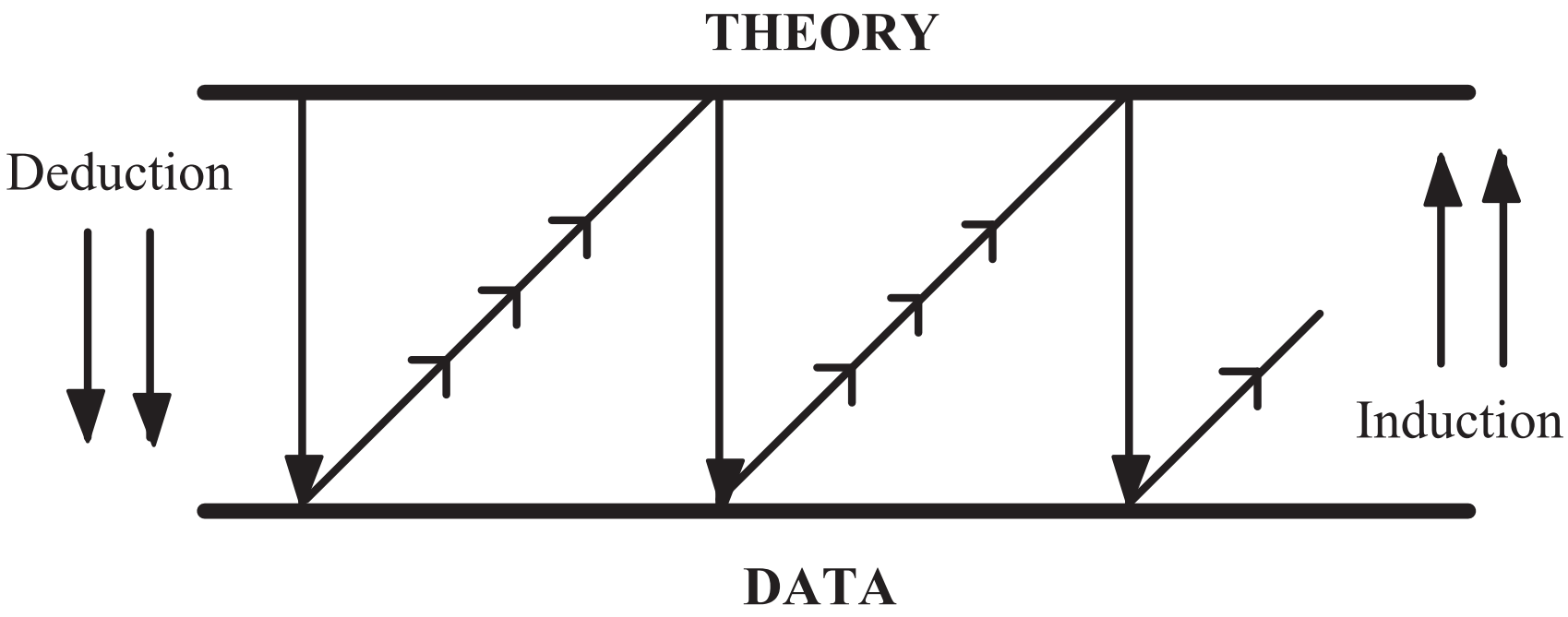
Aim

- general notions of exploratory (preliminary) statistics
- data variable types
- statistical characteristics and methods of graphical presentation qualitative variables
- statistical characteristics and methods of graphical presentation quantitative variables

Statistics – Science which Involves Study of Data

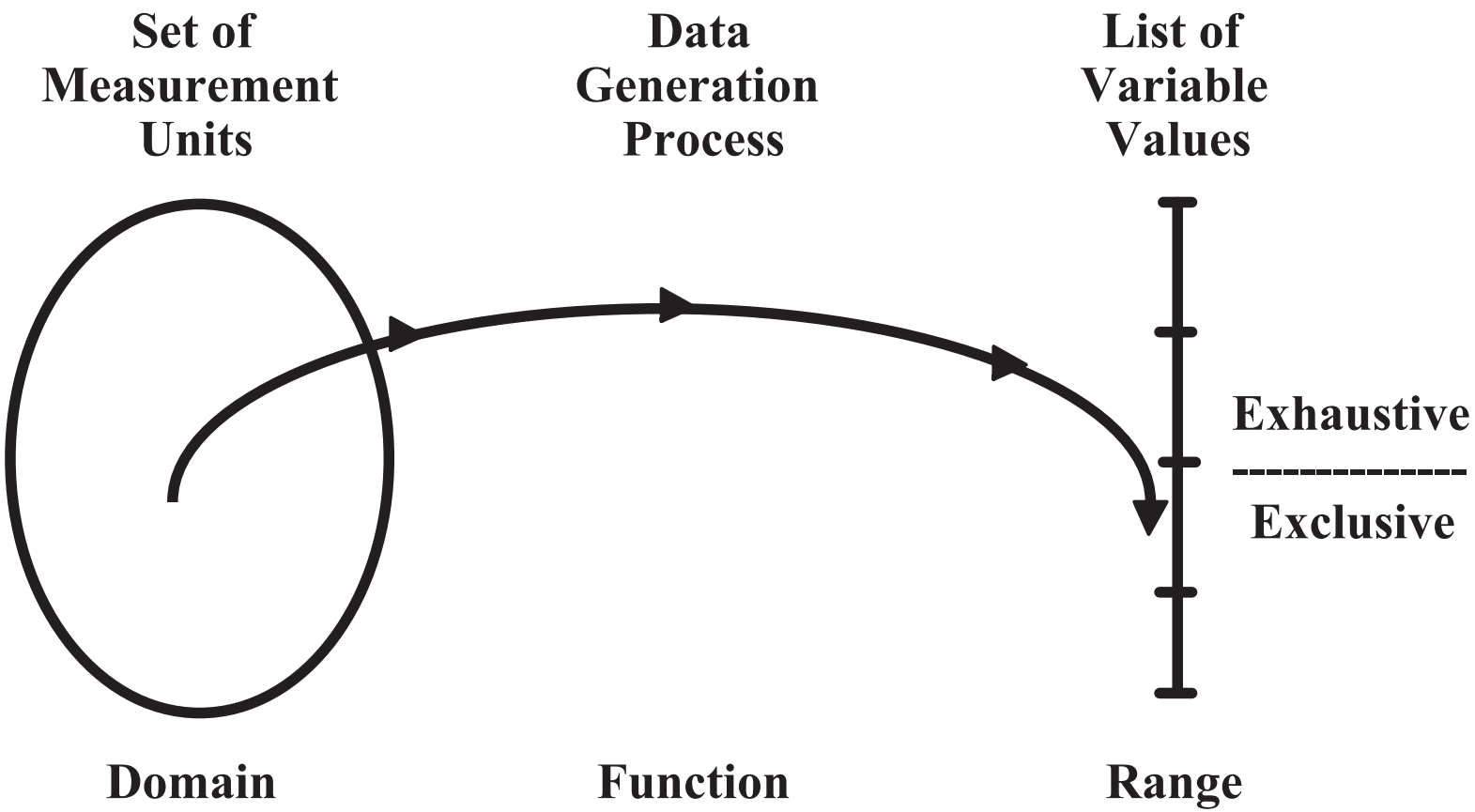
- The Objective of the Course
- A Definition of Statistics
- The Dual Role of Statistical Science

