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

PROBABILITY AND STATISTCS

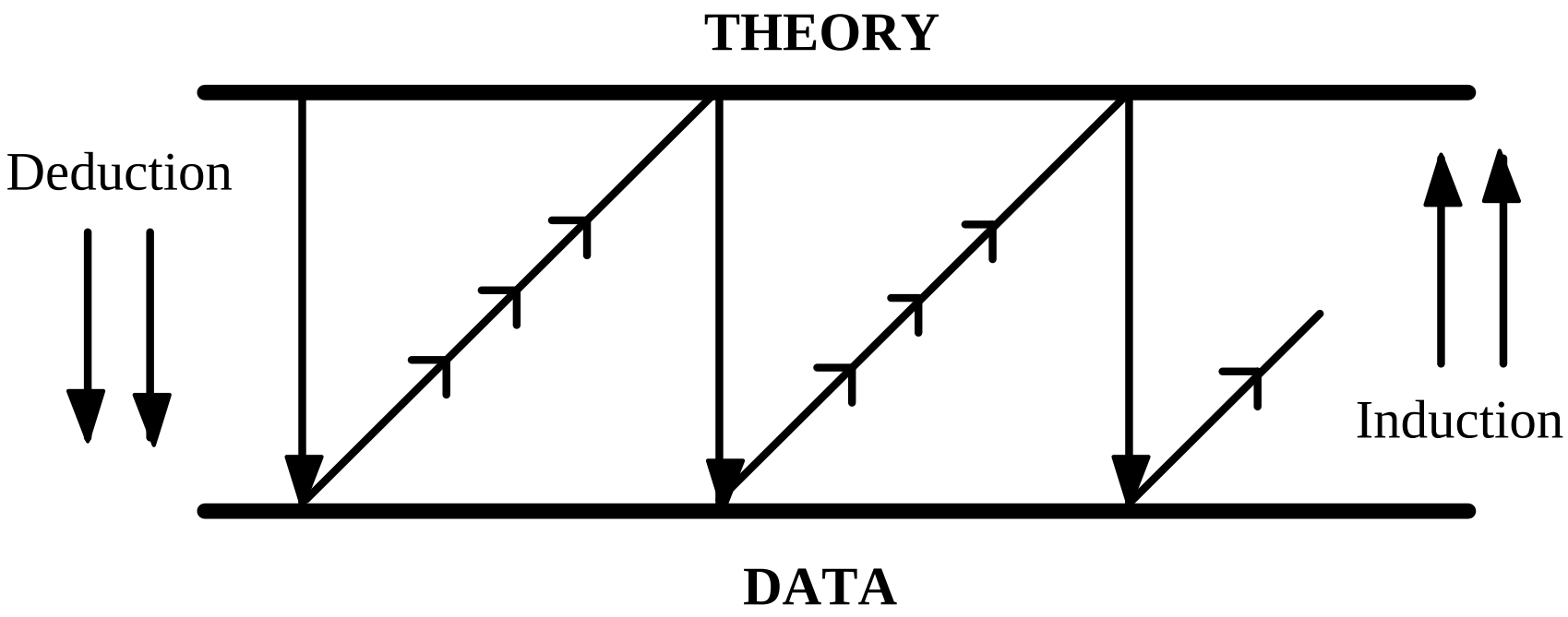
prof. Ing. Radim Briš, CSc.
www.am.vsb.cz/bris/

Teaching
Statistics 1.zip

Statistics – Science which Involves Study of Data

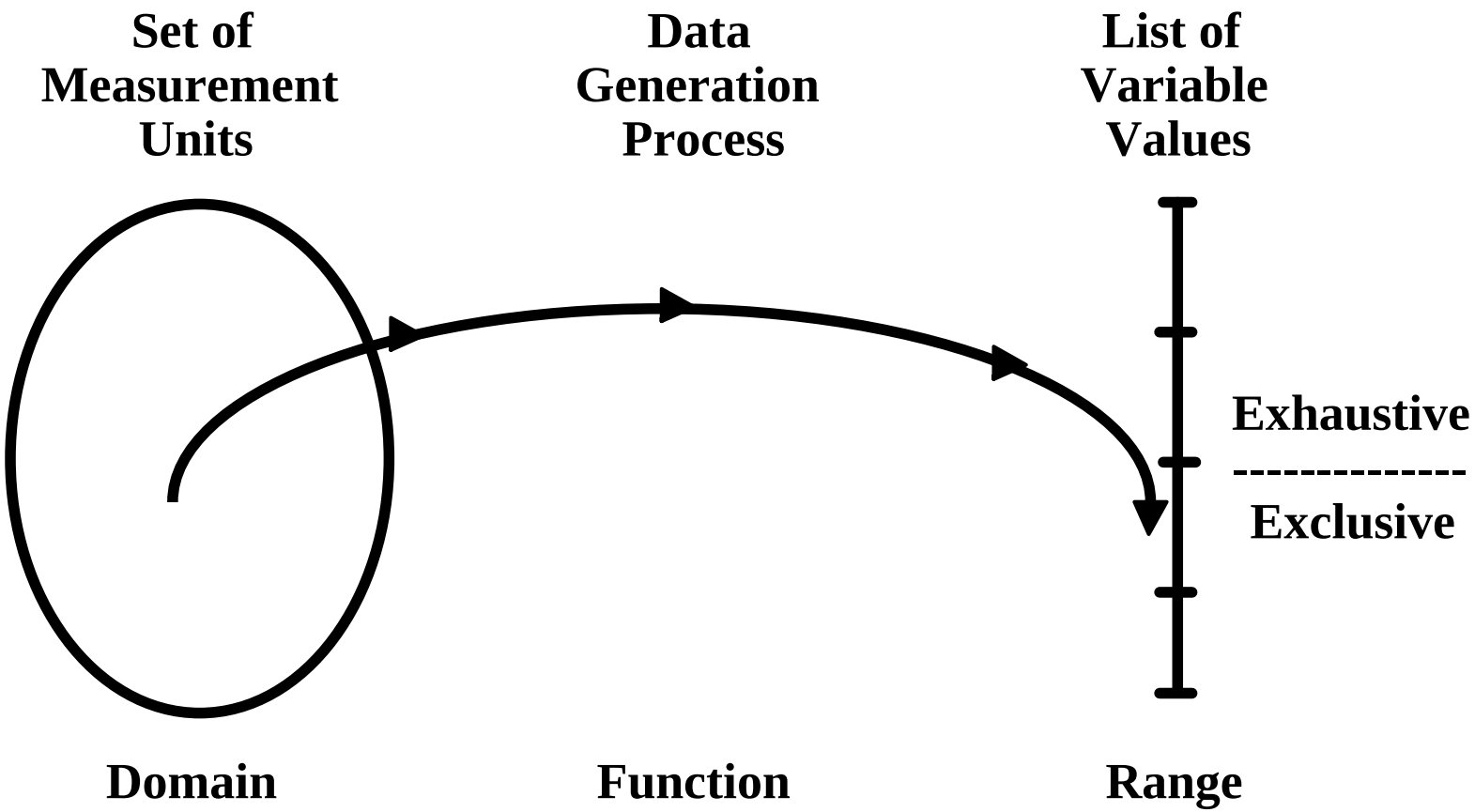
- The Objective of the Course
- A Definition of Statistics
- The Dual Role of Statistical Science

