# Computer Science and Mathematics. Part Statistics

1. Types and levels of variables
2. Descriptive statistics:
    - Frequency Distribution: Histograms
    - Quantiles
    - Measures of location and dispersion
    - Box-plot
3. Confirmative statistics:
    - Population and sample
    - Expected value
    - Probability
    - Normal distribution
    - Confidence intervals
    - Basics on statistical testing

## 4. Two-Variable Statistics

- **Scatterplots**
- **Correlation**
- **Regression**

Slides for the statistics part of the course base partially on the scripts and on the slides for the lecture course by Prof. Dr. Hans-Peter Piepho, „Statistik“ and „Biometrie“ for students in B.Sc. Agricultural Biology, Agricultural Sciences und Renewable Raw Materials at the University of Hohenheim, 2011

# Types of Variables

Variables used is statistical analyses can belong to different types of scale

## Qualitative or Categorical

- colour (white, pink)
- sex (male, female)
- school marks (in Germany : 1-6)
- health or disease classes

## Quantitative or Metric

- temperature ($^0$C)
- height of trees (m)
- number of plants (in a plot)
- crop yield (in dt/ha)

The statistic data analysis depends on the type of variable!

## Categorical data

- Finite number of categories
- Are categories rank-ordered?

  Yes $\Longrightarrow$ ordinal

  No $\Longrightarrow$ nominal

- Distances between categories are not quantifiable
- Calculating of means, sums or differences is impossible or problematic

## Metric data

Representable on the number line [is a picture of a straight line on which every point is assumed to correspond to a real number and every real number to a point]

**Example:**


**Continuous data:** Infinitely many different values between any two points on the real line

**Example:** Crop yield

**Discrete data**: Only certain fixed numerical values, intermediate values are not possible

**Example**: Number of plants on the plot

## Properties of metric data

- Distances are quantifiable
- Calculating of means, sums or differences makes sense

**Interval scale**

The "zero point" on an interval scale is arbitrary

Negative values can be used

Ratios between numbers on the scale are not meaningful

**Example**: Celsius scale

**Ratio scale:**

Possesses a zero value

**Example**: Kelvin scale

## Information content

metric > ordinal > nominal

 Losing of information

**Transformation of the scale level:**

metric $\Rightarrow$ ordinal $\Rightarrow$ nominal

nominal $\not\Rightarrow$ ordinal $\not\Rightarrow$ metric

## Descriptive Statistics: Description of data sample

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of every quantitative analysis of data.

Univariate analysis involves the examination across cases of one variable at a time. There are three major characteristics of a single variable that we tend to look at:

- the distribution
- the location (mean, median, and mode)
- the dispersion (range and quantiles of the data-set, and measures of spread such as the variance and standard deviation)

## Frequency Distribution: Discrete data (categorical & metric)

**Bar diagram** is a graphic representation of the frequency distribution of **discrete** data (Data that can only take certain values).

It is a chart with rectangular bars with lengths proportional to the values that they represent.

**Example (J.Saborowski. Biometric Data Analysis and Experiment Planning):**

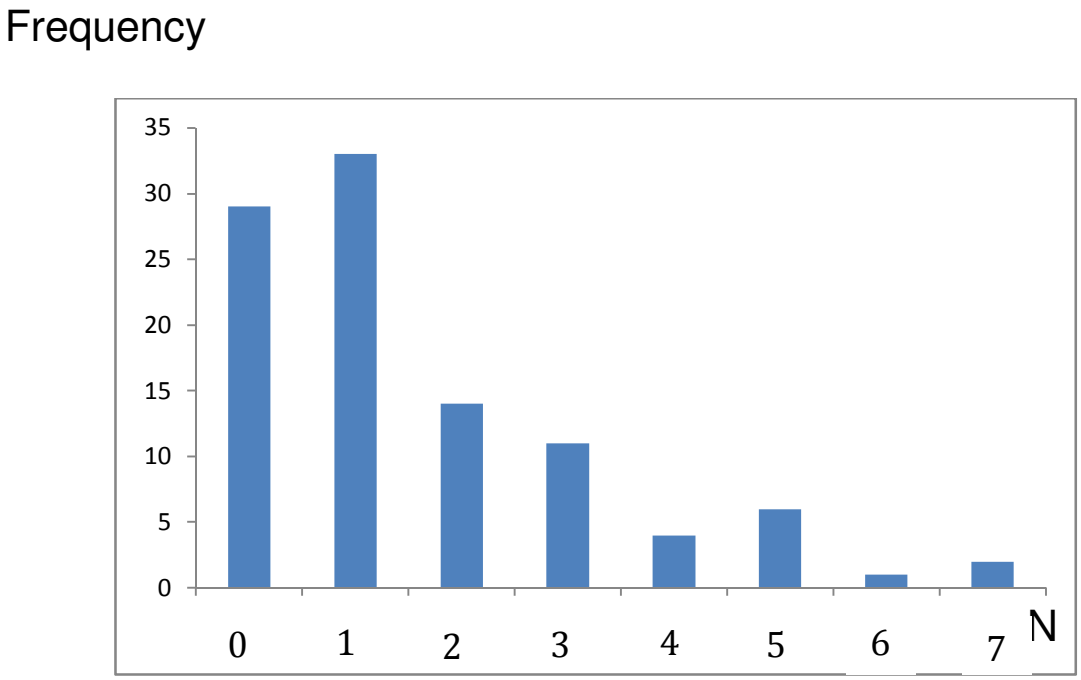
$n = 100$ plots, each $1m^2$, the number of plants is counted:

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 3 | 1 | 5 | 1 | 2 | 2 | 0 | 1 | 2 | 5 | 2 | 1 | 0 | 1 | 0 | 0 | 4 | 0 | 1 | 1 | 3 | 0 |
| 1 | 1 | 1 | 3 | 1 | 0 | 1 | 4 | 2 | 0 | 3 | 1 | 1 | 7 | 0 | 0 | 2 | 1 | 3 | 0 | 0 | 0 | 0 | 6 | 1 |
| 1 | 2 | 1 | 0 | 1 | 0 | 3 | 0 | 1 | 3 | 5 | 3 | 2 | 1 | 0 | 2 | 4 | 0 | 1 | 1 | 3 | 0 | 1 | 2 | 1 |
| 1 | 1 | 1 | 2 | 2 | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 5 | 0 | 5 | 1 | 2 | 2 | 7 | 4 | 1 | 3 | 1 | 5 | 0 |