Deduction and induction

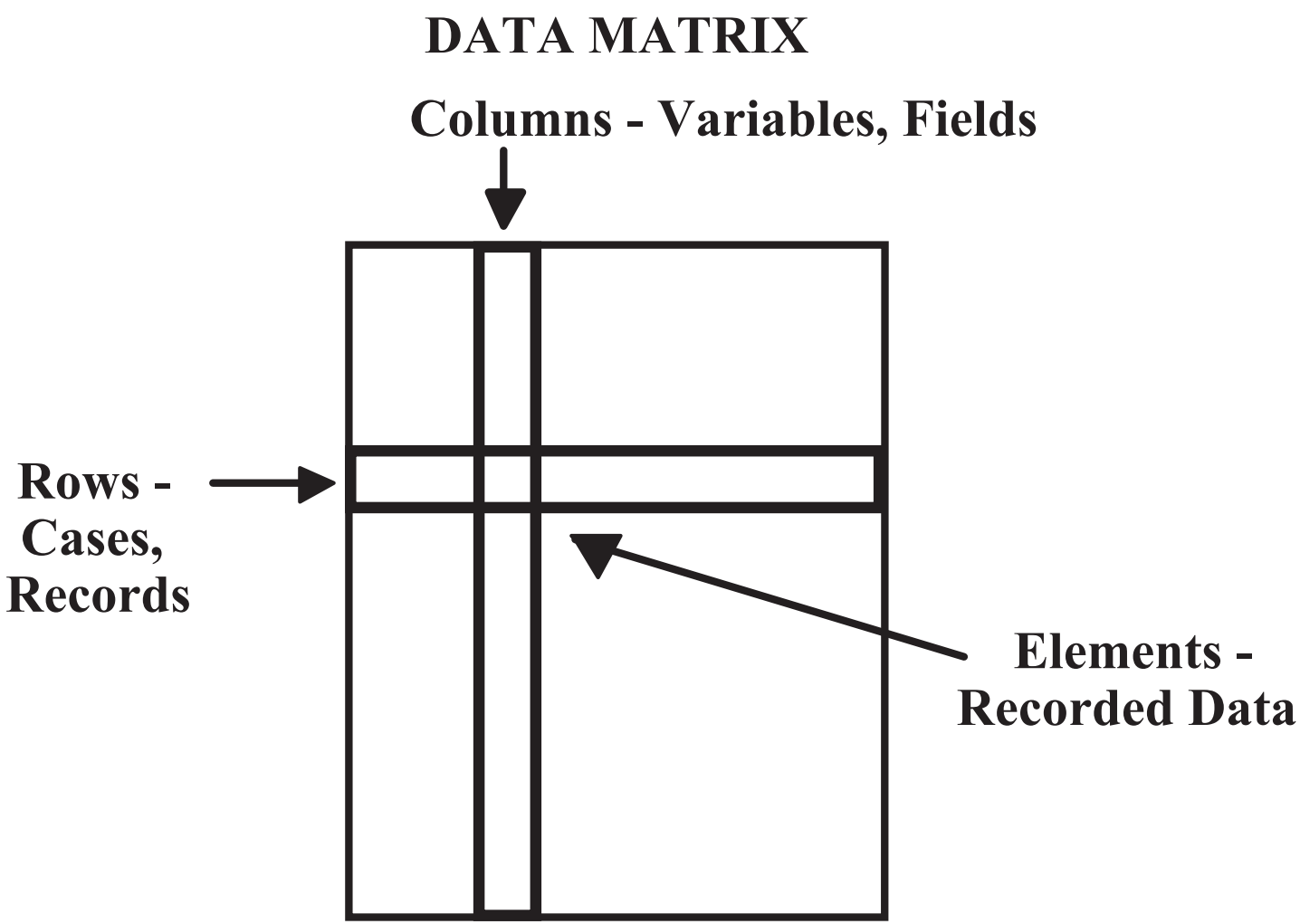


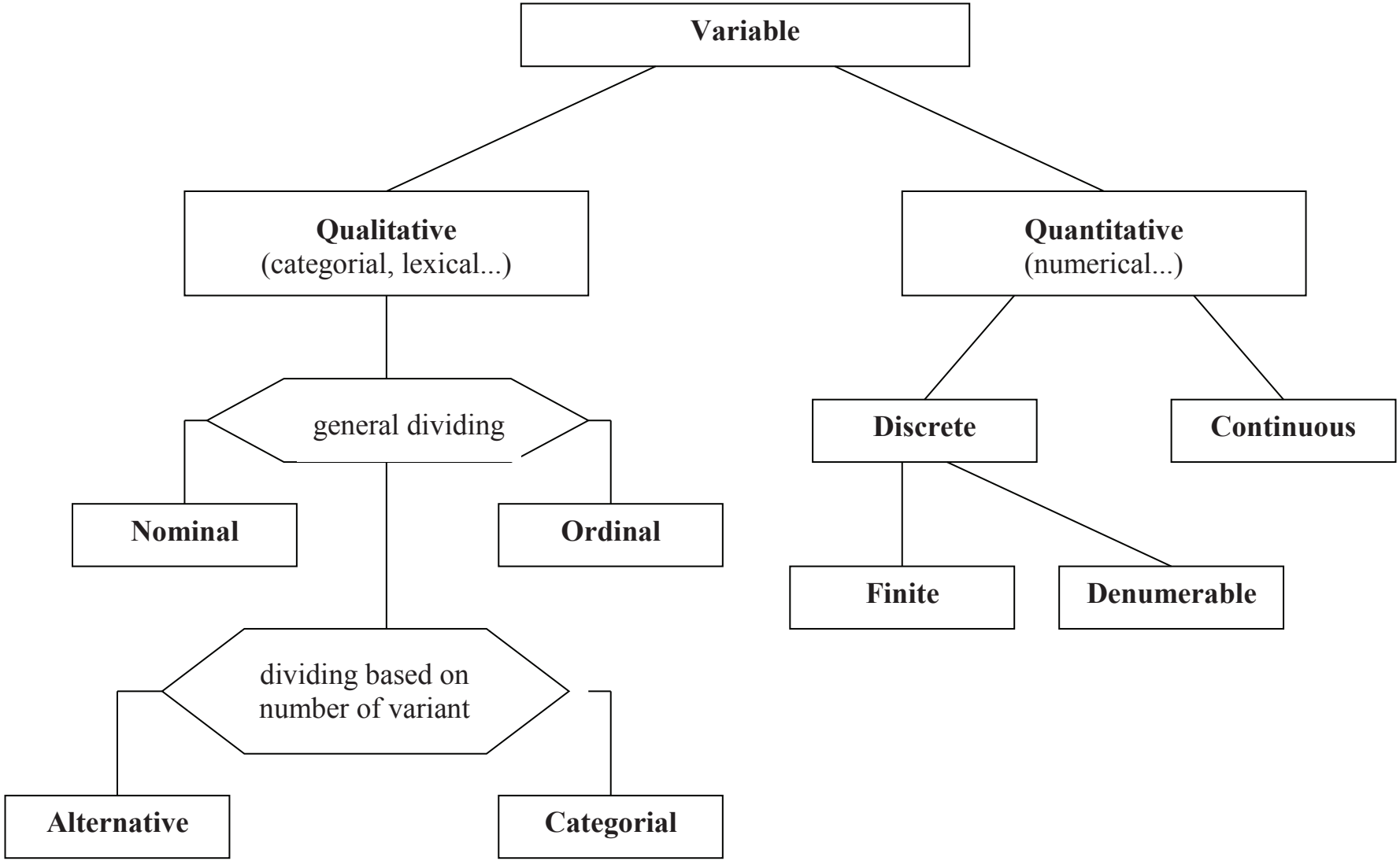
Data Generation Process

DATA DEFINITION AND CREATION



Data matrix


DATA MATRIX
Columns - Variables, Fields
Rows - Cases, Records
Elements - Recorded Data


Variable
Qualitative
(categorial, lexical...)
Quantitative
(numerical...)
general dividing
Nominal
Ordinal
Discrete
Continuous
Finite
Denumerable
dividing based on
number of variant
Alternative
Categorial

Statistical characteristics of qualitative variables

We know that qualitative variable has two basic types - nominal and ordinal.

Nominal variables

Nominal variable has different but equivalent variants in one group. Number of these variants is usually low and that's why the first statistical characteristics we use to describe it will be its **frequency**.

- **Frequency n_i** (absolute frequency)

  - is defined as number of occurrence variant of the qualitative variable

  In case that qualitative variable has *k* different variants (we describe their frequency n_1, n_2… n_k) in the statistical group (n values large) it must hold true:

$$n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = n$$

If we want express what part of the group forms variables with any variant we use relative frequency for description of variable.

- **Relative frequency p_i**

  - is defined as:

$$p_i = \frac{n_i}{n}, \textit{eventually}, p_i = \frac{n_i}{n} \cdot 100 \quad [\%]$$

(We use second formula in case if we want express of the relative frequency in percents).
It must hold true for relative frequency:

$$p_1 + p_2 + \ldots + p_k = \sum_{i=1}^{k} p_i = 1$$

When qualitative variables are processed it is good to order frequency and relative frequency into so-called **frequency table**:

FREQUENCY TABLE		
Value x_i	**Absolute frequency**	**Relative frequency**
	n_i	**p_i**
x_1	n_1	p_1
x_2	n_2	p_2
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	p_k
Total	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$

The last characteristic for nominal variable is mode.

- **Mode**

  - is defined as a variant name that have for the variable the most frequency

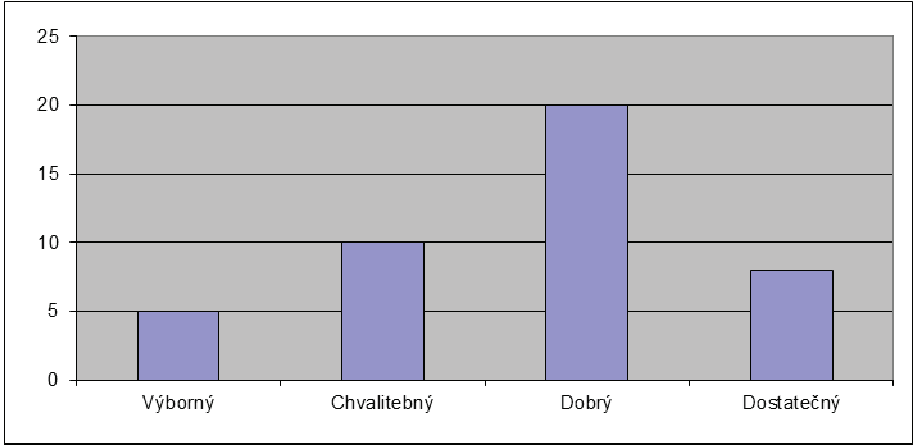
  The mode represented a typical element of the group. We don't determine mode in case that there is more variants with maximum frequency in the statistical group.

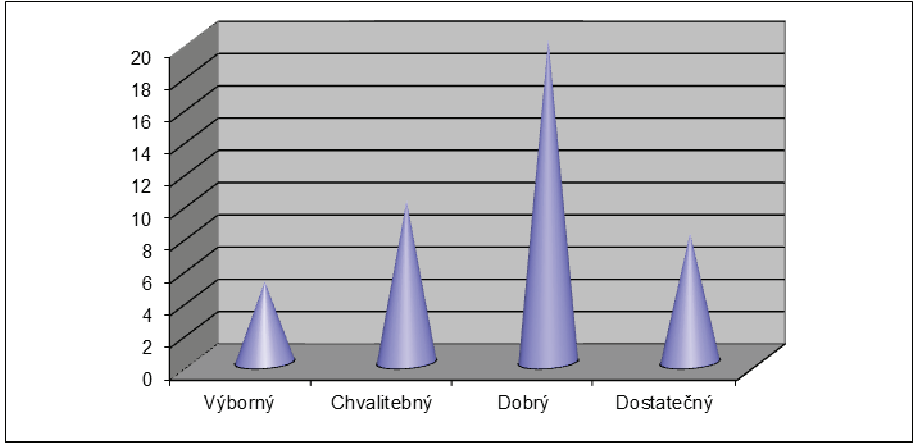
Graphical presentation of qualitative variables

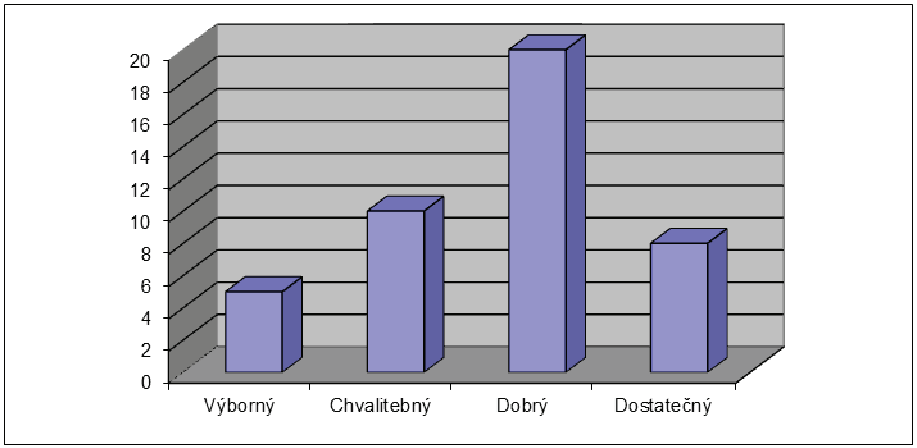
The statistics often use **graphs** for better plasticity of variables analysis. They are these two types for nominal variable:

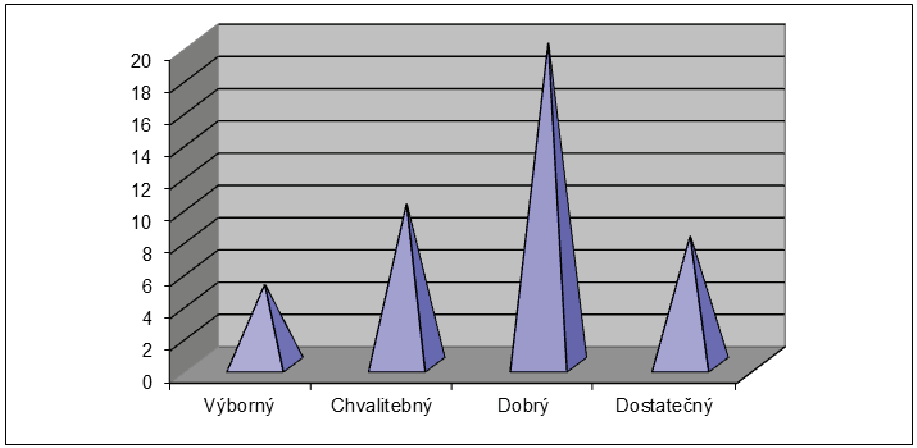
- **Histogram** (bar chart)
- **Pie chart**

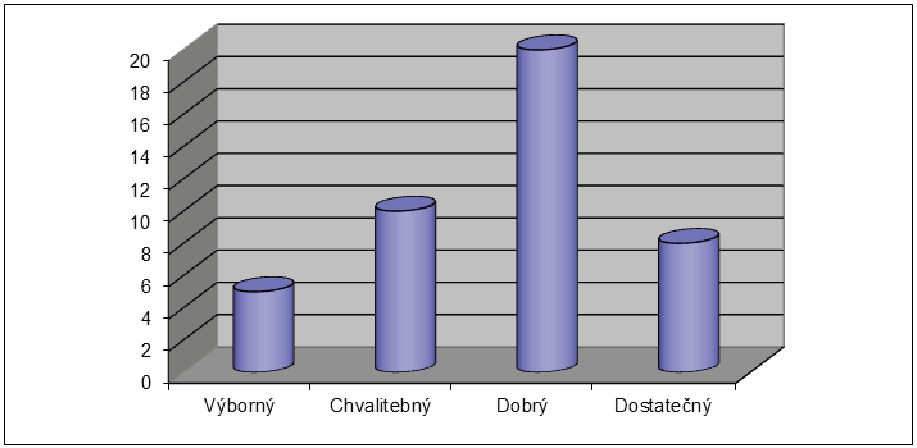
Histogram is a classical graph whereof we take variants of the variable on one axis and variable frequencies on the second one. Individual values of the frequency are then displayed as bars (boxes or vectors, squared logs, cones ...)