Deduction and induction

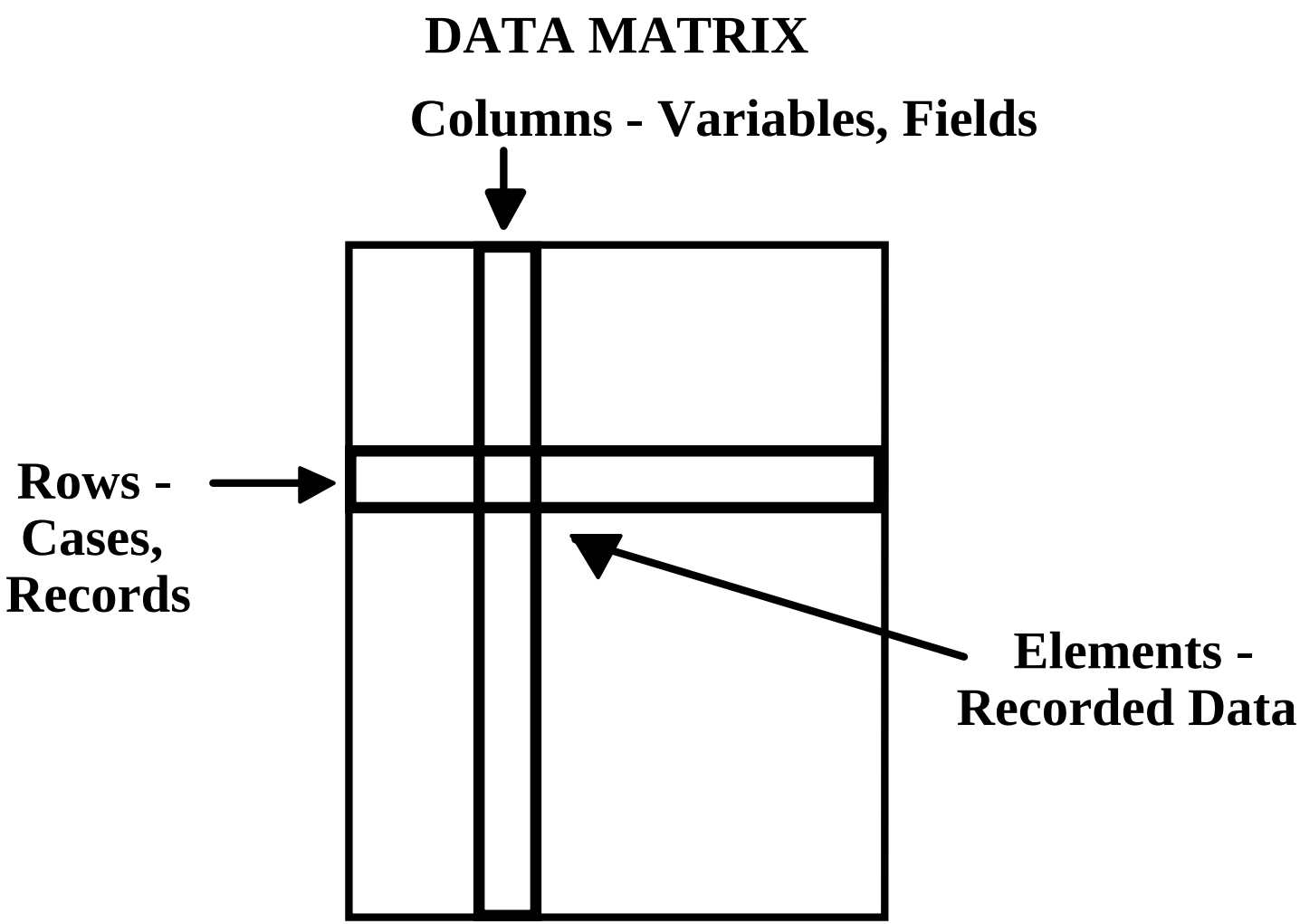


Data Generation Process

DATA DEFINITION AND CREATION



Data matrix

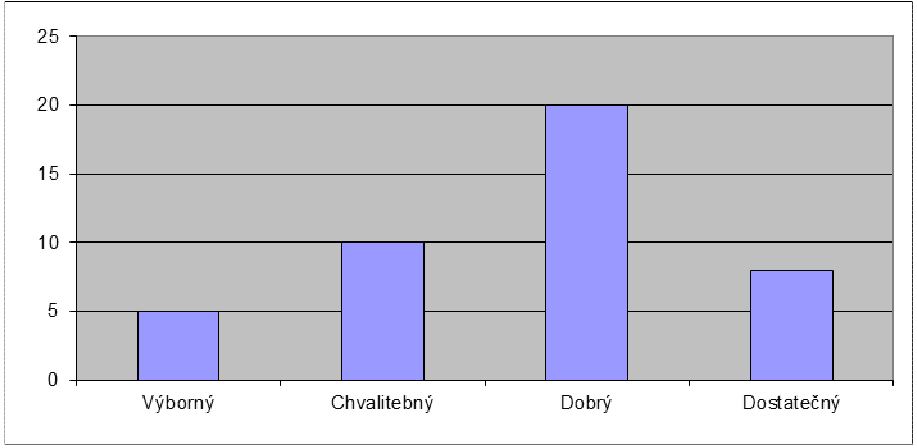

DATA MATRIX
Columns - Variables, Fields
Rows - Cases, Records
Elements - Recorded Data

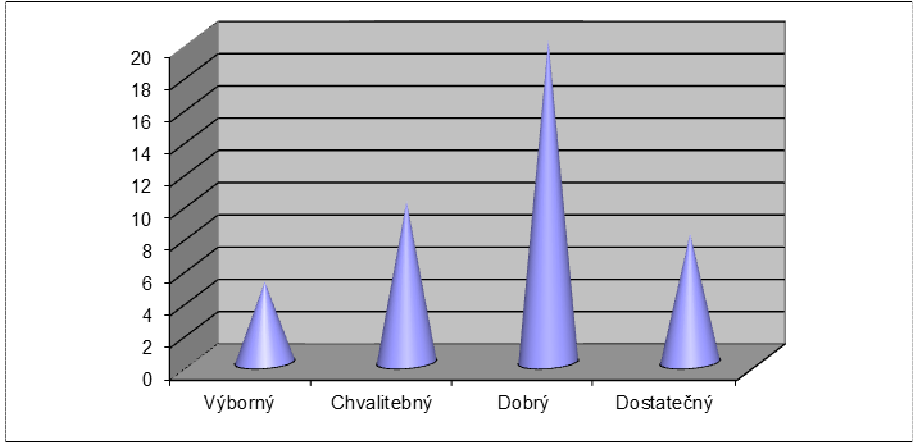
Graphical presentation of qualitative variables