## Check list of frequencies

| Plants | plot counts | number of plants/ absolute frequency | Percent/ relative frequency |
|---|---|---|---|
| 0 | 𝍸 𝍸 𝍸 𝍸 𝍸 \|\|\|\| | 29 | 29 |
| 1 | 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 \|\|\| | 33 | 33 |
| 2 | 𝍸 𝍸 \|\|\|\| | 14 | 14 |
| 3 | 𝍸 𝍸 \| | 11 | 11 |
| 4 | \|\|\|\| | 4 | 4 |
| 5 | 𝍸 \| | 6 | 6 |
| 6 | \| | 1 | 1 |
| 7 | \|\| | 2 | 2 |

# Bar diagram



## Frequency Distributions: Continuous data

**Example:**

- Corn field with 100.000 plants
- Number of plants $n = 50$
- Plants height measured in cm ($x$)

175 172 179 167 163 154 163 164 157 177

186 165 175 187 176 162 166 169 170 181

168 166 180 156 181 170 149 188 170 168

170 150 174 157 182 188 157 165 166 168

179 179 156 161 178 162 180 166 182 176

Forming of classes:

Number of classes $k$: rules of thumb

$k \geq (2n)^{1/3}$ (Terrel & Scott, 1985)

or

$k = 1 + 3.32 log_{10}(n) \approx 1 + 1.44 \ln(n)$ (Sturge's formula)

In our case:

Terrel & Scott formula: $k \geq (2 \cdot 50)^{1/3} = 4.64$

Sturge's formula: $k = 1 + 3.32 log_{10}(n) = 1 + 3.32 log_{10}(50) = 6.64$

We choose $k = 5$

Width of classes:

$$b > \frac{x_{max} - x_{min}}{k} = \frac{188 - 149}{5} = 7.8$$

We choose $b = 10$

| height (cm) |
| --- |
| 145 – 154.9 |
| 155 – 164.9 |
| 165 – 174.9 |
| 175 – 184.9 |
| 185 – 194.9 |

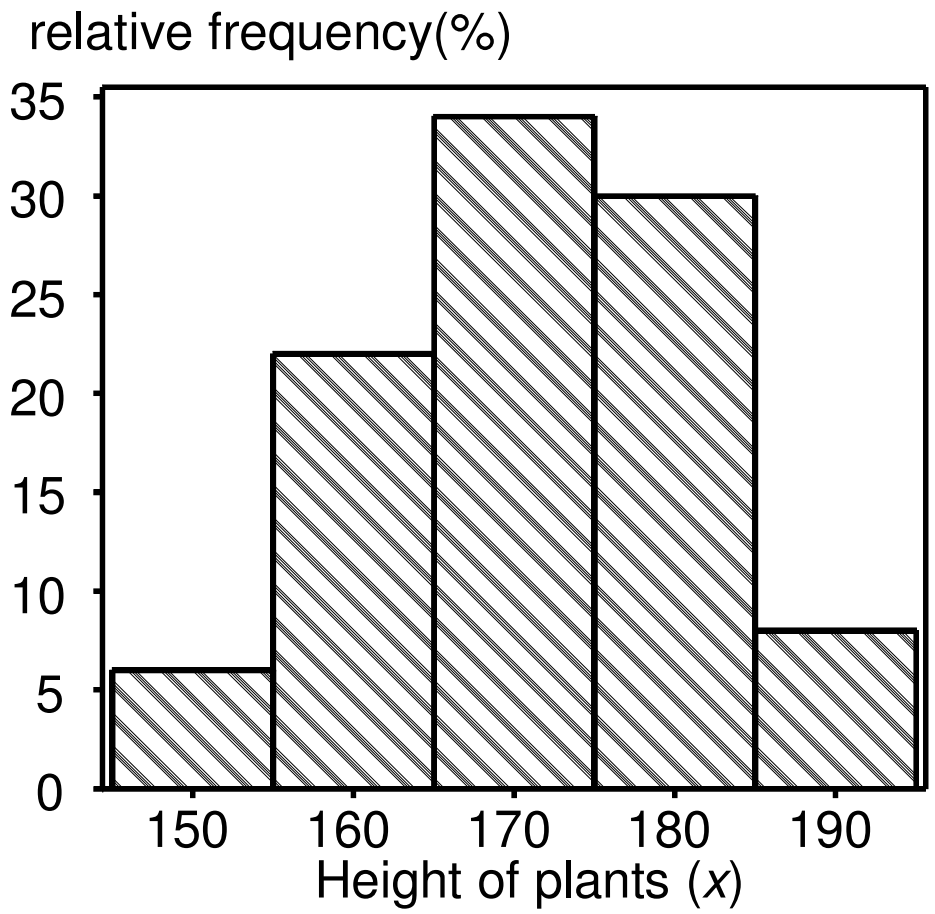
Precise mathematical notation of classes (without overlapping!):

$145 \leq X < 155$
$155 \leq X < 165$
$165 \leq X < 175$
$175 \leq X < 185$
$185 \leq X < 195$

## Check list of frequencies

| Height (cm) | plot counts | number of plants/ absolute frequency | Percent/ relative frequency |
|---|---|---|---|
| 145 – 154.9 | \|\|\| | 3 | 6 |
| 155 – 164.9 | 𝍸 𝍸 \| | 11 | 22 |
| 165 – 174.9 | 𝍸 𝍸 𝍸 \|\| | 17 | 34 |
| 175 – 184.9 | 𝍸 𝍸 𝍸 | 15 | 30 |
| 185 – 194.9 | \|\|\|\| | 4 | 8 |

Histogram


relative frequency(%)
35
30
25
20
15
10
5
0
150
160
170
180
190
Height of plants (x)

# Measures of Location and Dispersion

## Quantiles

The $t$% quantiles is such a value that at least $t$% of data values are less than or equal to $Q_t$

**Median** = $Q_{50}$= 50%-quantile

**Example**: Sample of $n = 16$ corn plants, height of plants $(x)$ was measured

175 172 179 167 163 154 163 164 157 177 186 165 175 194 176 162

Asending sequence of sorted value of the data set:

154 157 162 163 163 164 165 167 | 172 175 175 176 177 179 186 194

$Q_{50}$

$$Q_{50} = \frac{167 + 172}{2} = 169.5$$

25%- and 75%-quantiles:

154 157 162 163|163 164 165 167|172 175 175 176|177 179 186 194

$Q_{25}$ $Q_{50}$ $Q_{75}$

$$Q_{25} = \frac{163 + 163}{2} = 163$$

$$Q_{75} = \frac{176 + 177}{2} = 176.5$$

25% of values lie below $Q_{25} = 163$

75% of values lie below $Q_{75} = 176.5$

The general rule to compute quantiles:

$n$ – sample size

$t$ – $t\%$-quantile

$$p = \frac{t\%}{100\%}$$

The ordered sample:

$$x_{[1]} \leq x_{[2]} \leq x_{[3]} \leq \cdots \leq x_{[n]}$$

$x_{[j]}$- the „ $j$ -th order statistic", the $j$'th value of the ascending sequence of sorted values

Compute: $np = j + g$

Where $j$ – a integer part of the $np$

$g$ – a rest after subtraction $np - j$

The $t\%$ quantile is given as:

$Q_t = \frac{x_{[j]} + x_{[j+1]}}{2}$ if $g = 0$

$Q_t = x_{[j+1]}$ if $g > 0$

Notice:

$x_{[0]} = x_{[1]}$

and

$x_{[100]} = x_{[n]}$

**Example**: Sample of corn plants $n = 16$, height of plants $(x)$ was measured:

asending sequence of sorted values of the data set:

| 154 | 157 | 162 | 163 | 163 | 164 | 165 | 167 | 172 | 175 | 175 | 176 | 177 | 179 | 186 | 194 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_{[1]}$ | $x_{[2]}$ | $x_{[3]}$ | $x_{[4]}$ | $x_{[5]}$ | $x_{[6]}$ | $x_{[7]}$ | $x_{[8]}$ | $x_{[9]}$ | $x_{[10]}$ | $x_{[11]}$ | $x_{[12]}$ | $x_{[13]}$ | $x_{[14]}$ | $x_{[15]}$ | $x_{[16]}$ |

$Q_0 = x_{[1]} = 154$

$Q_{10}: n\frac{t}{100\%} = 16 \cdot 0.1 = 1.6; j = 1; g = 0.6 > 0;\ Q_{10} = x_{[2]}$

$Q_{25}: n\frac{t}{100\%} = 16 \cdot 0.25 = 4; j = 4; g = 0;\ Q_{25} = \frac{x_{[4]}+x_{[5]}}{2} = \frac{163+163}{2} = 163$

$Q_{50}: n\frac{t}{100\%} = 16 \cdot 0.5 = 8; j = 8; g = 0;\ Q_{50} = \frac{x_{[8]}+x_{[9]}}{2} = \frac{167+172}{2} = 169.5$

$Q_{70}: n\frac{t}{100\%} = 16 \cdot 0.7 = 11.2; j = 11; g = 0.2 > 0;\ Q_{70} = x_{[12]} = 176$

$Q_{100} = x_{[16]} = 194$

## Measures of location

**Median**= 50% quantile

A value left and right of which are 50% of all values of the data set

**Arithmetic mean:**

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

[pronounced as "$x$-bar"]

**Example**:

Height of corn plants:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{175 + 172 + 179 + \cdots + 194 + 176 + 162}{16} = 170.56$$

$(Median = 169.5)$

**Median**: more robust against outliers than the mean

**Example:** The same sample, in which the 1st value contains a typing error!

**1**175 172 179 167 163 154 163 164 157 177 186 165 175 194 176 162

$Q_{50} = 169.5$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1175 + 172 + 179 + \cdots + 194 + 176 + 162}{16} = 233.0625$$

**Mode:** A value or class with highest relative frequency, mainly used for discrete variables

## Measures of Dispersion

**Example:**

- On-farm trial with sorghum in Africa,
- 28 farms,
- 3 fertilisation systems:

  Without fertiliser (control)

  NPK (Nitrogen-Phosphor-Potassium)

  DAP (Di-Ammon-Phosphate)

- Crop yield in dt/ha

**Data**

| Farm | Control | NPK | DAP |
|---|---|---|---|
| 1 | 0.30 | 0.80 | 1.64 |
| 2 | 0.34 | 1.12 | 1.38 |
| 3 | 0.39 | 1.12 | 1.70 |
| 4 | 0.40 | 1.60 | 2.80 |
| 5 | 0.40 | 2.80 | 2.40 |
| … | … | … | … |
| 24 | 2.40 | 4.48 | 3.84 |
| 25 | 2.40 | 9.60 | 3.84 |
| 26 | 2.56 | 5.28 | 3.24 |
| 27 | 3.60 | 4.80 | 5.60 |
| 28 | 4.50 | 5.50 | 6.75 |

**Question**: Which system is the most stable concerning a crop yield?

## Range

$$V = Q_{100} - Q_0 = x_{max} - x_{min}$$

**Example** (Control)

$x_{max} = 4.50,\ x_{min} = 0.30,\ V = 4.50 - 0.30 = 4.20$

## Interquartile range

$$R_{IQ} = Q_{75} - Q_{25}$$

## **Example** (Control)

| $x_{[i]}$ | $[i]$ | | $x_{[i]}$ | $[i]$ |
|---|---|---|---|---|
| 0.30 | 1 | | 0.78 | 15 |
| 0.34 | 2 | | 0.82 | 16 |
| 0.39 | 3 | | 0.96 | 17 |
| 0.40 | 4 | | 1.02 | 18 |
| 0.40 | 5 | | 1.06 | 19 |
| 0.42 | 6 | | 1.10 | 20 |
| 0.48 | 7 | | 1.44 | 21 |
| 0.54 | 8 | | 1.60 | 22 |
| 0.56 | 9 | | 1.68 | 23 |
| 0.58 | 10 | | 2.40 | 24 |
| 0.62 | 11 | | 2.40 | 25 |
| 0.68 | 12 | | 2.56 | 26 |
| 0.74 | 13 | | 3.60 | 27 |
| 0.74 | 14 | | 4.50 | 28 |

$$Q_{75}:\ t = 75\%;\ p = 0.75;\ n = 28;\ np = 21;\ j = 21;\ g = 0;$$
$$Q_{75} = (x_{[21]} + x_{[22]})/2 = (1.44 + 1.60)/2 = 1.52$$
$$Q_{25}:\ t = 25\%;\ p = 0.25;\ n = 28;\ np = 7;\ j = 7;\ g = 0;$$
$$Q_{25} = (x_{[7]} + x_{[8]})/2 = (0.48 + 0.54)/2 = 0.51$$
$$R_{IQ} = 1.52 - 0.51 = 1.01$$

**Variance** is a measure of the 'spread' of a distribution about its average value.

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

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

$$\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i \cdot \bar{x} + n \cdot \bar{x}^2 =$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i \cdot \frac{\sum_{i=1}^{n} x_i}{n} + n\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^2} =$$

$$\sum_{i=1}^{n} x_i^2 - 2\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} + \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

Thus,

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

**Example** (Control)

$$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{68.2841 - \frac{(33.11)^2}{28}}{27} = 1.079$$