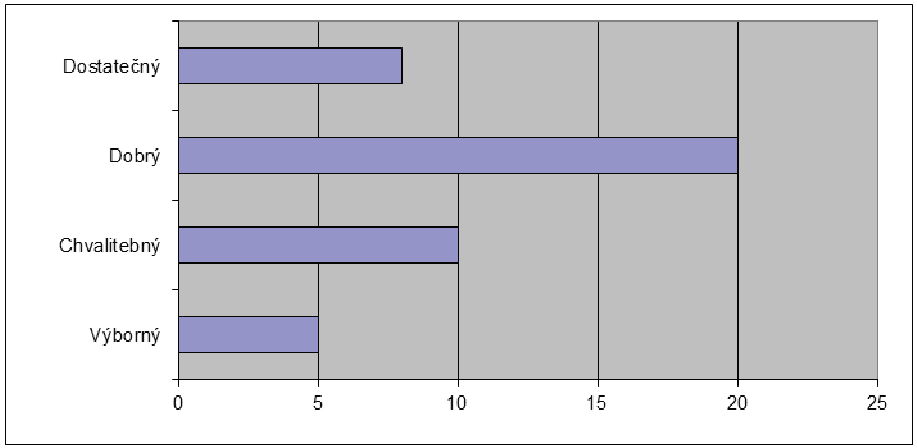

25
20
15
10
5
0
Výborný
Chvalitebný
Dobrý
Dostatečný

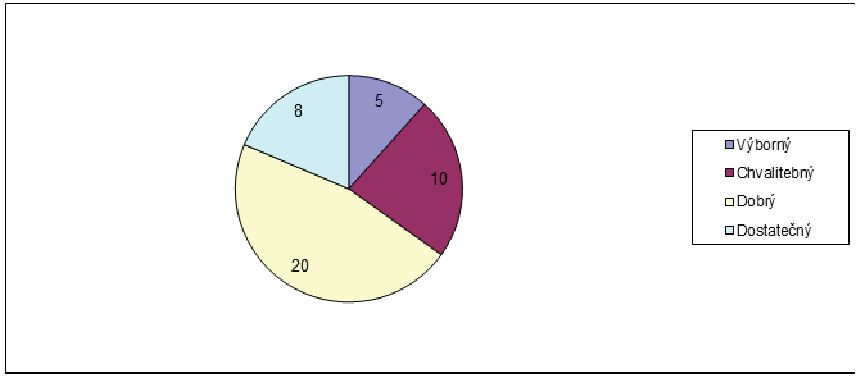

20
18
16
14
12
10
8
6
4
2
0
Výborný
Chvalitebný
Dobrý
Dostatečný

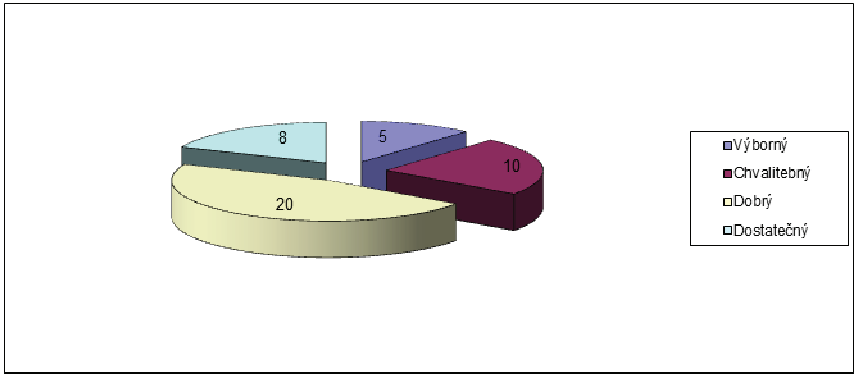

20
18
16
14
12
10
8
6
4
2
0
Výborný
Chvalitebný
Dobrý
Dostatečný

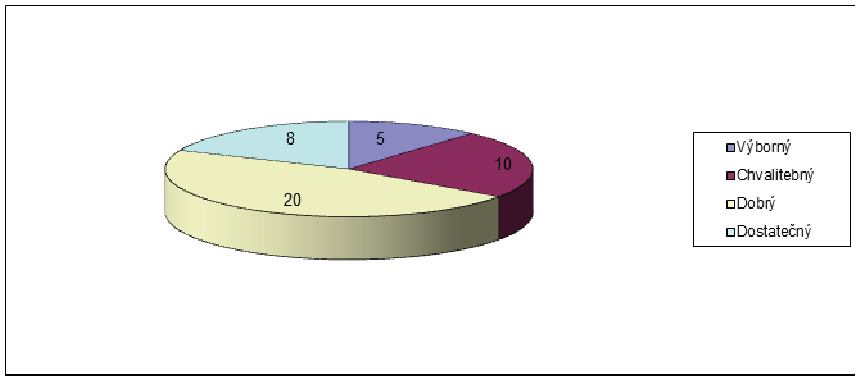

20
18
16
14
12
10
8
6
4
2
0
Výborný
Chvalitebný
Dobrý
Dostatečný

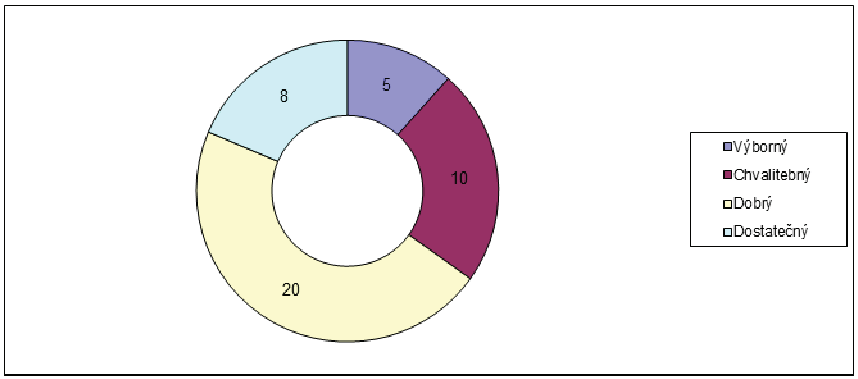

20
18
16
14
12
10
8
6
4
2
0
Výborný
Chvalitebný
Dobrý
Dostatečný


Dostatečný
Dobrý
Chvalitebný
Výborný
0
5
10
15
20
25









Pie graph represents relative frequencies of the individual variants of the variable. Individual relative frequencies are proportionally represent as a sector of a circle (when we change a circle to an ellipse we obtain three-dimensional effect).

Solved example

We made crossroad usage research. Obtained data are in following table. They represent color of cars that pass through crossroad. Analyze these data and represent results in graphical form.

red	blue	red	green
blue	red	red	white
green	green	blue	red

Solution:

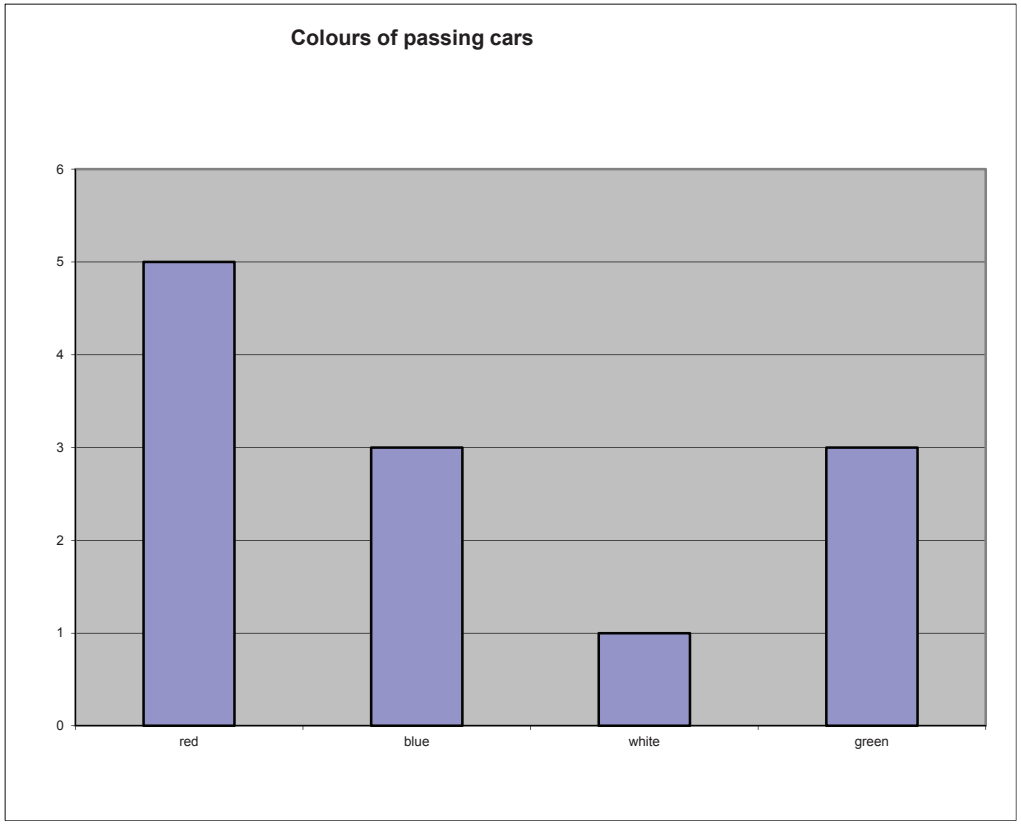
It's obvious that it's a qualitative (lexical) variable and considering the fact that there is no point in ordering or comparing colors of cars we can say it's a nominal variable.