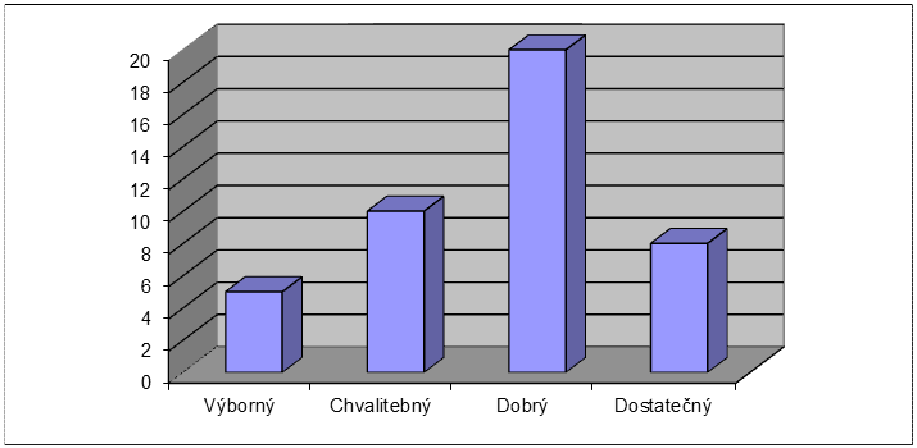
The statistics often use **graphs** for better plasticity of variables analysis. They are these two types for nominal variable:

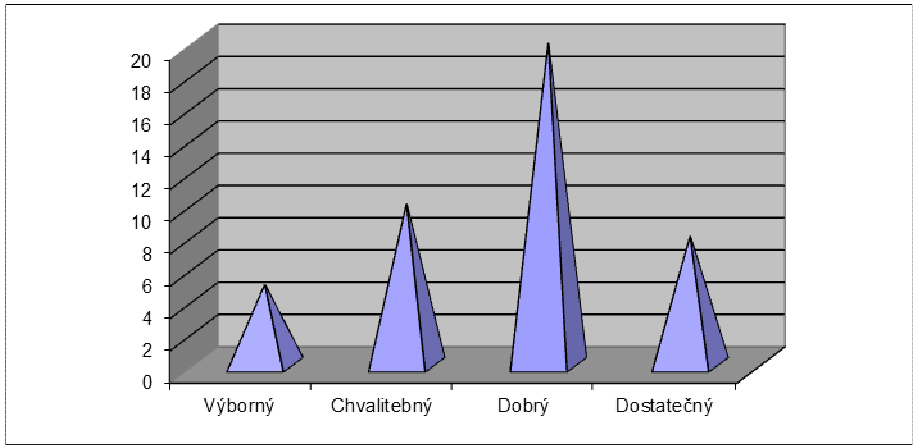
- **Histogram** (bar chart)
- **Pie chart**

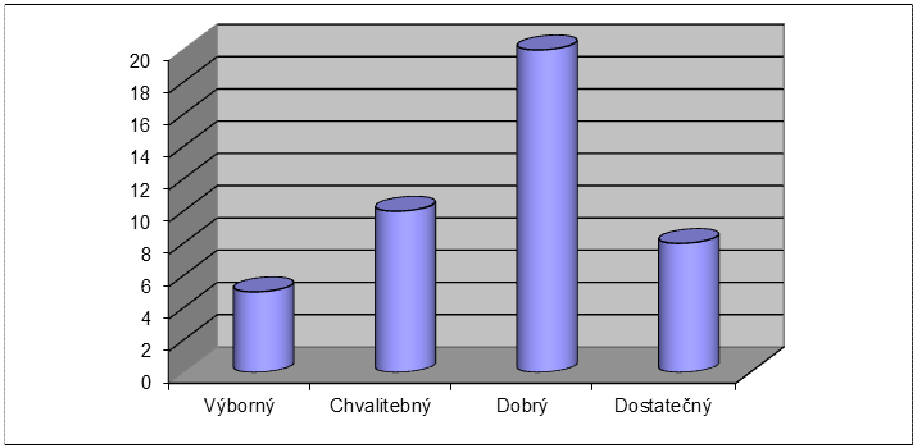
Histogram is a classical graph whereof we take variants of the variable on one axis and variable frequencies on the second one. Individual values of the frequency are then displayed as bars (boxes or vectors, squared logs, cones ...)