**Standard deviation** (the mean squared deviation of the $x_i$ from their mean)

$$s_x = \sqrt{s_x^2}$$

**Example** (Control)

$$s_x = \sqrt{1.079} = 1.039$$

**Coefficient of variation**

$$CV = \frac{s_x}{\bar{x}}$$

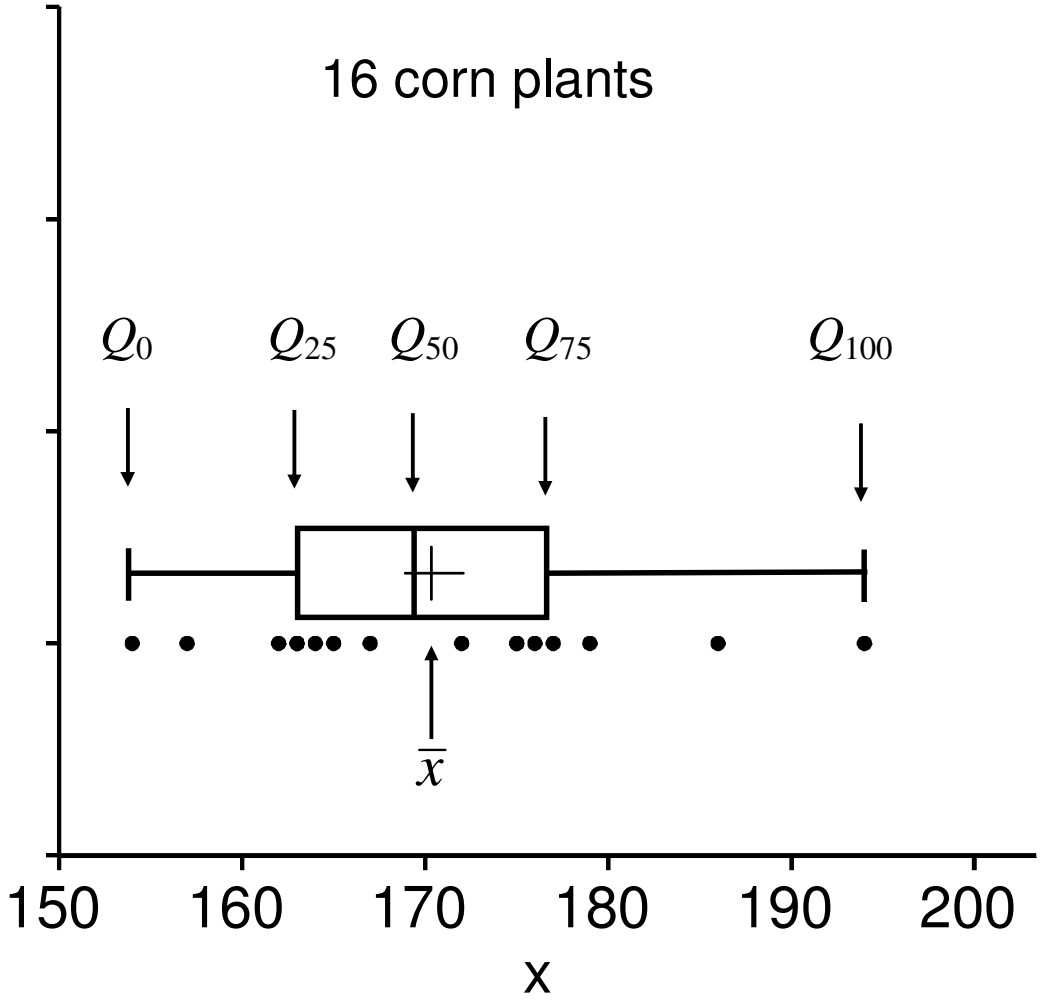
**Example** (Control)

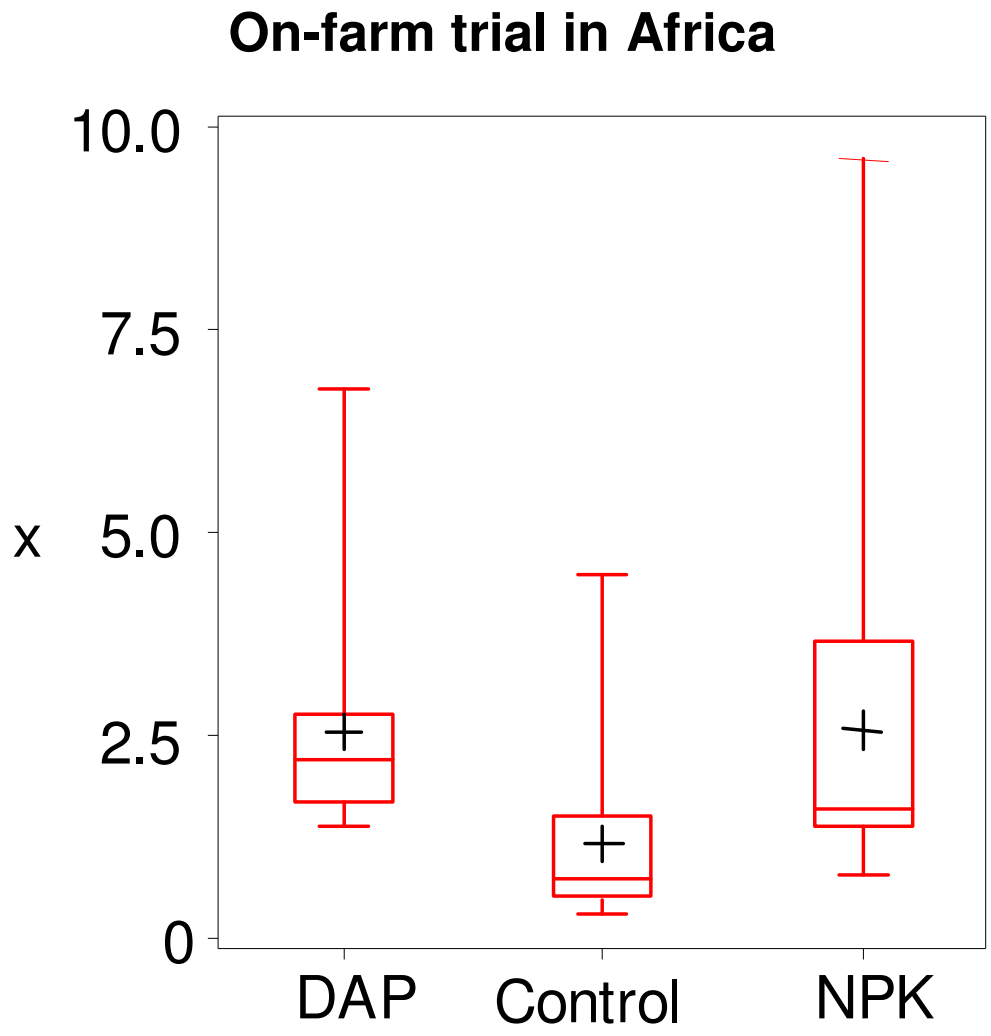
$$\bar{x} = 1.183;\ s_x = 1.039;\ CV = \frac{1.039}{1.183} = 0.8784 = 87.84\%$$

| Statistic Measures | Control | NPK | DAP |
|---|---|---|---|
| Range | 4.20 | 8.80 | 5.37 |
| Interquartile Range | 1.01 | 2.31 | 1.10 |
| Variance | 1.079 | 3.995 | 1.588 |
| Standard Deviation | 1.039 | 1.999 | 1.260 |
| Coefficient of Variance (%) | 87.84 | 77.62 | 49.47 |
| Mean | 1.183 | 2.575 | 2.548 |