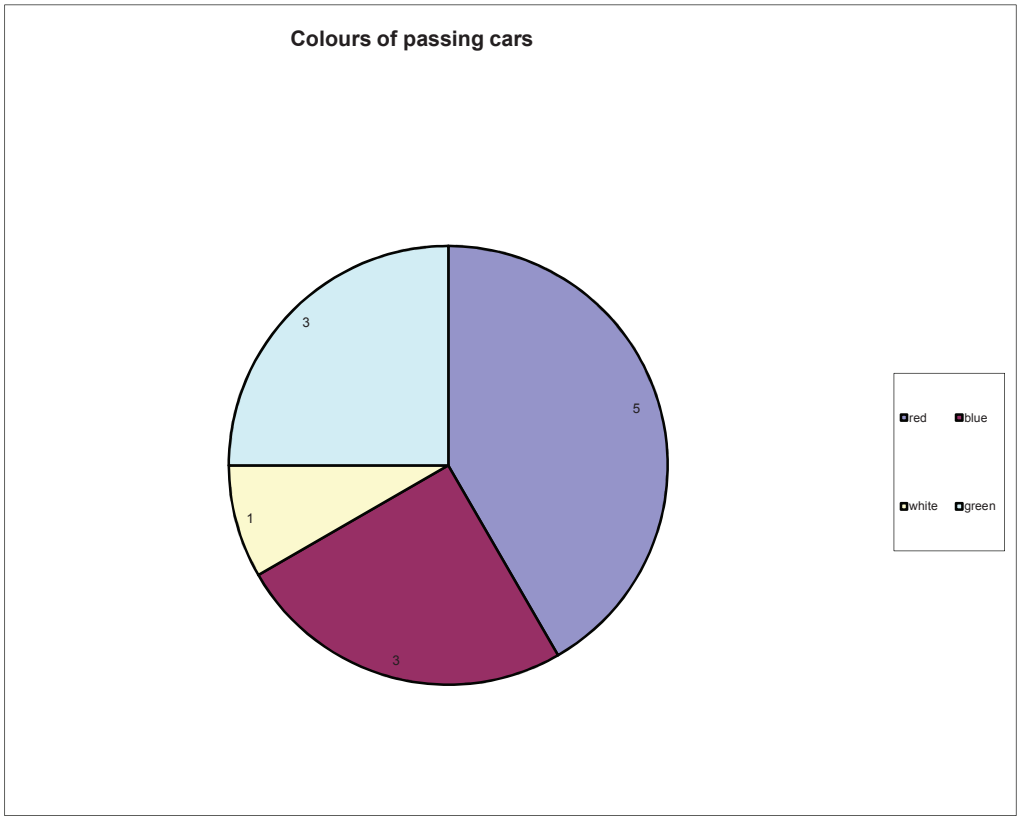
For its descriptions we choose frequency table we determine mode and color of passing vehicles we represent by histogram and pie graph.

FREQUENCY TABLE		
Colors of passing cars	**Absolute frequency**	**Relative frequency**
	$\mathbf{n_i}$	$\mathbf{p_i}$
red	5	$5/12 = 0{,}42$
blue	3	$3/12 = 0{,}25$
white	1	$1/12 = 0{,}08$
green	3	$3/12 = 0{,}25$
Total	12	1,00

We observed 12 cars total.

Mode = red (i.e. in our sample there were mostly red cars)


Colours of passing cars
6
5
4
3
2
1
0
red
blue
white
green


Colours of passing cars
3
5
1
3
red
blue
white
green

Ordinal variable

Now we will continue to ordinal variable description. Ordinal variable (as well nominal variable) has various lexical variants into group but these variants can be sort i.e. we can define which variant is "smaller" or "bigger".

For description ordinal variable we use same statistical characteristics and graphs such as for description nominal variable (frequency, relative frequency, mode + histogram, pie graph) extended two others characteristics (cumulative frequency and cumulative relative frequency) expressing sorting of ordinal variable.

- **Cumulative frequency of i-th variant m_i**

  - it's a number of values of variable showing the frequency of variants less or equal i-th variant

  E.g. we have a variable "classification from statistics". That has these variants: "1", "2", "3" or "4". Then for example cumulative frequency for variant "3" will be equal number of students who got classification "3" or better.

  If there are individual variants sort by their "size"

  ("$x_1 < x_2 < \ldots < x_k$") then it must holds true:

$$m_i = \sum_{j=1}^{i} n_j$$

  So it's obvious that cumulative frequency k-th ("the highest") variant is equal measure of the variable - n.

$$m_k = n$$

The second special characteristic for ordinal variable is cumulative relative frequency.

- **Cumulative relative frequency of i-th variant F_i**

  - a part of group are values gaining i-th and lower variant. It is expressed by this characteristic.

$$F_i = \sum_{j=1}^{i} p_j$$

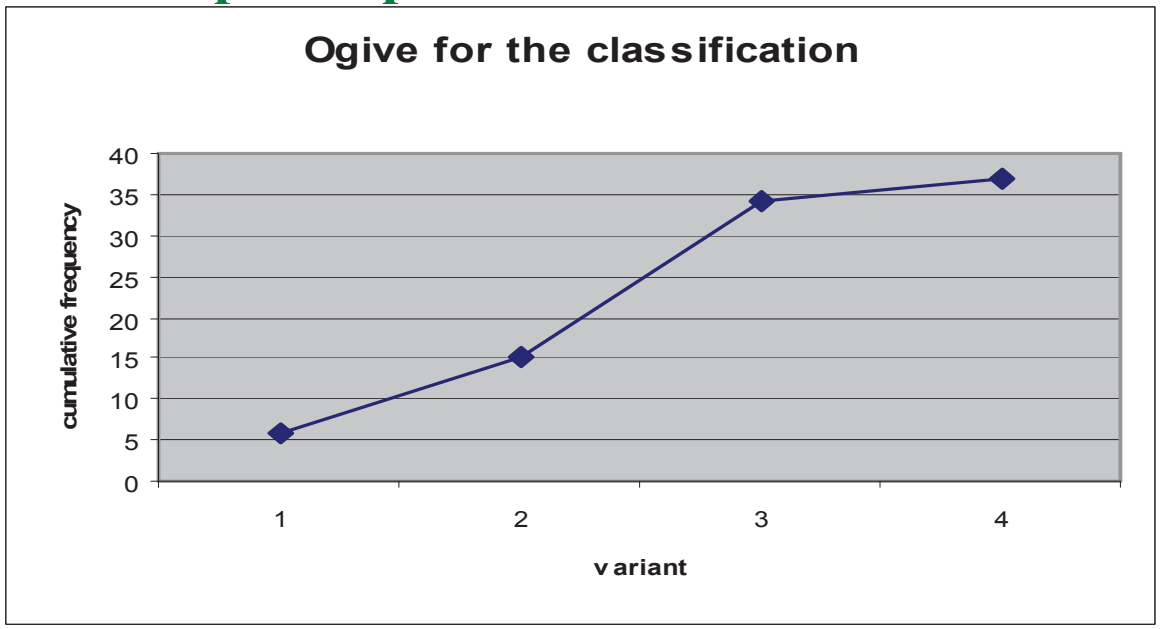
This is nothing else then relative expression of the cumulative frequency:

$$F_i = \frac{m_i}{n}$$

As well as at nominal variable we can present statical characteristics using frequency table at ordinal variable. It contains comparing with frequency table of nominal variable also values of cumulative and cumulative relative frequencies.

FREQUENCY TABLE				
Value x_i	**Absolute frequency**	**Cumulative frequencyt**	**Relative frequency**	**Relative cumulative frequency**
	n_i	**m_i**	**p_i**	**F_i**
x_1	n_1	$m_1 = n_1$	p_1	$F_1 = p_1$
x_2	n_2	$m_2 = n_1 + n_2 = m_1 + n_2$	p_2	$F_2 = p_1 + p_2 = F_1 + p_2$
x_k	n_k	$m_k = n_{k-1} + n_k = n$	p_k	$F_k = F_{k-1} + p_k = 1$
Total	$\sum_{i=1}^{k} n_i = n$	-----	$\sum_{i=1}^{k} p_i = 1$	-----

Graphical presentation of ordinal variables

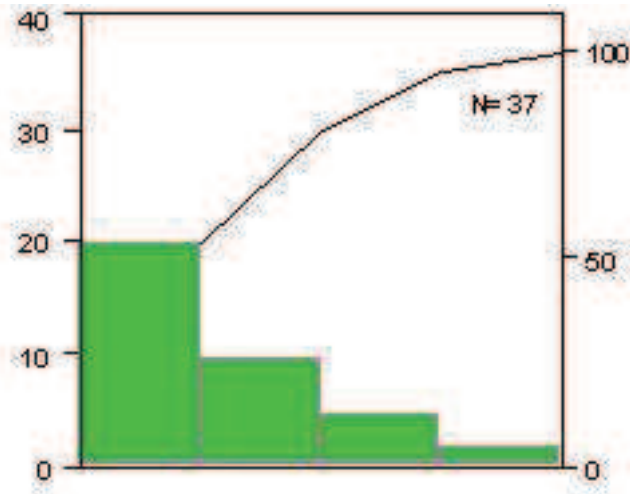


Ogive (cumulative frequency polygon)

- it's a frequency polygon of the cumulative frequency or the relative cumulative frequency. The vertical axis is the cumulative frequency or relative cumulative frequency. The horizontal axis represents possible variants. The graph always starts at zero at the lowest variant and will end up at the total frequency (for a cumulative frequency) or 1.00 (for a relative cumulative frequency).

Pareto graph

- it's a bar chart for qualitative variable with the bars arranged according to frequency
- there are particular variants on horizontal axis ordered from the one with "the biggest" importance to "the smallest one"



Statistical characteristics of quantitative variables

For description quantitative variable we can use most of the statistical characteristics that are used for ordinal variable description (frequency, relative frequency, cumulative frequency and cumulative relative frequency). With these characteristics we add another two characteristics:

- **measures of position** – those indicate a typical distribution of the variable values (dislocation on the numerical axis)

and

- **measures of variability** – those indicate a variability (variance) of the values round their typical position

Measures of position and variability

The most used measure of position is a mean of variable. The mean represents average or typical value of the sampling population. The most famous mean for quantitative variable is:

- **Arithmetical mean** $\bar{x}$

  Its value we obtain by means of this formula:

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

where: x_i ... particular values of the variable
n ... size of the sampling population (number of the values of the variable)

Mode

$\hat{x}$

In case of mode we will discern between discrete and continuous quantitative variable. **For discrete variable** we define **mode** $\hat{x}$ as value of the most frequency of the variable (analogous to by the qualitative variable).

But by the **continuous variable** we think of mode $\hat{x}$ as value around which is the most concentration of variable values.

For assessment of this value we use **shorth**. Short is the shortest interval whereof lies at least 50% of variable values. In case of sample large as $n = 2k\ (k \in \mathrm{N})$ **(**even number of values) k values lies in short - what is n/2 (50%) variable values. In case of sample large as $n = 2k+1\ (k \in \mathrm{N})$ (odd number of values) $k+1$ values lies in short - what is about 1/2 more then 50% variable values (n/2+1/2).