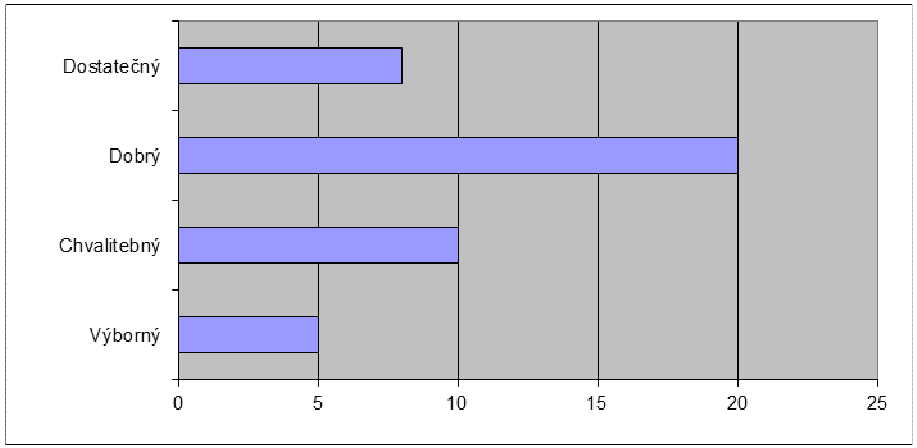

25
20
15
10
5
0
Výborný
Chvalitebný
Dobrý
Dostatečný

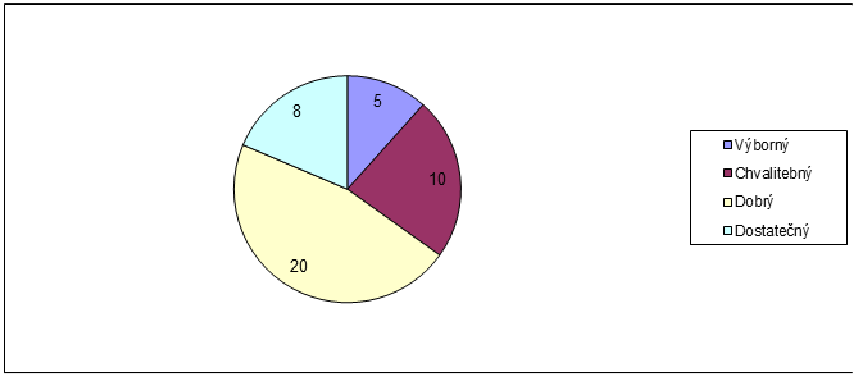

20
18
16
14
12
10
8
6
4
2
0
Výborný
Chvalitebný
Dobrý
Dostatečný

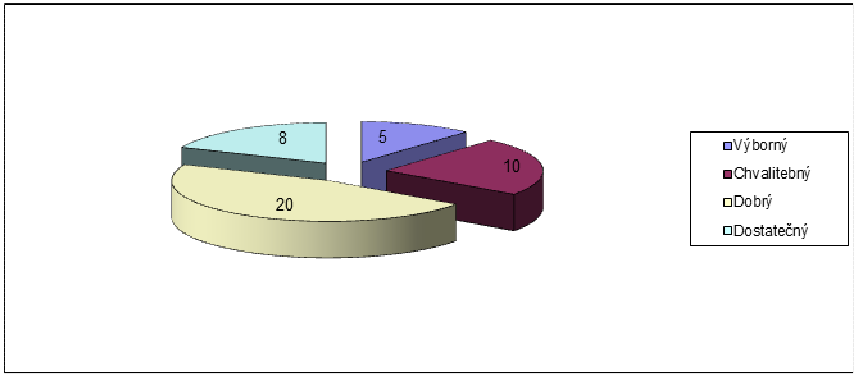

20
18
16
14
12
10
8
6
4
2
0
Výborný
Chvalitebný
Dobrý
Dostatečný

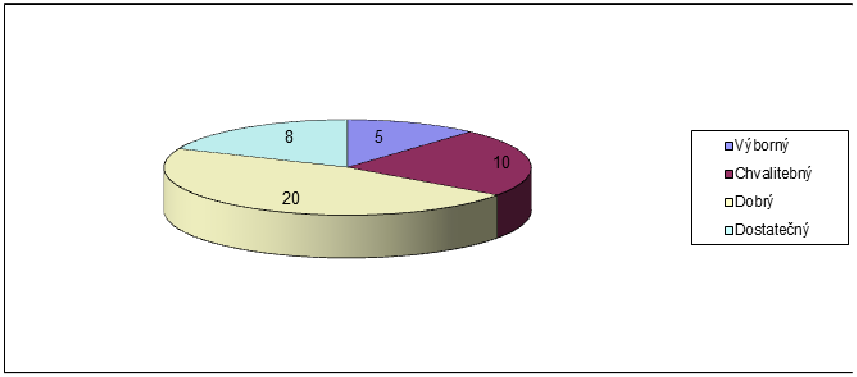

20
18
16
14
12
10
8
6
4
2
0
Výborný
Chvalitebný
Dobrý
Dostatečný

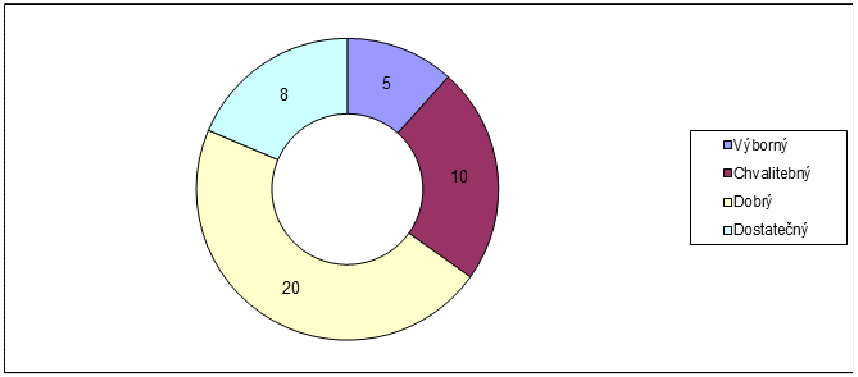

20
18
16
14
12
10
8
6
4
2
0
Výborný
Chvalitebný
Dobrý
Dostatečný


Dostatečný
Dobrý
Chvalitebný
Výborný
0
5
10
15
20
25



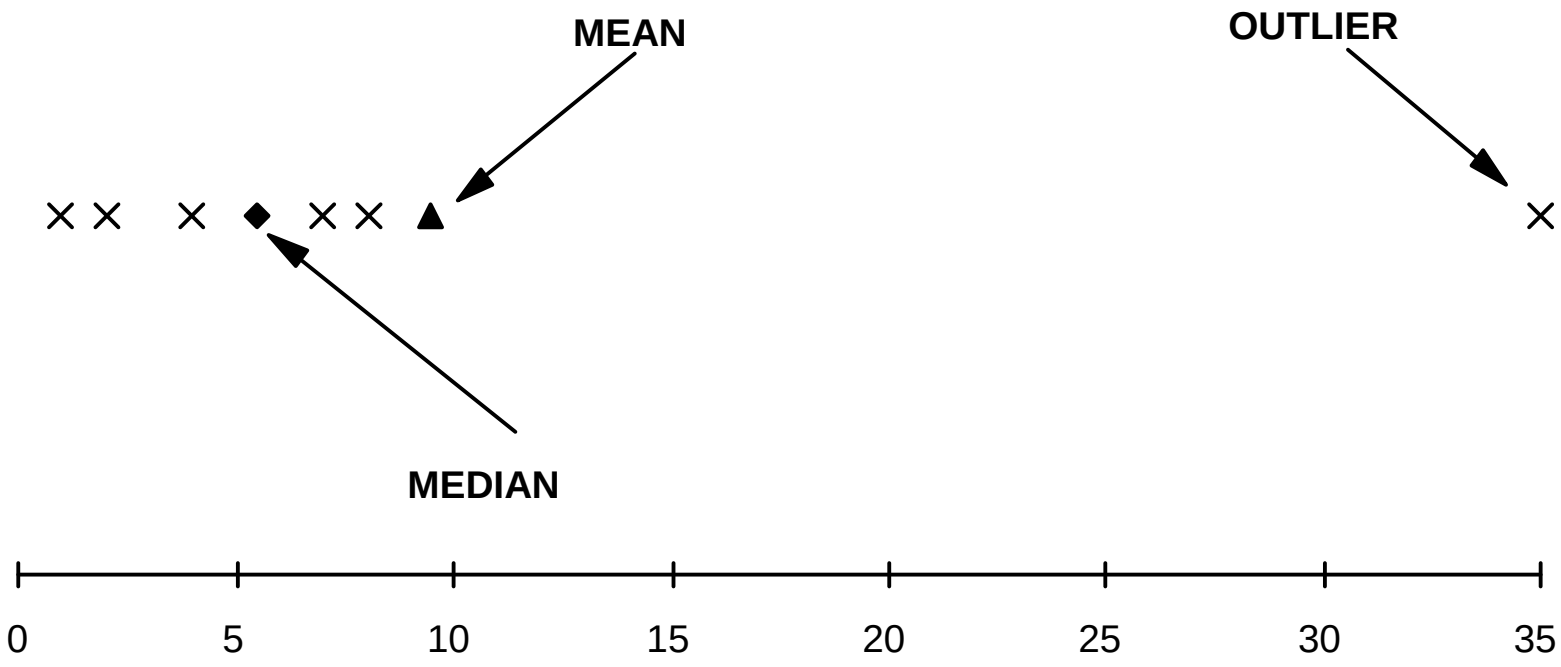






Pie graph represents relative frequencies of the individual variants of the variable. Individual relative frequencies are proportionally represent as a sector of a circle (when we change a circle to an ellipse we obtain three-dimensional effect).