DAP is the most stable system

**Box-and-Whiskers-Plots** or **Box-Plot:** Visualization of statistic measures




On-farm trial in Africa
10.0
7.5
x 5.0
2.5
0
DAP
Control
NPK

## Confirmative Statistics

**Main task:** Drawing conclusions about a population using information from a sample

**Statistical population:**
A population is any entire collection of individuals, items, or data from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about.

- Can be finite or infinite

**Examples:**

finite: All plants on the corn field
infinite: All possible results that can be reached under the same conditions of some experiment

- Typically, the population is very large. Making a complete enumeration of all the values in the population is impractical or impossible.

**Sample:**

- A subset of a population of manageable size, drawn out of the population

**Examples:**

50 plants from the corn field

4 plots with same system of fertilizer

- are collected so that one can make inferences or extrapolations from the sample to the population

## Same notations of statistical measures

| | **population** | **sample** |
|---|---|---|
| | Greek Letters | Latin Letters |
| mean | $\mu$ | $\bar{x}$ |
| variance | $\sigma^2$ | $s^2$ |

**Question**: How to describe the population using information about the sample?

**Answer**: We need **models** that define the relationships between population and sample.

Such models are delivered by the **probability theory**.

Model: Simplified description of reality

## Probability theory: fundamental concepts

**Random Variable $X$** [is always written by an uppercase letter]

A **random variable** or **stochastic variable** is a variable whose value is subject to variations due to chance (i.e. randomness, in a mathematical sense).

- does not have a single, fixed value
- can take on a set of possible different values, each with an associated **probability**.

There are two types of random variables - **discrete** and **continuous**.

A **discrete** random variable is one which may take on only a countable number of distinct values such as 0, 1, 2, 3, 4, ... Discrete random variables are usually (but not necessarily) counts. If a random variable can take only a finite number of distinct values, then it must be discrete.

**Examples:** The number of children in a family, number of plants on the field, number of  in a herd etc.

A **continuous** random variable is one which takes an infinite number of possible values. Continuous random variables are usually measurements.

**Examples:** Height, weight, crop yield etc.

**Realization** or **observed value** of the random Variable $x$ [is always written with lowercase letter]

- is the value that is actually observed (what actually happened).

The observed value of a random variable is an outcome of **random experiment.**

The concrete outcome of a random variable **cannot be predicted exactly**. It is only one of many possible realizations.

**Examples:**

**Dice throw**

Random Variable $Y = \{1,2,3,4,5,6\}$

Realization: outcome 3

**Chose the corn plant from the field with 100.000 plants and measure its height**

Random Variable $Y$ – height of the corn plant

Realization $y = 110$ cm

The random variable can be characterized as follows:

Possible outcomes of random experiments can be associated with **probabilities**.

Characterizing a random variable means defining probabilities for its different realizations, it means **probability distribution**.

**Probability distribution** is a function that describes the probability of a random variable taking certain values.

The probability distribution of a **discrete random** variable is a list of probabilities associated with each of its possible values.

More formally, the probability distribution of a discrete random variable $X$ is a function which gives the probability $p(x_i)$ that the random variable equals $x_i$, for each value $x_i$:

$$p(x_i) = P(X = x_i)$$

It satisfies the following conditions:

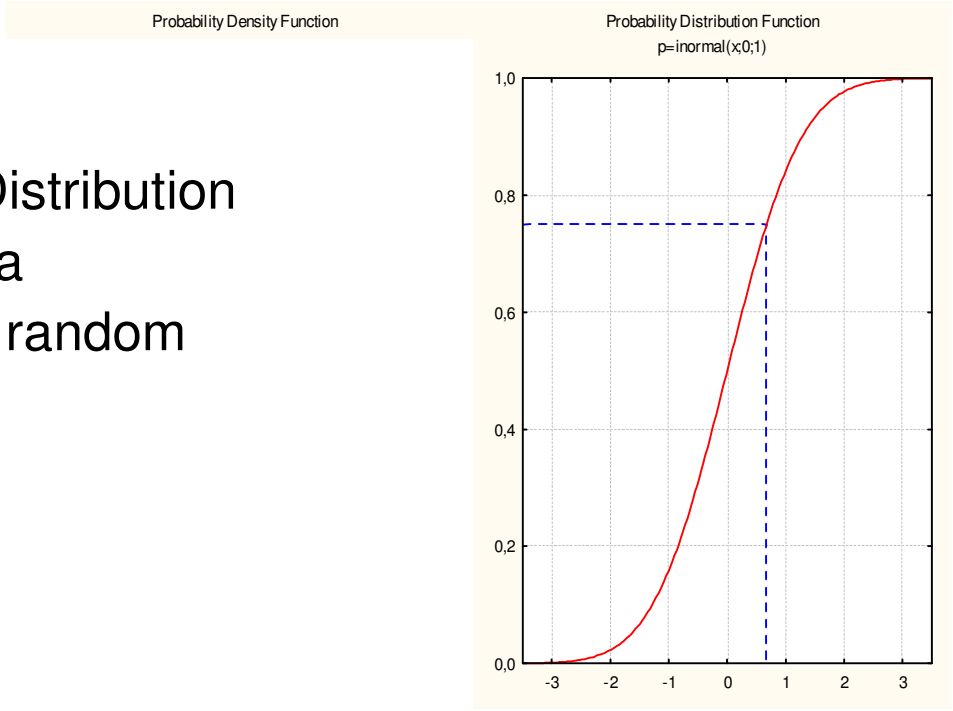
$$0 \leq p(x_i) \leq 1$$

$$\sum p(x_i) = 1$$

The Probability distribution of a random variable (discrete and continuous) can be described by a cumulative distribution function or just **distribution function** $F$. The distribution function $F$ shows, with which probabilities a random variable $X$ takes a value **less than or equal to** $x$:

$$F(x) = P(X \leq x)$$

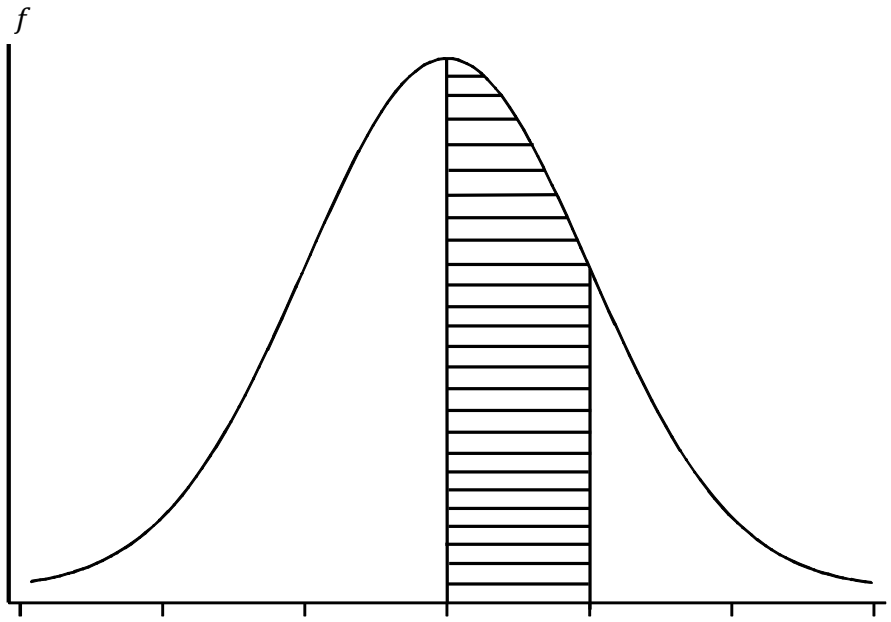
Example: Distribution function of a continuous random variable




$$P(X \leq 0{,}75) = 0{,}7734$$

**The Probability** distribution of a **continuous random** variable can also be described by the **Probability density function (*f*).**

The probability density function of a continuous random variable is a function which can be integrated to obtain the probability that the random variable takes a value in a given interval.




$$f(x) = F'(x)$$

$$P(X \in [a, b]) = \int_a^b f(x)dx = F(b) - F(a)$$

The most important continuous probability distribution is the so called **Normal distribution** or **Gaussian distribution** (named after Carl-Friedrich Gauss).

**Density function of Normal distribution:**

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

**Notation of the Normal distribution:** $N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

**The Normal distribution depends on two parameters**:

The **expected value** (or **expectation**, or **mathematical expectation**, or **mean**) of a random variable

$$E[X] = \mu$$

- Is the weighted average of all possible values that this random variable can take
- In case of a discrete random variable the weights correspond to the probabilities.
- In case of a continuous random variable the weights correspond to the density of the probability density function ($f$).

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

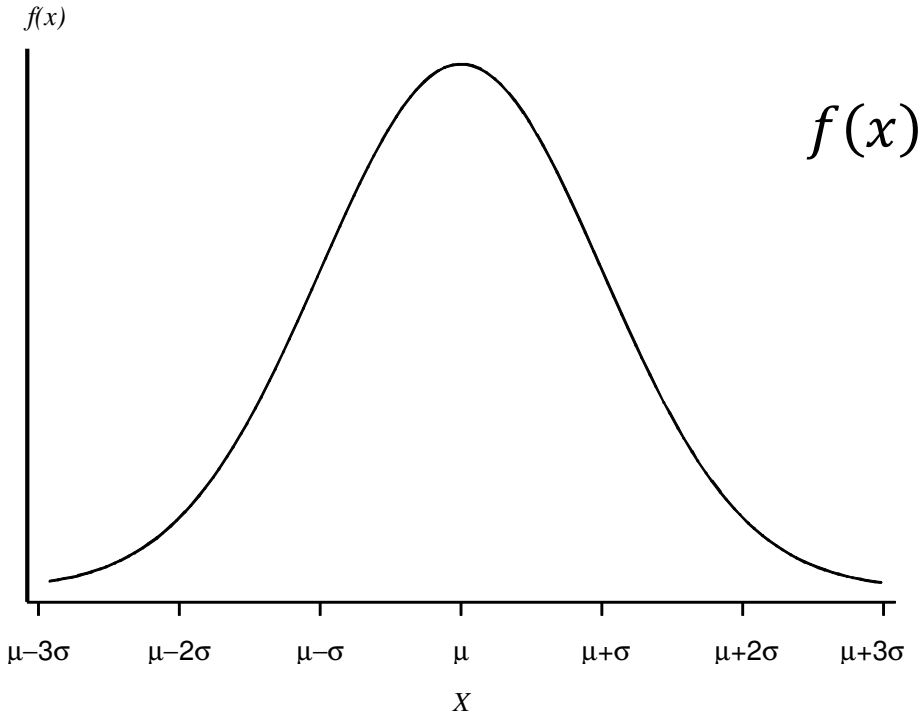
**The variance**: Gives an impression of how closely the distribution is concentrated around the expected value; it is a measure of the 'spread' of a distribution about its average value.

$$Var[X] = E[(X - E[X])^2] = \sigma^2$$

## Normal distribution $N(\mu, \sigma)$

This probability density function is a symmetrical, bell-shaped curve, centred at its expected value $\mu$. The variance is $\sigma^2$.

$N(\mu, \sigma^2)$



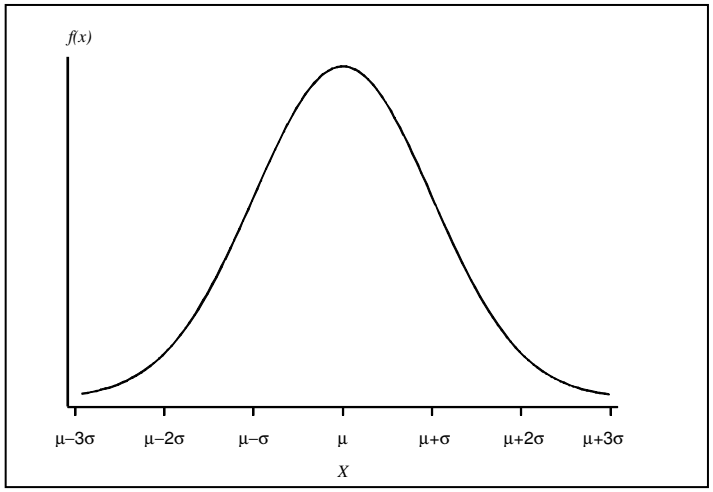

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Parameter of Normal distribution