Then we define **mode** $\hat{x}$ as centre of the short.

Solved example

The following data represent age of the musicians which played on the concert. The variable age is a continuous. Determine mean, short and mode for the variable.

22 82 27 43 19 47 41 34 34 42 35

Solution:

a) Mean:

In this case we use arithmetical mean:

Average age is 38,7 year for musician played on the concert.

$$\bar{x} = \frac{\sum_{i=!}^{n} x_i}{n} = \frac{22+82+27+43+19+47+41+34+34+42+35}{11} = 38{,}7\ let$$

b) Shorth:

Our sample population has 11 values. 11 is odd number of values. 50% of this is 5,5 and the nearest higher natural number is 6 - otherwise: n/2+1/2 = 11/2+1/2 = 12/2 = 6. There out imply that 6 values will be lies in the shorth.

- we sort variable
- we determine size of all intervals (having 6 elements) in which $x_i < x_{i+1} < \ldots < x_{i+5}$
- the shortest of these intervals is shorth (size of the interval = $x_{i+5} - x_i$

Original data	Sorting data	Size of intervals (having 6 elements)
22	19	16 (= 35 – 19)
82	22	19 (= 41 – 22)
27	27	15 (= 42 – 27)
43	***34***	***9*** (= 43 – 34)…min
19	***34***	13 (= 47 – 34)
47	***35***	47 (= 82 – 35)
41	***41***	
34	***42***	
34	***43***	
42	47	
35	82	

From table we see that the shortest interval has size 9. The only one interval corresponds to this size: $\langle 34;43 \rangle$.

Shorth = $\langle 34;43 \rangle$. This we will interpret as half of musicians are 34 to 43 years old.

c) Mode:

Mode is defined as a center of the shorth:

$$\hat{x} = \frac{34 + 43}{2} = 38{,}5$$

Mode = 38,5 year, i.e. typical age is 38,5 year for musician played on the concert.

Other characteristics based on ordering

- **Quantiles**

Quantiles are characteristics which describe of location of individual values (within the scope variable). The quantiles are resistant to outlier observation analogous to mode. Generally the **quantile is defined as a value which divides a sample into two parts - the first one contains values that are less than given quantile and the second one contains values that are bigger or equal than given quantile.** We must have got sorted data (from the least to the biggest value).

Quantile of variable x which separates 100p % smaller values from rest of sample (i.e. from 100(1-p)% values) we call **100p % quantile** and we mark it $\mathbf{x_p}$.

In work we most often meet these quantiles:

- **Quartiles**

When division is into four parts the values of the variate corresponding to 25%, 50% and 75% of the total distribution are called quartiles.

Lower quartile $\mathbf{x_{0,25}}$ = 25% quantile (it divides a sample of data so that 25% of values is less than this quartil, i.e. 75% is bigger (or equal))

Median $\mathbf{x_{0,5}}$ = 50% quantile (it divides a sample of data so that 50% of values is less than median and 50% of values is bigger (or equal))

Upper quartile $\mathbf{x_{0,75}}$ = 75% quantile (it divides a sample of data so that 75% of values is less than this quartil, i.e. 25% is bigger (or equal))

Example:

Data	6 47 49 15 43 41 7 39 43 41 36
Ordered Data	6 7 15 36 39 41 41 43 43 47 49
Median	41
Upper quartile	43
Lower quartile	15

The difference between the 1st and 3rd quartiles is called the **inter-quartile range (IQR)**.

$IQR = x_{0,75} - x_{0,25}$

Example:

Data	2 3 4 5 6 6 6 7 7 8 9
Upper quartile	7
Lower quartile	4
IQR	7 - 4 = 3

- **Deciles** – $x_{0,1}; x_{0,2}; \ldots ; x_{0,9}$
  The deciles divide the data into 10 equal regions.

- **Percentiles** – $x_{0,01}; x_{0,02}; \ldots; x_{0,99}$

The percentiles divide the data into 100 equal regions.

For example, the 80th percentile is the number which has 80% below it and 20% above it. Rather tan counting 80% from the bottom, count 20% from the top.

Note: The 50th percentile is the median.

- **Minimum x_{min} and Maximum x_{max}**

$x_{\min} = x_0$, i.e. 0% of values are less than minimum

$x_{\max} = x_1$, i.e. 100% of values are less than maximum

Now we talk about **relation** between **quantiles and cumulative relative frequency**. The value p denotes cumulative relative frequency of quantile x_p i.e. relative frequency of those variable values that are lesser than quantile x_p. Quantile and cumulative relative frequency are inverse notions.

Graphical or tabular representation of the ordered variable and appropriate cumulative frequencies is designated as **distribution function of the cumulative frequency** or **empirical distribution function**.

- **Empirical distribution function F(x) for the quantitative variable**

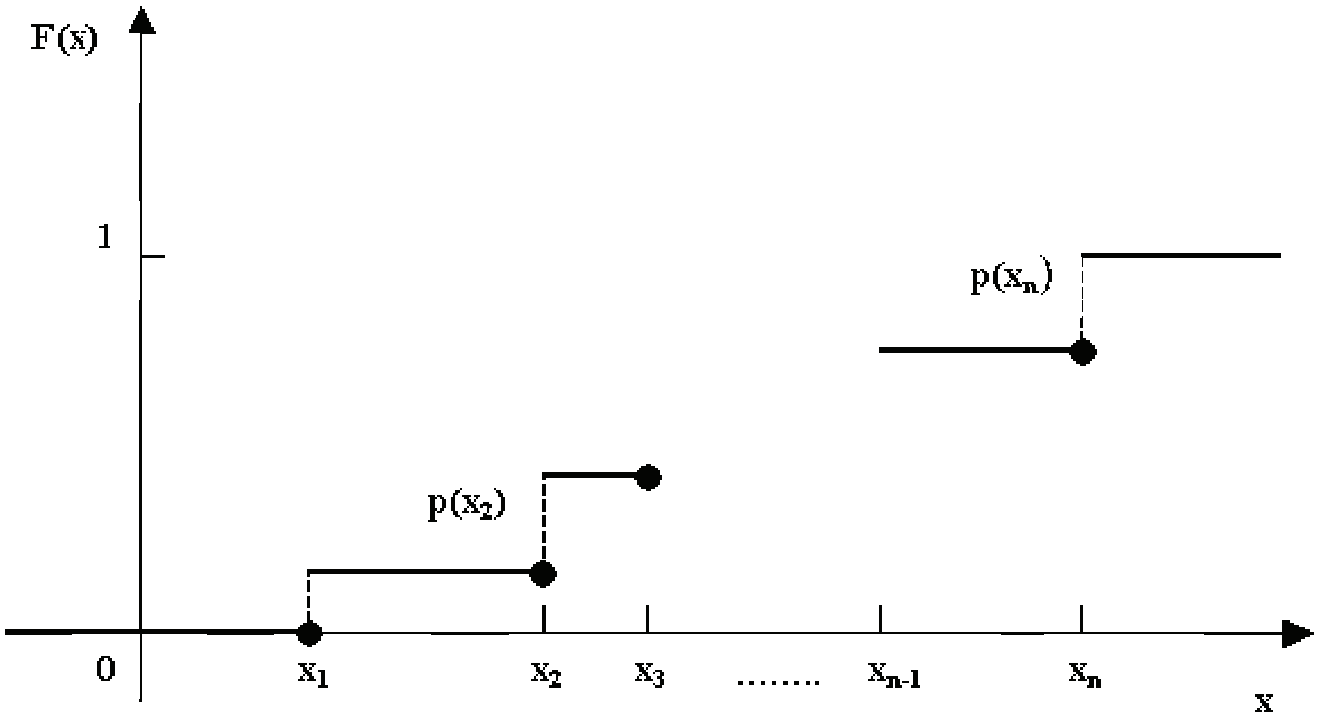
We have ordered sample population ($x_1 < x_2 < \ldots < x_n$) and we denote $p(x_i)$ as relative frequency of the value x_i. Then it must hold true for empirical distribution function F(x):

$$F(x) = \begin{cases} 0 & for \quad x \le x_1 \\ \sum_{i=1}^{j} p(x_i) & for\ x_j < x \le x_{j+1}, 1 \le j \le n-1 \\ 1 & for\ x_n < x \end{cases}$$

The empirical distribution function is monotonous increasing function and it is continuous from the left.

We see:

$$p(x_i) = \lim_{x \to x_{i+}} F(x) - F(x_i)$$



Measures of Variability

- **IQR**

$$IQR = x_{0,75} - x_{0,25}$$

MAD

MAD is a name for **m**edian **a**bsolute **d**eviation from the median.

We determine MAD in this way:

1. we order the sampling population by size
2. we determine a median of the sampling population
3. for each value we determine absolute value of its deviation from the median
4. the absolute deviations from the median we order by size
5. now we determine a median of the absolute deviations from the median i.e. MAD

Solved example

We have these data: 22, 82, 27, 43, 19, 47, 41, 34, 34, 42, 35 (these are the same data as the previous solved example).
Determine: MAD
Solution:

Original data	Ordered data	Sequence
22	19	1
82	22	2
27	**27**	**3**
43	34	4
19	34	5
47	**35**	**6**
41	41	7
34	42	8
34	**43**	**9**
42	47	10
35	82	11

Median $x_{0,5}$:

$$p = 0{,}5;\ n = 11 \Rightarrow z_p = 11.0{,}5 + 0{,}5 = 6 \Rightarrow x_{0,5} = 35$$

i.e. a half of the musician is younger then 35 years (50% of them have 35 years and

Original data x_i	Ordered data y_i $\Rightarrow x_{0,5} = 35$	Absolute deviations from median $\lvert y_i - x_{0,5} \rvert$	Ordered absolute values M_i
22	19	$16 = \lvert 19 - 35 \rvert$	0
82	22	$13 = \lvert 22 - 35 \rvert$	1
27	27	$8 = \lvert 27 - 35 \rvert$	1
43	34	$1 = \lvert 34 - 35 \rvert$	6
19	34	$1 = \lvert 34 - 35 \rvert$	7
47	**35 (median)**	$0 = \lvert 35 - 35 \rvert$	**8 (MAD)**
41	41	$6 = \lvert 41 - 35 \rvert$	8
34	42	$7 = \lvert 42 - 35 \rvert$	12
34	43	$8 = \lvert 43 - 35 \rvert$	13
42	47	$12 = \lvert 47 - 35 \rvert$	16
35	82	$47 = \lvert 82 - 35 \rvert$	47

(MAD is a median absolute deviation from the median i.e. 6th value of ordered absolute deviations from the median)

MAD = 8.