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:

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

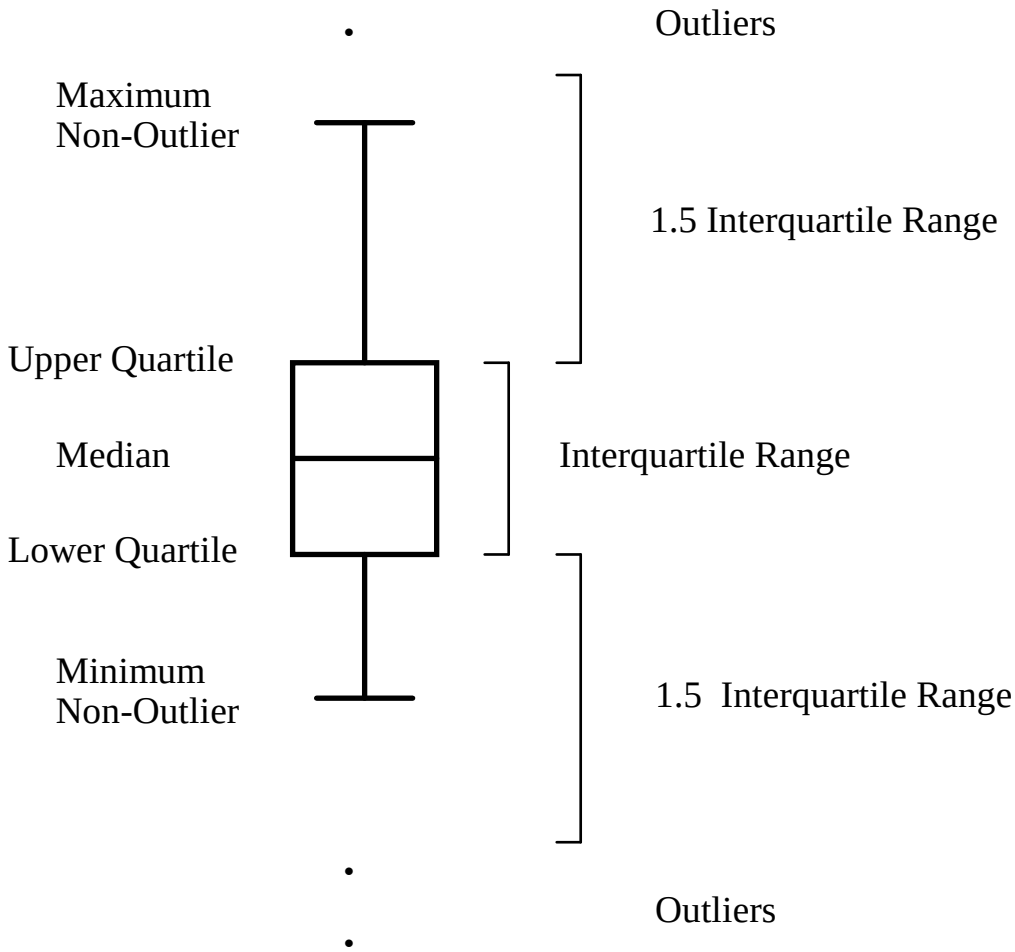
2. The outlier can be such value x_i where absolute value of median-axis is greater then 3:

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

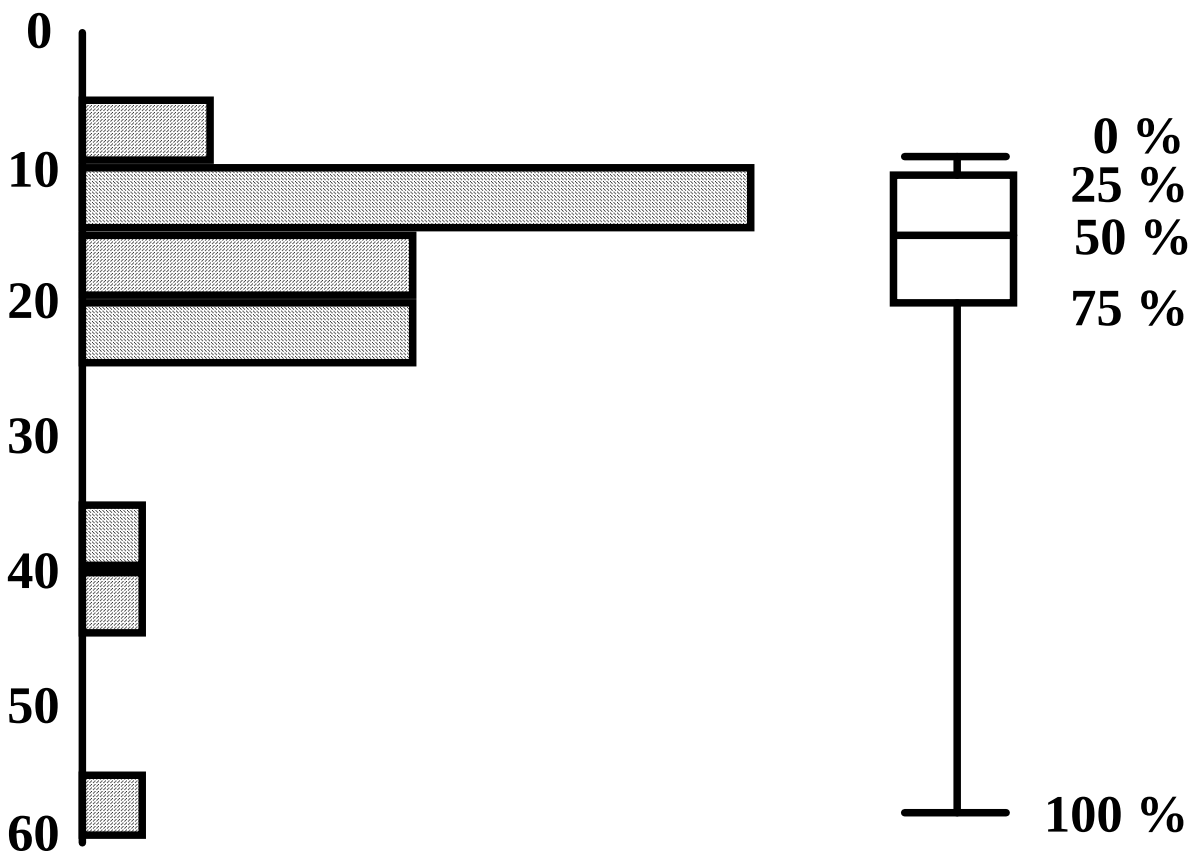
3. The outlier can be such value x_i that is far more then 1,5 IQR from lower (or upper) quantile.

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

Graphical presentation of quantitative variable – Box Plot



Graphical presentation of quantitative variable



Axiomatic probability definition

Probability space is a triad (Ω, S, P) where

(i) P is function from S to $< 0,1 >$ such that it holds:

a) $P(A) \geq 0$ for $A \in S$, … **P is non negative function**

b) $P(\Omega) = 1$ … probabilities are scaled to lie in the interval [0,1];

c) If

$$A \cap B = \emptyset, \text{ then}$$

$$P\{A \cup B\} = P\{A\} + P\{B\}$$

Function P is called **probability measure** or shorter **probability**.

Definition of a random variable

Let us consider a probability space (Ω, *S*, P). A **random variable X** (RV) on a sample space Ω is such real function *X(ω)* on Ω, which is chacterized by distribution function:

Distribution function

Definition: The distribution function of a random variable *X* is written *F(t)*

and, for each $t \in R$ has the value

$$F(t) = P\{\omega \in \Omega \mid X(\omega) < t)\} = P(X < t)$$

We call the *F(t)* as **distribution function** of *X*.