$$Expectation: E[X] = \int_{-\infty}^{+\infty} xf(x)dx = \mu$$

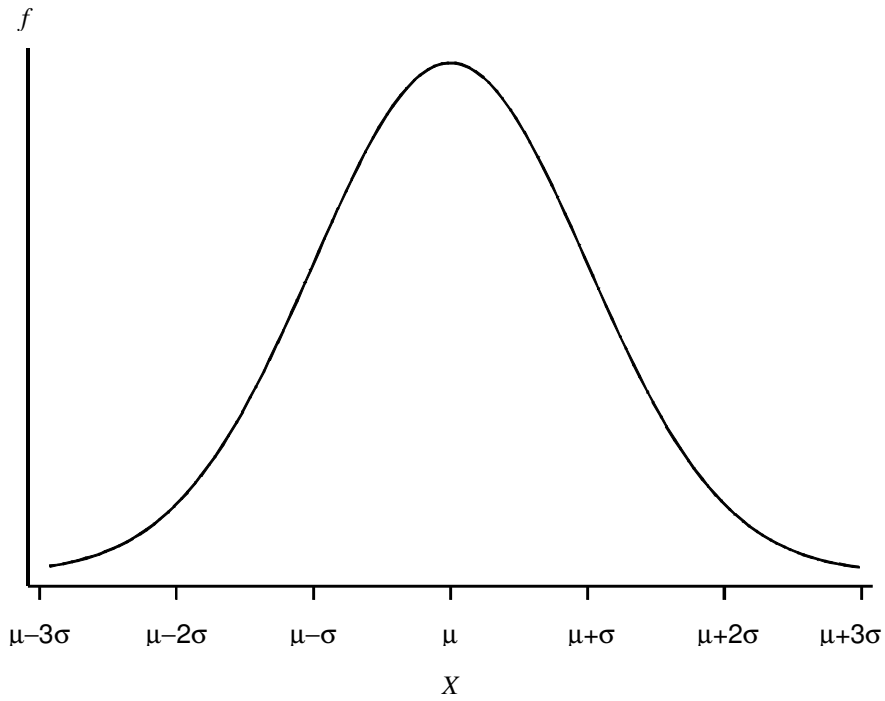
$$Variance: Var[X] = E[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx = \sigma^2$$

Many distributions arising in practice can be approximated by a Normal distribution. Other random variables may be transformed to normality.




Some properties of the normal distribution function:

- has a maximum at $x = \mu$ which is at the same time the mode, the median and the mean of the distribution
- a symmetric function around the point $x = \mu$
- has two inflection points at $x = \mu - \sigma$ and $x = \mu + \sigma$
- $\int_{-\infty}^{+\infty} f(x)dx = 1$
- $P(X = \boldsymbol{x}) = 0$




(1) about 68% of the values under the curve lie between $\pm\sigma$ around the mean $\mu$

(2) about 95% of the values under the curve lie between $\pm 2\sigma$ around the mean $\mu$ (more precisely 95% lie inside the boundaries $\pm 1.96\sigma$).

(3) about 99.7% of the values under the curve lie between $\pm 3\sigma$ around the mean $\mu$.

## Distribution model for population

The attribute $X$ which is being investigated is a **random variable**.

If every element from the population can be chosen according to a random selection (has the same chance to be chosen) and the value of the attribute $X$ can be measured, then this random process is considered as a **random experiment**.

The probabilities, with which the attribute $X$ assumes different values during the random selection, can be described by the **distribution function** $F(x) = P(X \leq x)$.

This distribution is denoted as the **distribution of population**.

The notions introduced for distribution are transferred to the population. These are for example the **expectation** or the **variance of population**.

## Choice of Population Distribution based on the Sample

The theoretical probability distribution models of the population are chosen based on information from the sample.

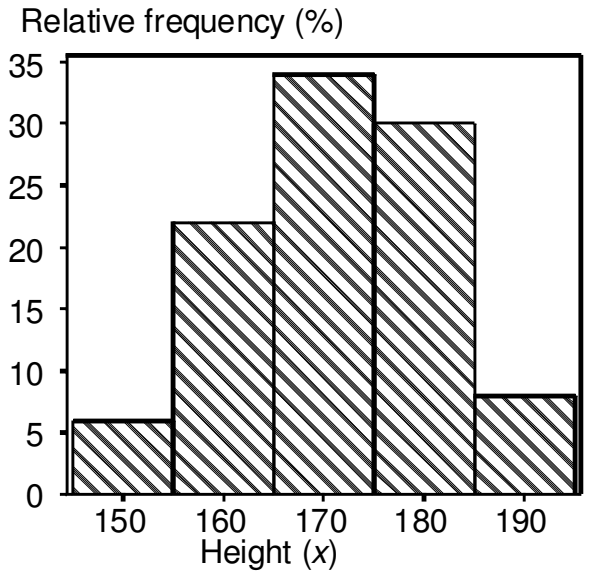
This is carried out by comparison of empiric distribution of values in the sample with theoretical distribution functions.

The theoretical density function can be considered as a limit function of an empirical histogram.

The notation of **probability distribution** corresponds to the notation of **frequencies distribution** in the empirical statistics**.**

# Histogram and Probabilities

Histogram of corn data, $n = 50$




Conclusion: To describe the distribution of data the normal distribution model is appropriate.

## Estimation of Population Parameters from the Sample

After the distribution model for the population is chosen, its parameters can be estimated from the data sample.

The goal of parameters estimation is to draw conclusion about unknown parameters of population from **only one** random sample.

Every random sample is not identical with the population. It presents only a random choice of the elements of population.

The characteristics calculated from the sample differ randomly from **true** parameter of the population.

Thus, the calculated sample characteristics are more or less good **estimates** of the true parameters.

To get the sample one proceeds as follows:

A **finite** number of elements are taken from the population according to the random selection and the values of attribute are measured.

The number of chosen elements is denoted as **a sample size $\boldsymbol{n}$**. By this one gets $n$ observed values: $x_1, x_2, \ldots, x_n$

Every observed value $x_i$ is considered a realization of one so called $i$-th **sample variable** $X_i$ $(i = 1,2, \ldots n)$.

All sample variables $X_1, X_2, \ldots, X_n$ are

- **Stochastically independent**
- **Identically distributed according to the distribution of the population**

The stochastical independence of attribute values $x_1, x_2, \ldots, x_n$ in model is expressed by the fact that they are realizations of $n$ stochastically independent random variables $X_1, X_2, \ldots, X_n$.