Other Measures of Variability

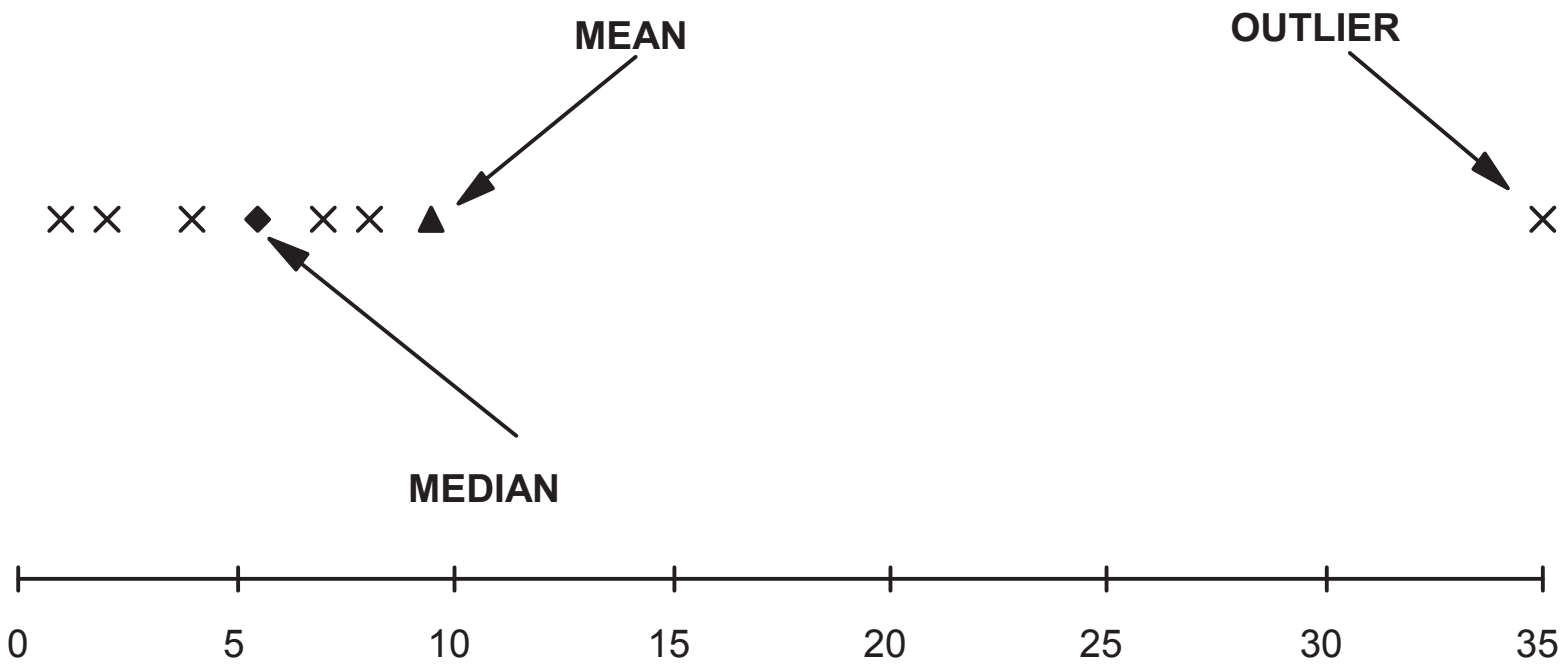
- **Sample variance s^2**

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

- **Sample standard deviation s**

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

Identification of outliers



Identification of outliers

1. The outlier can be such value x_i where absolute value of z-axis is greater then 3:

$$z_i = \frac{x_i - \overline{x}}{s}, \qquad (|z_i| > 3) \Rightarrow x_i \text{ is an outlier}$$

2. The outlier can be such value x_i where absolute value of median-axis is greater then 3:

$$m_i = \frac{x_i - x_{0,5}}{1,483.MAD} \qquad (|m_i| > 3) \Rightarrow x_i \text{ is an outlier}$$

3. The outlier can be such value x_i that is far more then 1,5 IQR from lower (or upper) quantile.

$$[(x_i < x_{0,25} - 1,5IQR) \vee (x_i < x_{0,75} + 1,5IQR)] \Rightarrow x_i \text{ is an outlier}$$

Sample skewness α

- **Skewness**

  - Skewness is defined as asymmetry in the distribution of the variable values. Values on one side of the distribution tend to by further from the "middle" than values on the other side.

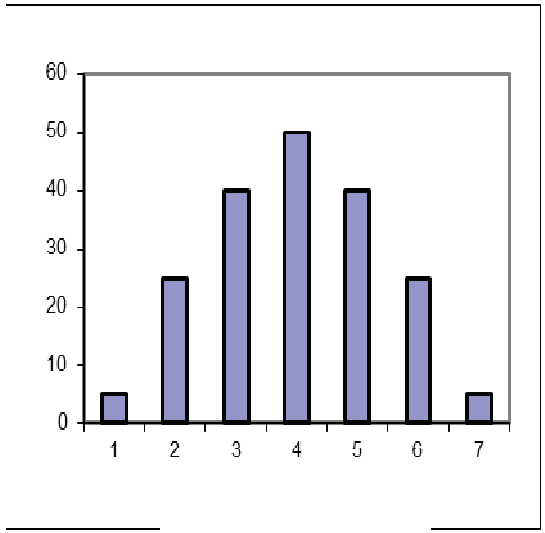
  - Its value we obtain by means of this formula:

$$\alpha = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3} \qquad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

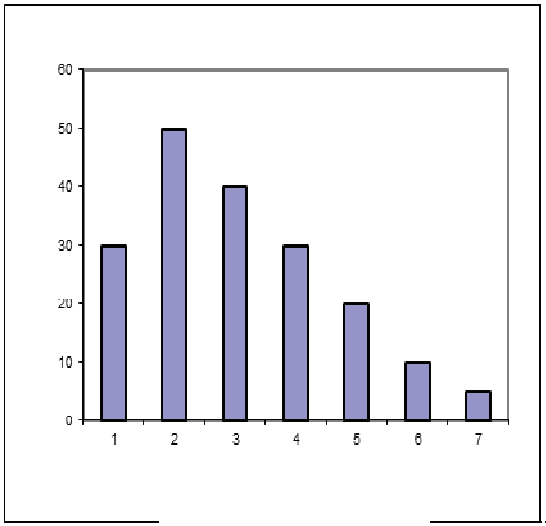
$\alpha = 0$ ... variable values are distributed symmetrically round the mean

$\alpha > 0$ ... predominate values are less then mean of the variable

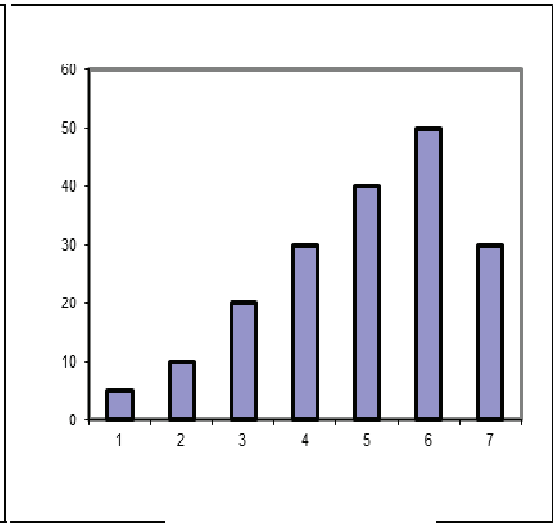
$\alpha < 0$ ... predominate values are greater then mean of the variable



$\alpha=0$



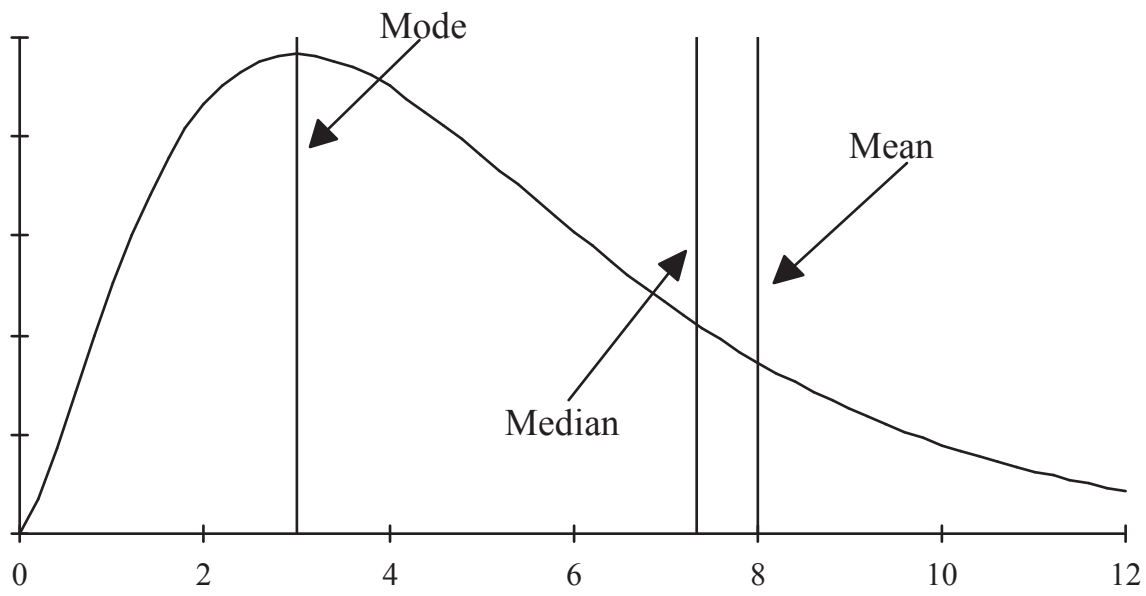
$\alpha>0$



$\alpha<0$

ILLUSTRATION OF SKEWNESS

The Gamma distribution, a theoretical distribution commonly used to study reliability of manufactured products, has **postive skew**. As a result, the mean is greater than the median and both are much greater than the mode.



Sample **kurtosis β**

- Kurtosis represents concentration of variable values round their mean.