Distribution function is a function *X*: R → R

General properties of distribution function

1. $0 \leq F(x) \leq 1$

2. The distribution function is a **monotonic increasing** function of *x*, i.e.
$\forall x_1, x_2 \in R: x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$

3. The distribution function *F(x)* is **left-continuous**

4. $\lim_{x \to +\infty} F(x) = 1 \quad \lim_{x \to -\infty} F(x) = 0$

5. $\forall$ a, b $\in$ R; a < b must be: $P(a \leq X < b) = F(b) - F(a)$

Discrete random variable

Definition

We say that a random **variable *X* has a discrete probability distribution** when:

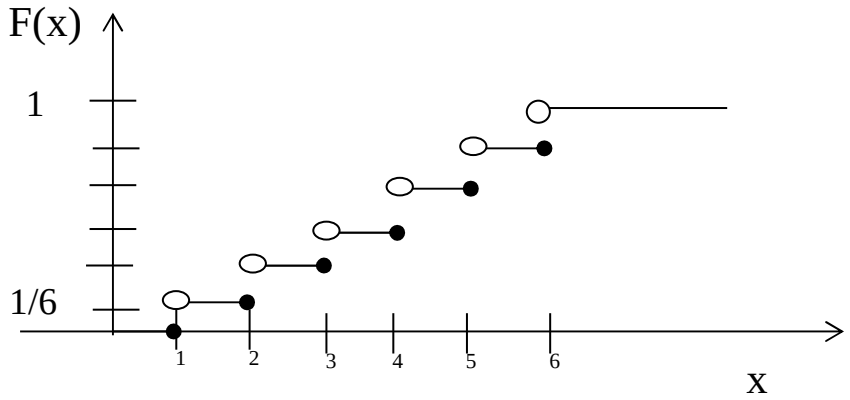
$\exists$ finite or enumerable set of real numbers M={ $x_1, \dots, x_n, \dots$} that

$P(X = x_i) > 0$ i = 1, 2, ...

$$\sum_i P(X = x_i) = 1$$

Function $P(X = x_i) \Leftrightarrow P(x_i)$ is called **probability function of random variable** *X*.

A distribution function of such distribution is a step function with steps in $x_1, \dots, x_n, ..$



Continuous random variable

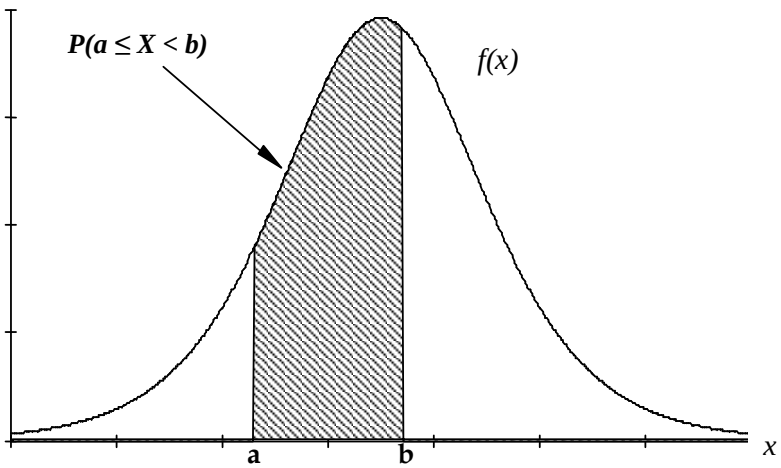
Definition

Random variable has a **continuous probability distribution** when a function *f(x)* exists that

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad -\infty < x < \infty$$

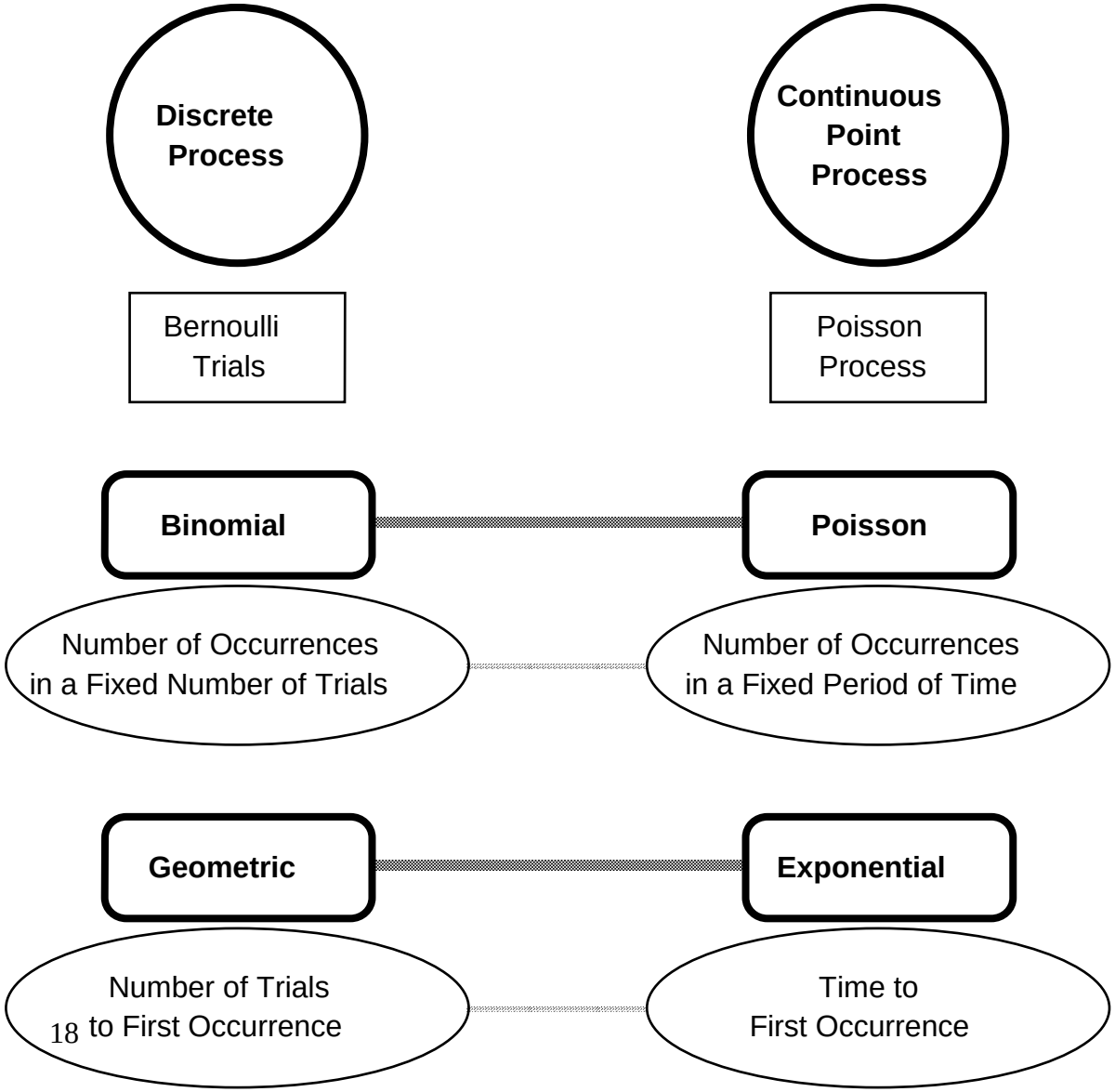
Function *f(x)* is called a **probability density function** of continuous random variable *X*. It is non-negative real function.