Under the given assumptions the sample can be used as a source of information about the underlying population.

# Estimation

- The base for the estimation of parameters of population is delivered by the observed values $x_i$ of some attribute $X$.

- The estimation function or shortly **estimator** is constructed as a function of the sample data. The estimator represents a **computation rule**.

- In accordance to the computation rule given by the estimator the single value, so called **estimate** of some unknown parameter of the population is computed.

- The single value of estimate is called a **point estimate**. This can be supplemented by the **interval estimate** (confidence interval).

Remember: Random variables are written with a large letter $(\boldsymbol{X})$, the measured value, the concrete attribute values are written with a small letter $(\boldsymbol{x})$

## Estimator vs. estimate

The estimator is a computation rule, in accordance to which the result variable is derived from the input variables. After insertion of realizations in the estimator one get estimation.

Example: **Mean**

**Estimator**: A random variable $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ ← Computation rule

$X_1, X_2, \ldots, X_n$ in estimator are $n$ stochastically independent and identically distributed random variables. That reflects the model assumptions about stochastic independence of the individual elements in the sample.

**Estimate:** A singe value

| $i$ | $x_i$ |
|---|---|
| 1 | 2 |
| 2 | 3 |
| 3 | 4 |
| 4 | 5 |

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4}{4} = \frac{2 + 3 + 4 + 5}{4} = 3.5$$ ← Single value

The observed values of attribute $x_1, x_2, \ldots, x_n$ are realizations of $n$ stochastically independent random variables $X_1, X_2, \ldots, X_n$.

Example: **Variance**

**Estimator**: A random variable

$$S^{*2} = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}$$

Computation rule

**Estimate:** A single value

| $i$ | $x_i$ | $(x_i - \mu)^2$ |
|---|---|---|
| 1 | 2 | 2.25 |
| 2 | 3 | 0.25 |
| 3 | 4 | 0.25 |
| 4 | 5 | 2.25 |

$$s^{*2} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_2 - \mu)^2 + (x_2 - \mu)^2}{4} = \frac{2.25 + 0.25 + 0.25 + 2.25}{4}$$

$$= \frac{1.25}{4} = 0.3125$$

Single value

**Notation: Parameter $\theta$ vs. estimate $\hat{\theta}$**

If the parameters of the population are estimated from the sample data, then the estimate is written with the “hat”.

**Important requirement to the estimator: Unbiasedness**

An estimator is called **unbiased**, if its expectation is equal to the true value of the parameter that is to be estimate.

$$E\left[\hat{\theta}\right] = \theta$$

Otherwise the estimator is biased.

The amount of deviation of its expectation from the true value of parameter is called the **bias**:

$$Bias\ \hat{\theta} = E\left[\hat{\theta}\right] - \theta$$

## Estimation of the mean μ of the normally distributed population

The sample mean is a unbiased estimator of the true value of the population

$$E[\bar{X}] = \mu$$

and

$$\hat{\mu} = \bar{x}$$

where

$\bar{X}$ – random variable mean

$\mu$ – true mean of the population

$\hat{\mu}$ – estimate of the mean $\mu$

$\bar{x}$ – sample mean

## Estimation of the variance $\sigma^2$ of the normally distributed population

Estimator of variance, **if $\mu$ is known**:

$$S^{*2} = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}$$

Estimator of variance, **if $\mu$ is unknown**

$$S^{*2} = \frac{\sum_{j=1}^{n}(X_i - \bar{X})^2}{n}$$

⇨ The unknown parameter of the population $\mu$ must be replaced by the sample mean $\bar{X}$

It is a **biased** estimator of the population variance, because

$$E\left[S^{*2}\right] = \frac{n-1}{n}\sigma^2$$

The estimate

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

is biased by the factor $(n-1)/n$.

Thus, the inverse of this factor must be inserted as a correction.

**Unbiased estimator of the variance:**

$$S^2 = \frac{\sum_{j=1}^{n}(X_i - \bar{X})^2}{n-1}$$

$$E[S^2] = \sigma^2$$

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

Unbiased estimate
of the population variance

Corrected sample variance.
Further: sample variance

## Probability statements about population based on the sample

1. **Choice of a distribution model**: Comparison of the empiric histogram with theoretical density functions.

2. **Estimation of parameters of the population.** Computation of the estimates of population parameters from the sample following computation rules for estimators.

**Example:**

- Corn field with 100.000 plants
- Plants height measured in cm ($x$)

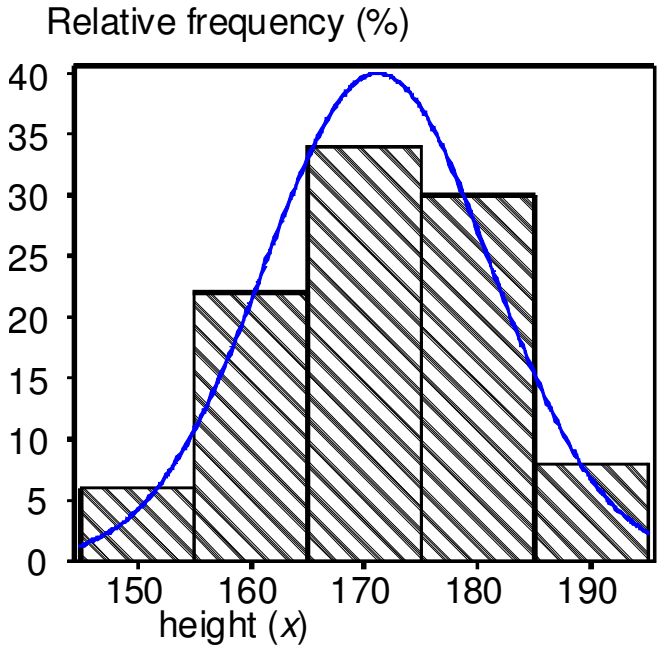
**Choice of a distribution model:** From comparison of the empiric histogram with theoretical density functions we assume, that the height are normally distributed.

**Estimation of parameters of the population:** Estimates of population parameters from the sample are as follows:

$$\hat{\mu} = 170;\ \hat{\sigma} = 10$$

Histogram of corn data with overlying density function of a normal distribution:

$\mu = 170;\ \sigma = 10$




(1) about 68% of the corn plants have a height between 170 – 10 and 170 + 10 cm, so between 160 and 180 cm.

(2) about 95% of the corn plants have a height between 170 – 2*10 and 170 + 2*10 cm, so between 150 and 190 cm.

(3) about 99.7% of the corn plants have a height between 170 – 3*10 and 170 + 3*10 cm, so between 140 and 200 cm.