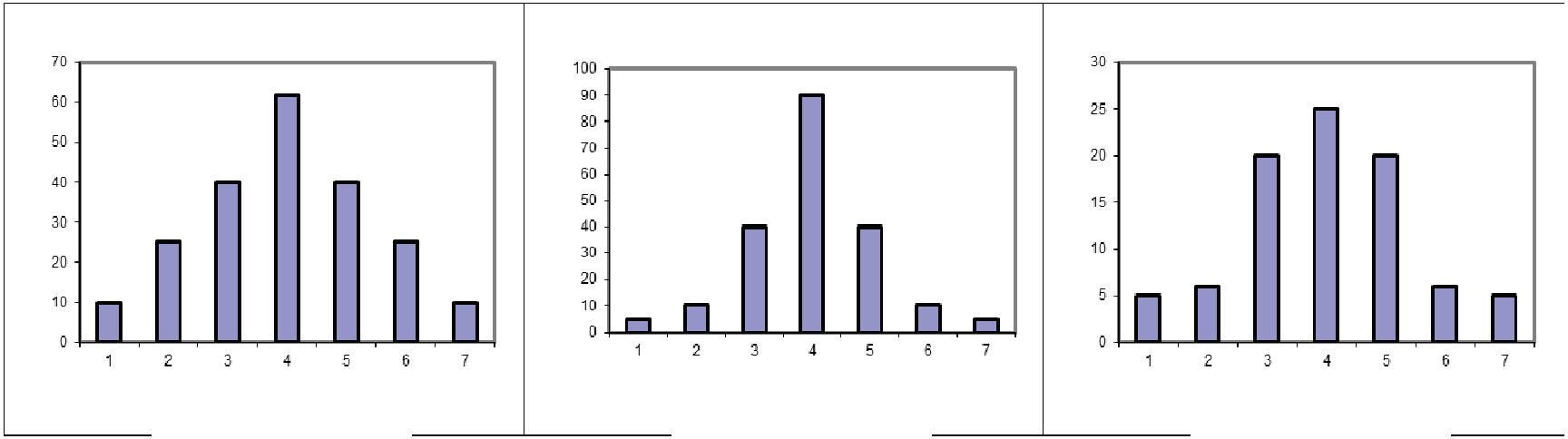
- Its value we obtain by means of this formula:

$$\beta = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^4}{s^4} - 3$$

$\beta = 0$ ... Kurtosis corresponds to normal distribution

$\beta > 0$ ... "peaked" distribution of the variable

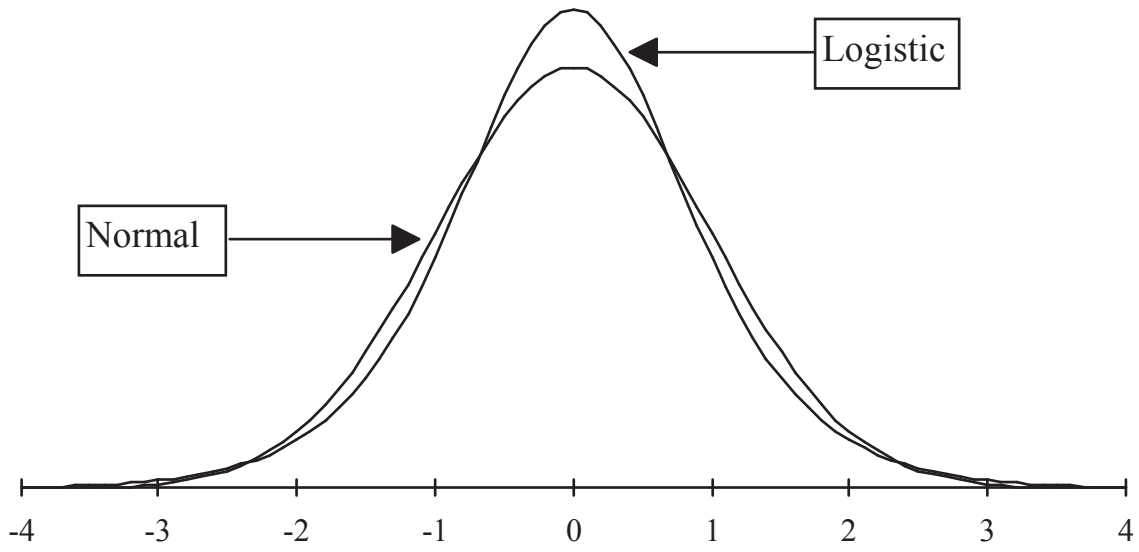
$\beta < 0$ ... "flat" distribution of the variable



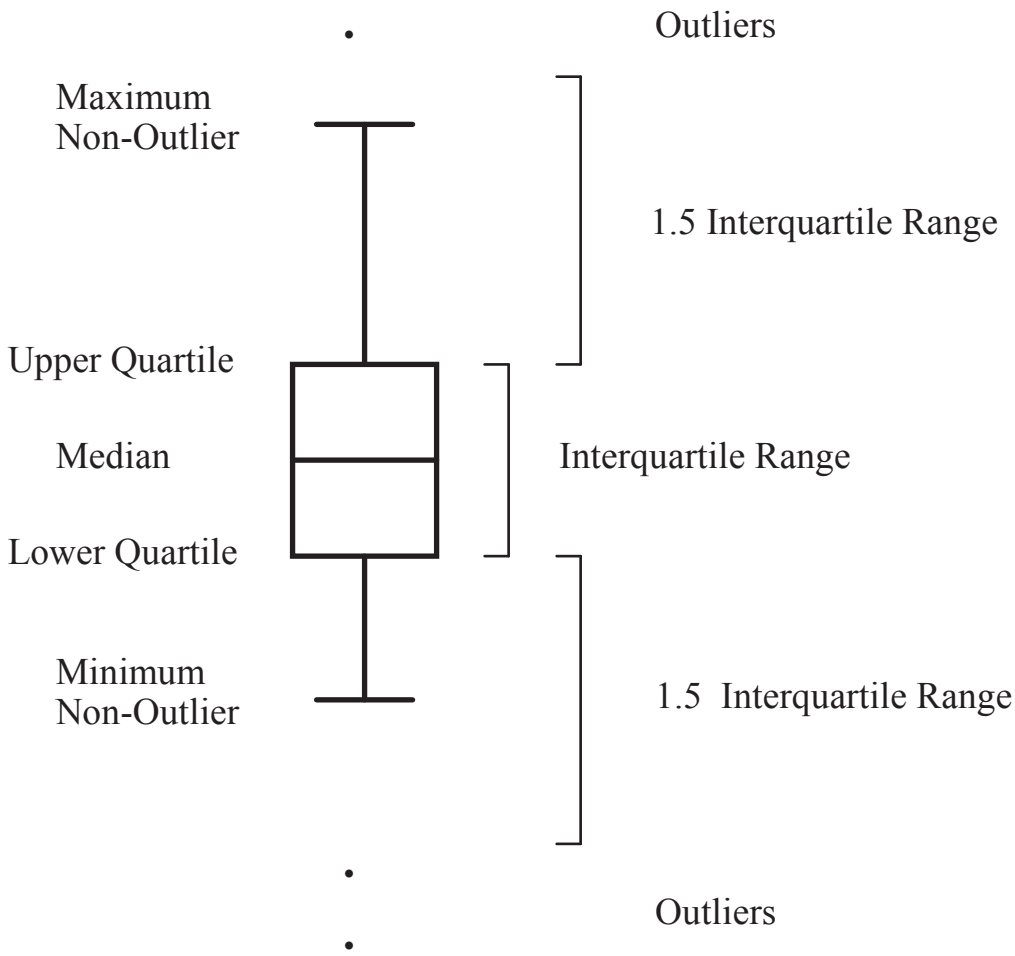
$\beta = 0$ $\beta > 0$ $\beta < 0$

ILLUSTRATION OF KURTOSIS

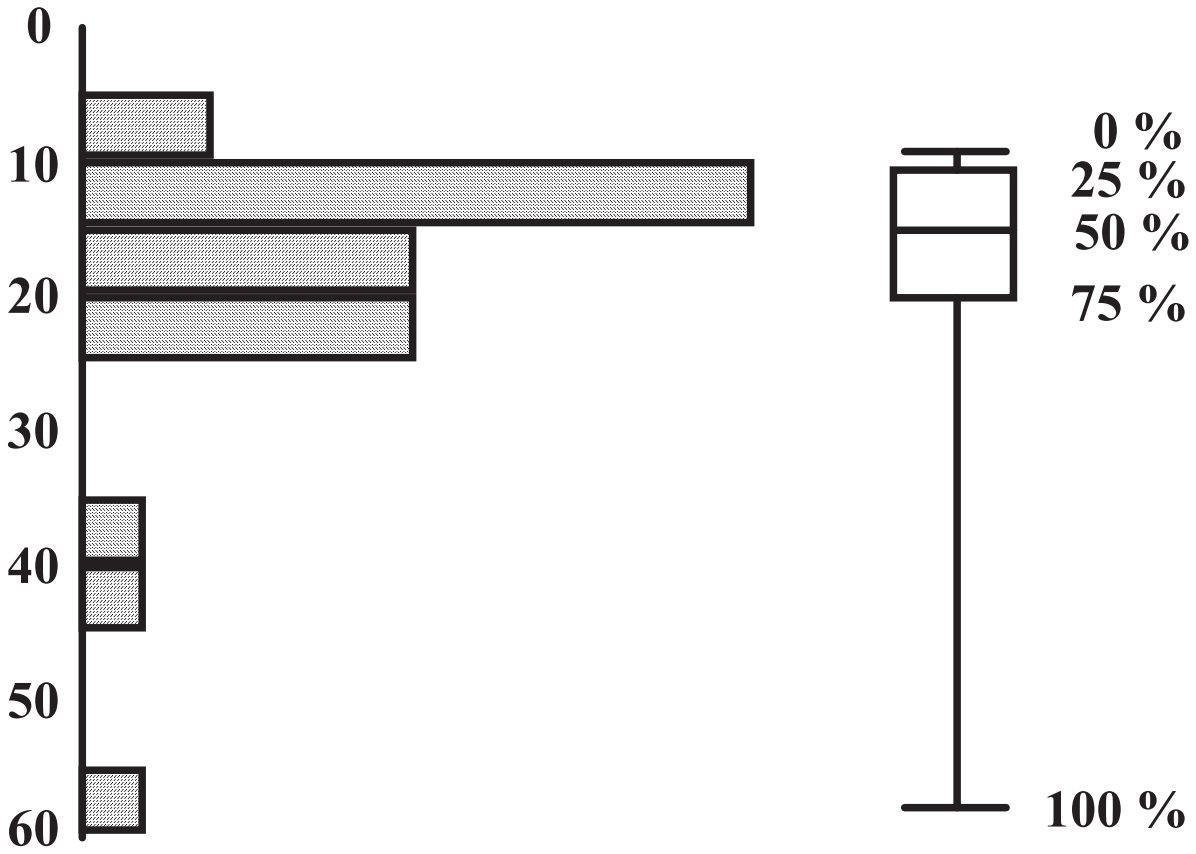
Two theoretical symmetric distributions with different kurtosis are the normal and the logistic. In the following illustration, both distributions have the same mean and standard deviation, but the logistic distribution has **greater kurtosis than the standard normal**.