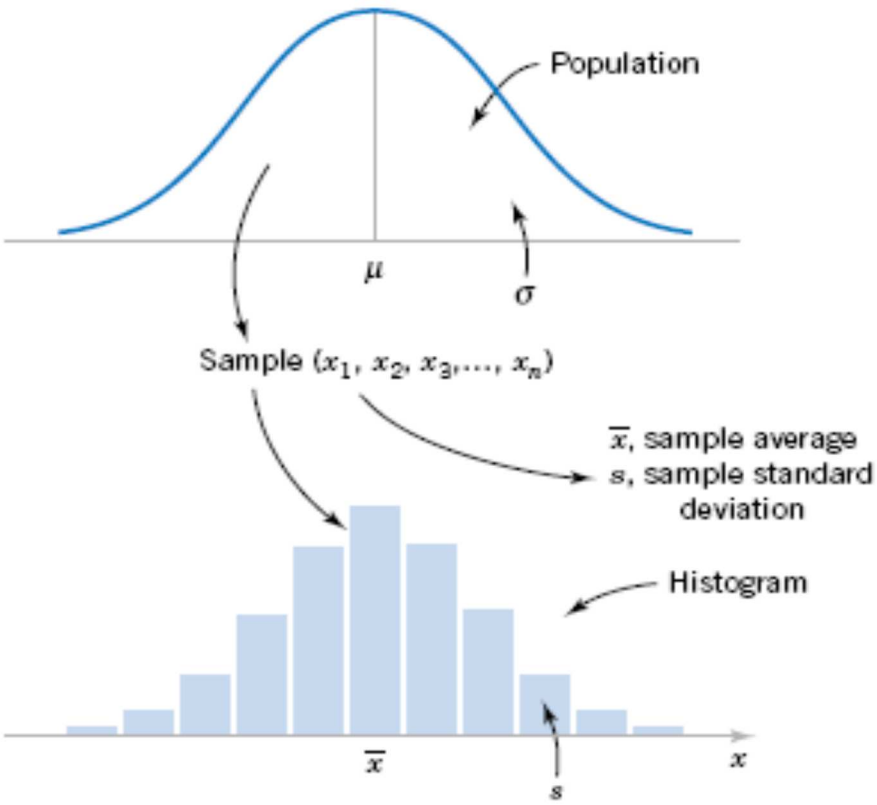
We can show that in all points where a derivation of distribution function exists it holds:



$$f(x) = \frac{dF(x)}{dx}$$

Relations between distributions





Definition: A **population** consists of the totality of the observations with which we are concerned.

Definition: A **sample** is a subset of observations selected from a population.

Random Sample $\underline{X}$ is special random vector: $\underline{X} = (X_1, \ldots, X_n)'$

(a) the X_i's are independent random variables, and

(b) every X_i has the same probability distribution.

Sampling distributions

Let's assume that given random sample comes from normal distribution:

$\underline{X}=(X_1, \dots , X_n)'$, $X_i \to N(\mu,\sigma^2)$

1. $\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n} \to N\left(\mu, \frac{\sigma^2}{n}\right)$ … comes from CLTfor large number n

2. $Z_n = \frac{\overline{X_n} - \mu}{\sigma} \cdot \sqrt{n} \to N(0,1)$ … comes from a transformation of previous distrib.

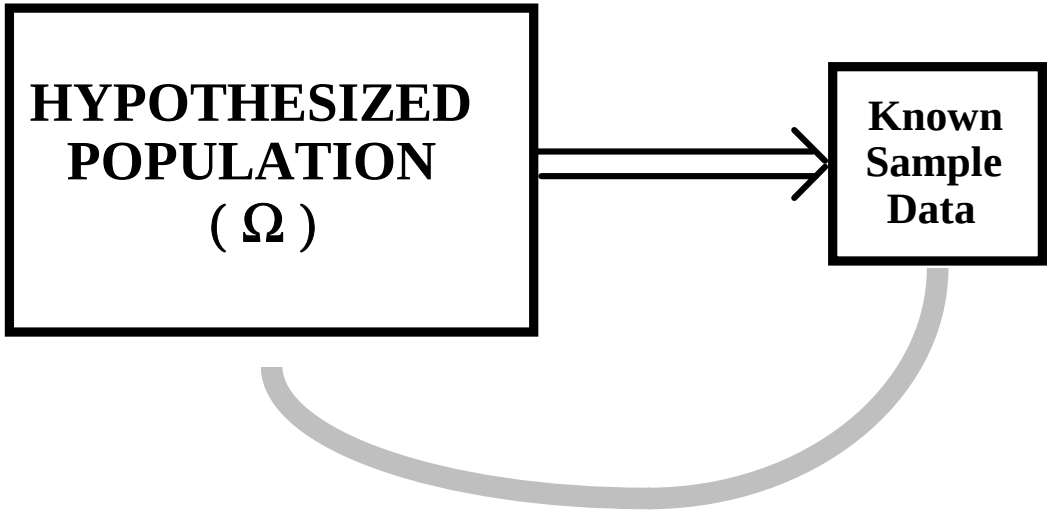
3. $\frac{S_n^2}{\sigma^2} \cdot (n-1) \to \chi^2(n-1)$ … was explained in context with χ^2discussion:

$$S_n^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2 \quad ; \frac{S_n^2}{\sigma^2} \cdot (n-1) = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2$$

The Method of Hypothesis Testing

Pure Significance Tests

The pure significance test asks whether the sample result is extreme with respect to some hypothesized distribution.



Can this data reasonably be assumed to have been generated by ? Ω

Is this data consistent with Ω?

The pure significance test consists of the following components:

1. **Null Hypothesis**: H_0

– The null hypothesis expresses some belief about the nature of the population. It must be specified precisely enough to define a probability measure on the population.

2. **Sample Statistic**: $T(\underline{X})$

– The sample statistic is a function of the sample data drawn from the population. The choice of sample statistic is determined by the characteristics of the population's probability distribution with which the null hypothesis is concerned.

3. **Null Distribution**: $F_0(x)$

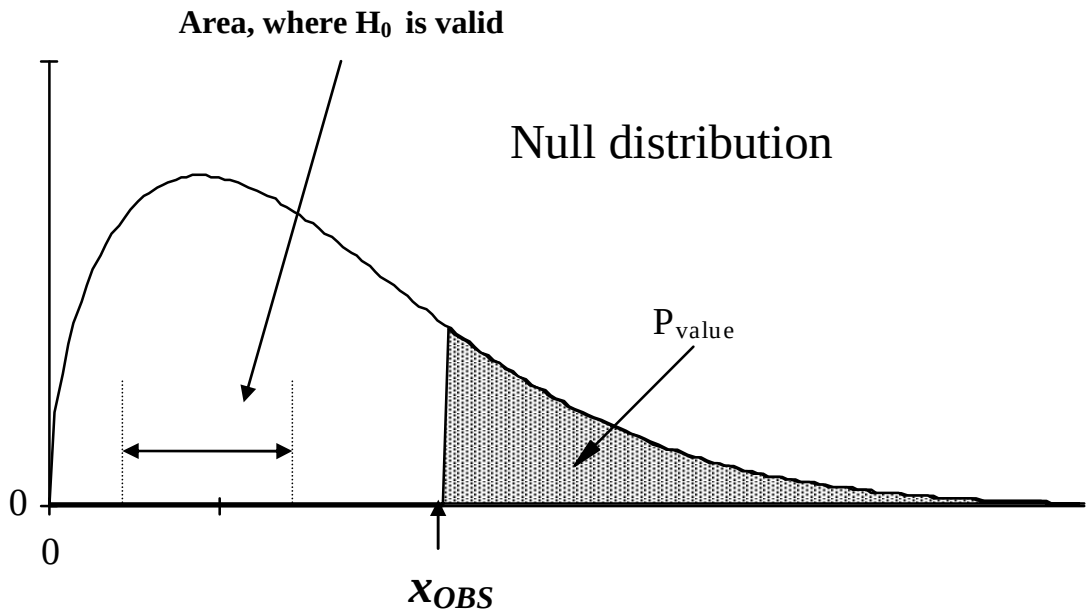
$$F_0(x) = P(T(\underline{X}) < x \mid H_0)$$