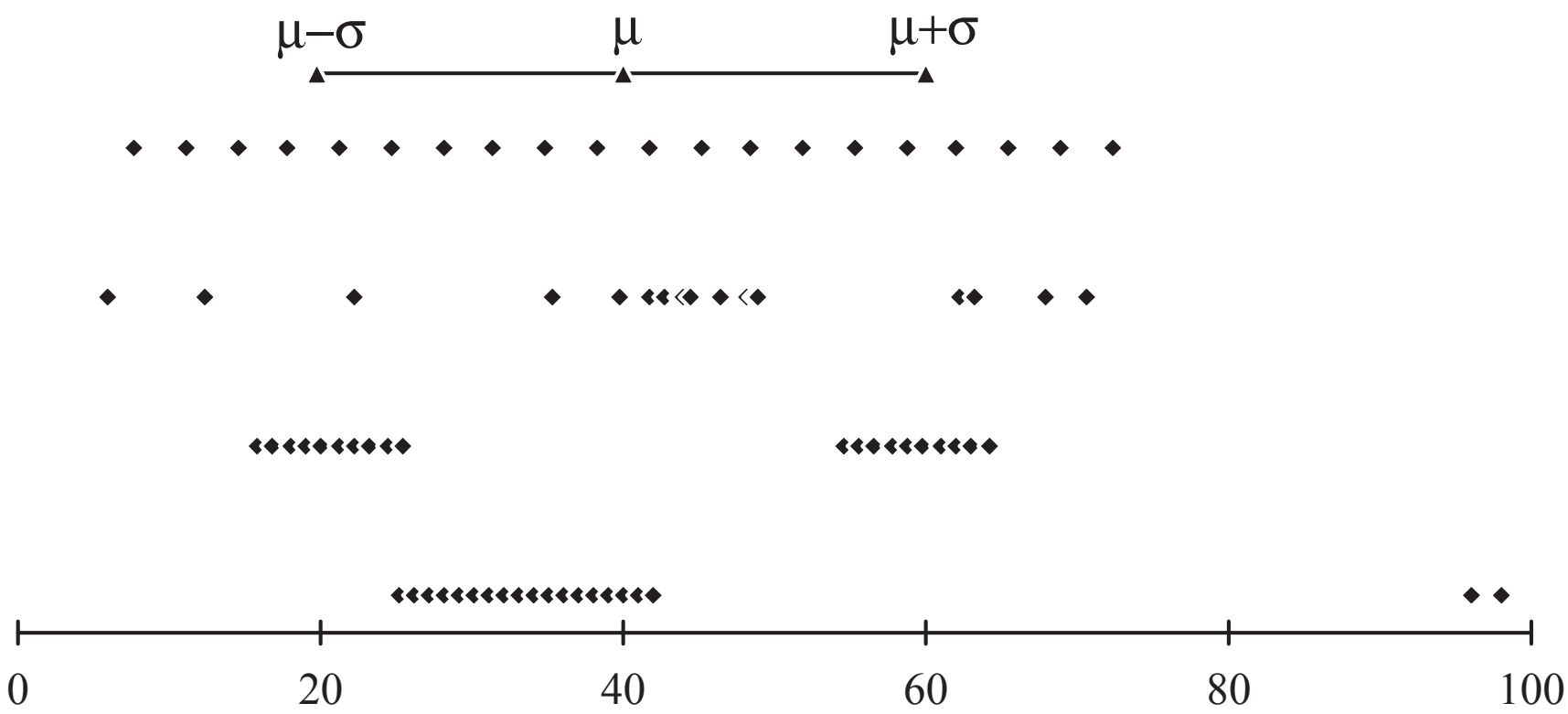
Graphical presentation of quantitative variable – Box Plot


Outliers
Maximum
Non-Outlier
1.5 Interquartile Range
Upper Quartile
Median
Interquartile Range
Lower Quartile
Minimum
Non-Outlier
1.5 Interquartile Range
Outliers

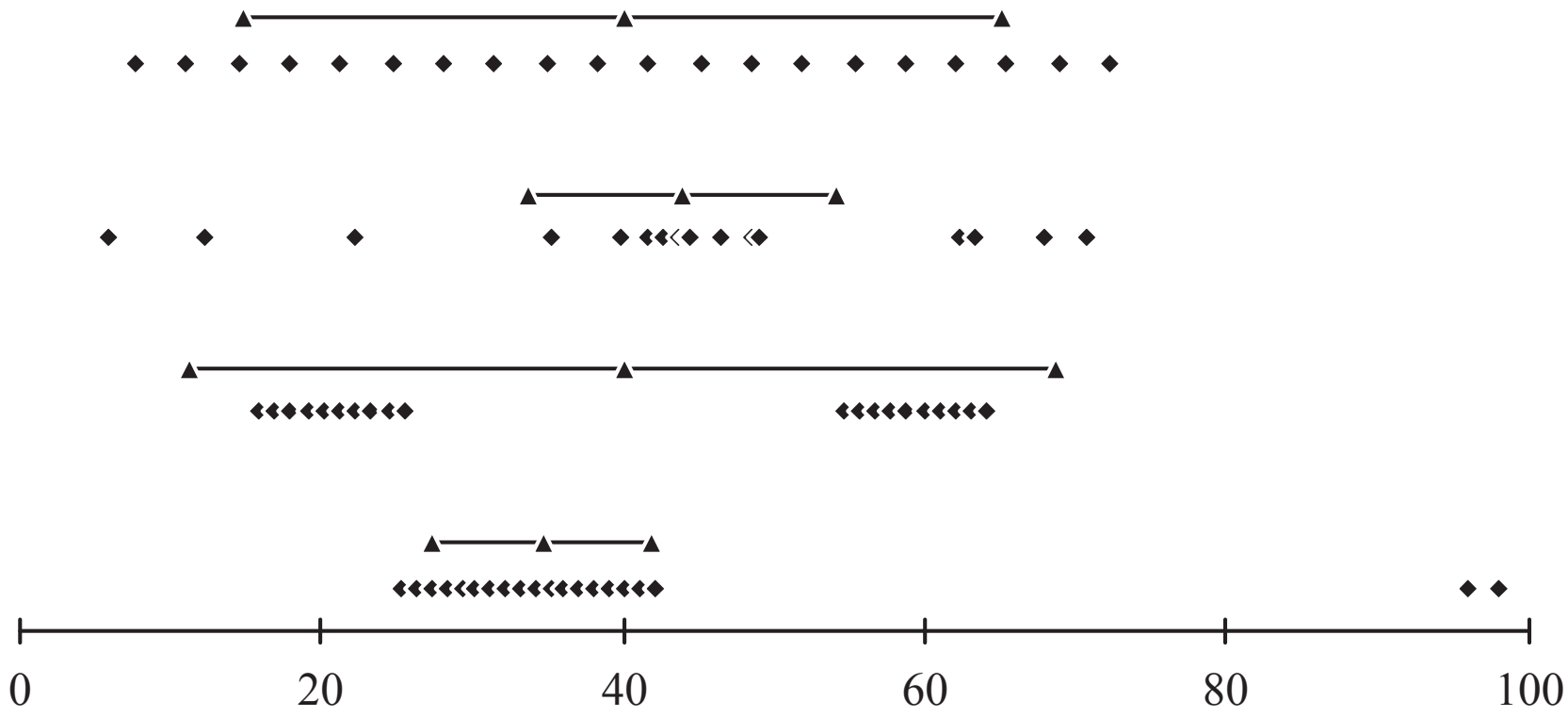
Graphical presentation of quantitative variable



All distributions have mean = 40 and standard deviation = 20.11



Median +/- 1.483 MAD For Each Distribution



Stem and leaf plot

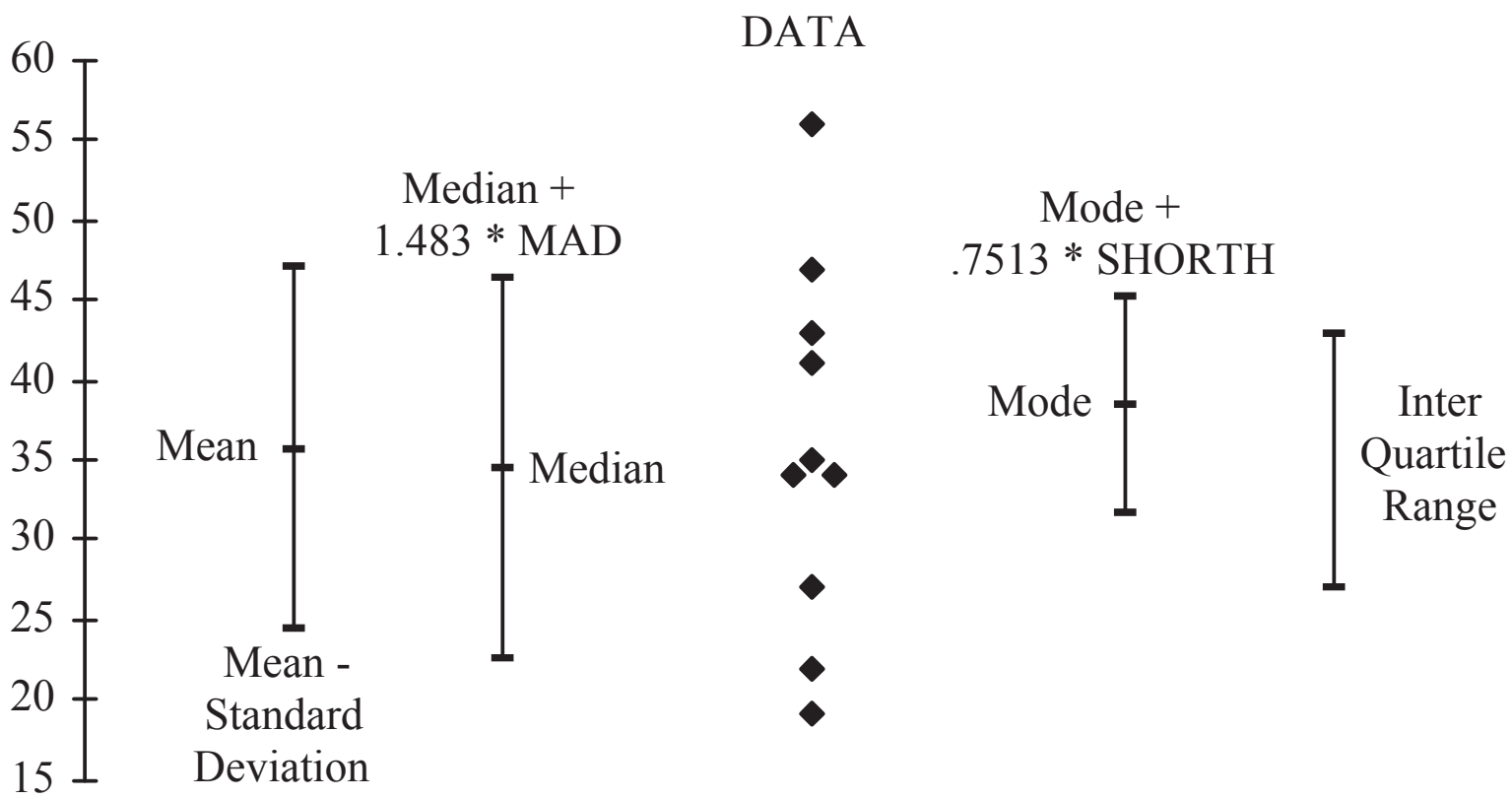
Table 1:

19	**8**	**12**	**10**	**42**
1	9	21	17	57
15	20	14	11	14
3	10	18	24	10
21	17	13	23	38

DATA POINT	SPLIT	STEM	LEAF
19	**1|9**	**1**	**9**

STEM	LEAVES	FREQUENCY
0	8	
1	0 2.....	
1	9.....	
2	.....	
2	.....	
3	.....	
3	.....	
4	2.....	
4	.....	
5	.....	
5	.....	

Stem and Leaf Plot for Table 1

STEM	LEAVES	FREQUENCY
0	8 9.....	2
1	0 0 0 1 1 2 3 3 4 4...	10
1	5 7 7 8 9.....	5
2	0 1 1 3 4.....	5
2	.....	0
3	.....	0
3	8.....	1
4	2.....	1
4	.....	0
5	.....	0
5	7.....	1

Summarization