- The null distribution is the probability distribution of the sample statistic when the null hypothesis is correct. The null hypothesis must be specified precisely enough to determine the null distribution.

The pure significance test - continuation

4. To determine whether the observed sample statistic $t=x_{OBS}$ is extreme with respect to the null distribution, a statistic known as the ***p-value*** is computed. The *p-value* has 3 definitions depending on the context of the null hypothesis, but in all cases, the interpretation of the *p-value* is the same.

$$P_{VALUE} = 1-F_0(x_{obs}),$$



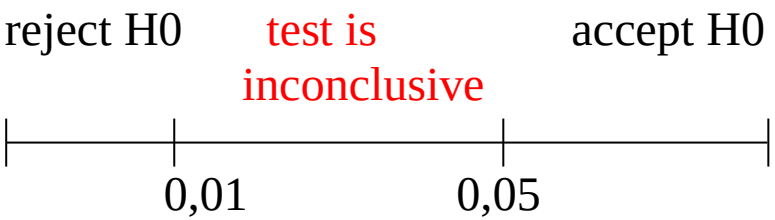
The pure significance test - continuation

5. Decision making based on P_{VALUE}

$P_{VALUE} < 0{,}01$ reject H0

$0{,}01 < P_{VALUE} < 0{,}05$ test is inconclusive

$P_{VALUE} > 0{,}05$ accept H0 (do not reject H0)



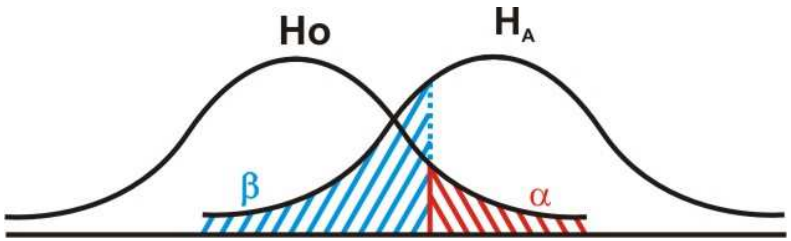
		True hypothesis	
		H_0	H_A
Decision	H_0	OK	Type II Error (Error II)
	H_A	Type I Error (Error I)	OK

Error I -probability

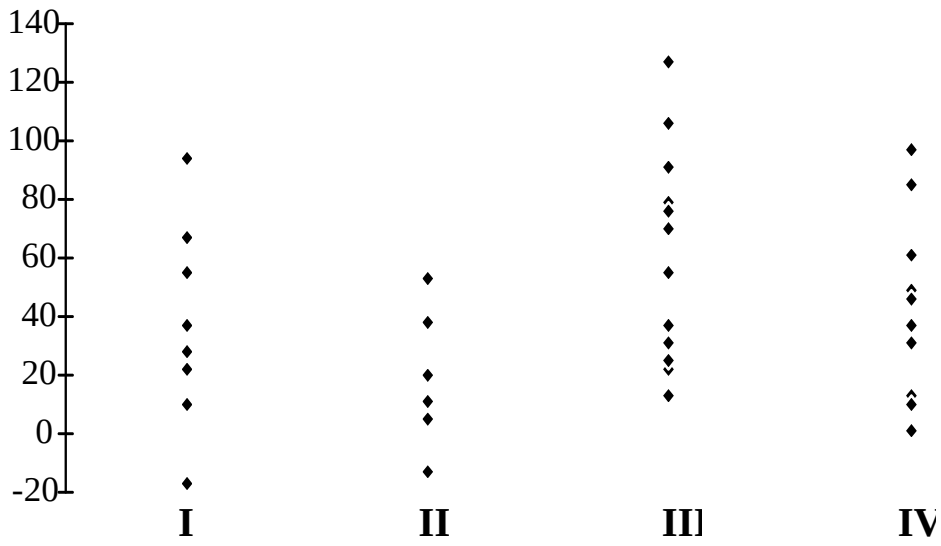
$$P(Error\ I) = P(P_{VALUE} < \alpha \mid H_0) = \alpha$$

Error II - probability

$$P(Error\ II) = P(P_{VALUE} > \alpha \mid H_A) = \beta$$



We faced the problem of testing hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_k = \mu$?



***Formulation of the problem*:**

The hypothesis of interest is H_0: $\mu_1 = \mu_2 = \ldots = \mu_k = \mu$

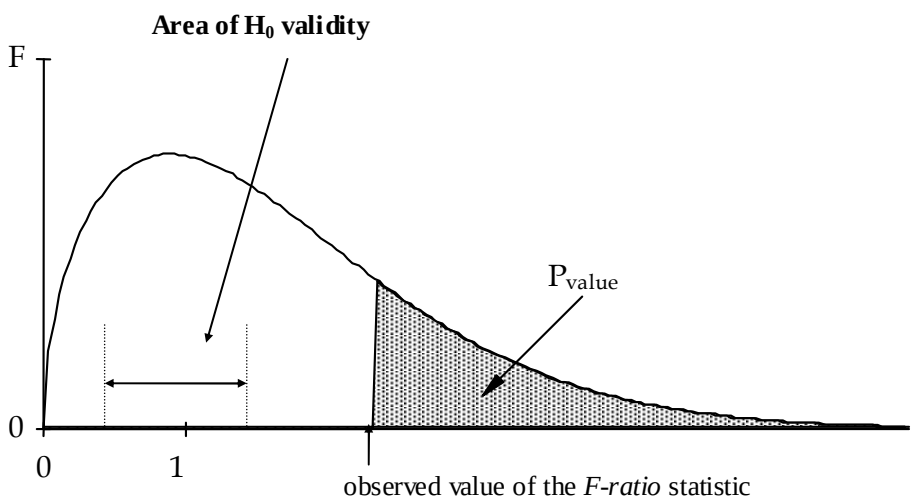
The alternate hypothesis is H_A: At least two μ_i's are different.

F-ratio

$$F = \frac{SS_B / k-1}{SS_W / N-k} = \frac{S_B^2}{S_W^2}$$

Why is useful use *F*-ratio as the test statistic?

i) If H_0 is true then *F-ratio* is a random number close to 1 ... $F \approx 1$.

ii) If H_0 is faslse then this number is markedly bigger than 1, see properties.

The statistic *F-ratio* is sensitive to validity of the hypothesis H_0.