Basics in 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
    - Basic on sampling
    - Expected value
    - Probability
    - Normal distribution
    - Confidence intervals
    - Basics on statistical testing

4. Two-Variable Statistics

- Scatterplots
- Correlation
- Regression

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 dh/ha)

The statistic data analysis depends on the type of variable!

Categorical data

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

  Yes $\Longrightarrow$ ordinal

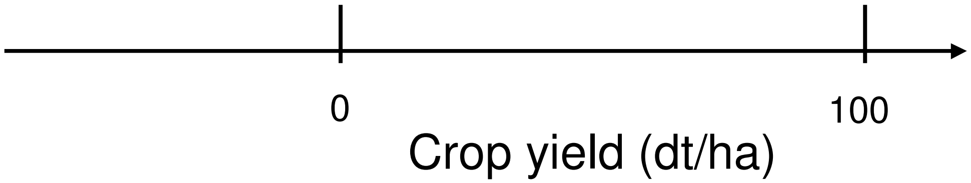
  No $\Longrightarrow$ nominal

- 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 sets

Frequency Distribution: discrete data (categorical & metric)

Bar diagram is a graphic representation of the frequency distribution of **discrete** data

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):

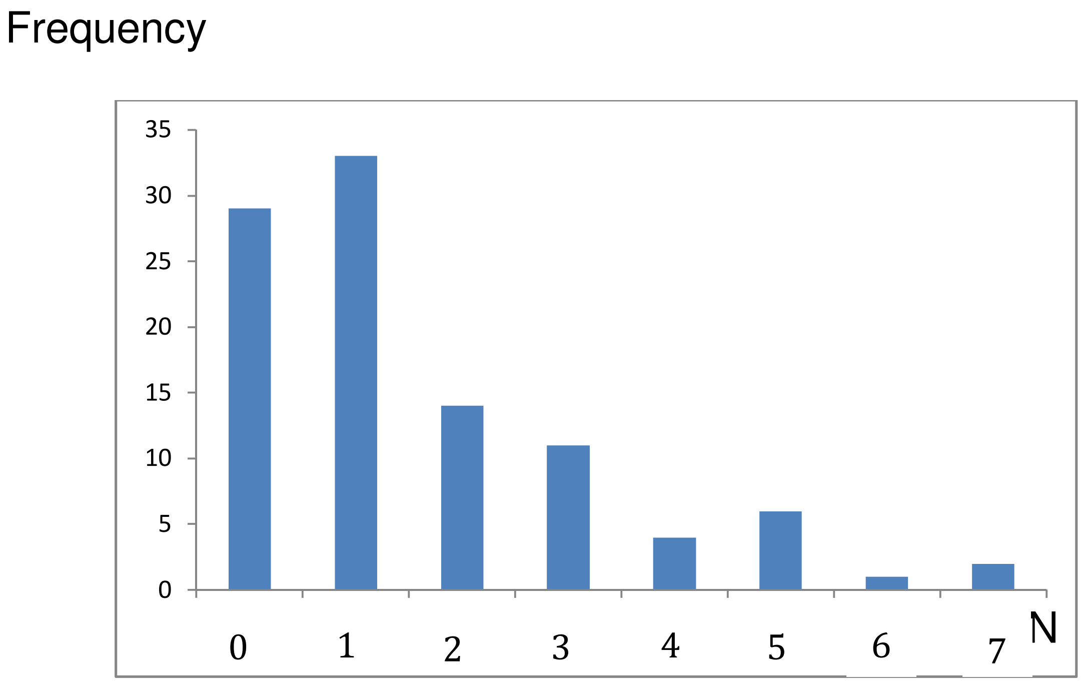
$n = 100$ plots, each $1m^2$, the number of plants is counted:

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/	Percent/
		absolute frequency	relative frequency
0	𝍸 𝍸 𝍸 𝍸 𝍸 ||||	29	0.29
1	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |||	33	0.33
2	𝍸 𝍸 ||||	14	0.14
3	𝍸 𝍸 |	11	0.11
4	||||	4	0.04
5	𝍸 |	6	0.06
6	|	1	0.01
7	||	2	0.02

Bar diagram



Frequency Distributions: continuous data

Example: (H.-P. Piepho. Statistics , University of Hohenheim):

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

175 172 179 167 163 154 163 164 157 177

186 165 175 194 176 162 166 169 170 181

168 166 180 164 179 170 150 192 170 173

170 150 174 164 182 188 157 165 172 168

179 179 164 162 178 162 182 171 182 183

Forming of classes:

Number of classes k: rules of thumb

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

or

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

In our case:

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

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

We choose $k = 5$

Width of classes:

$$b > \frac{x_{max} - x_{min}}{k} = \frac{194 - 150}{5} = 8.8$$

We choose $b = 10$

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

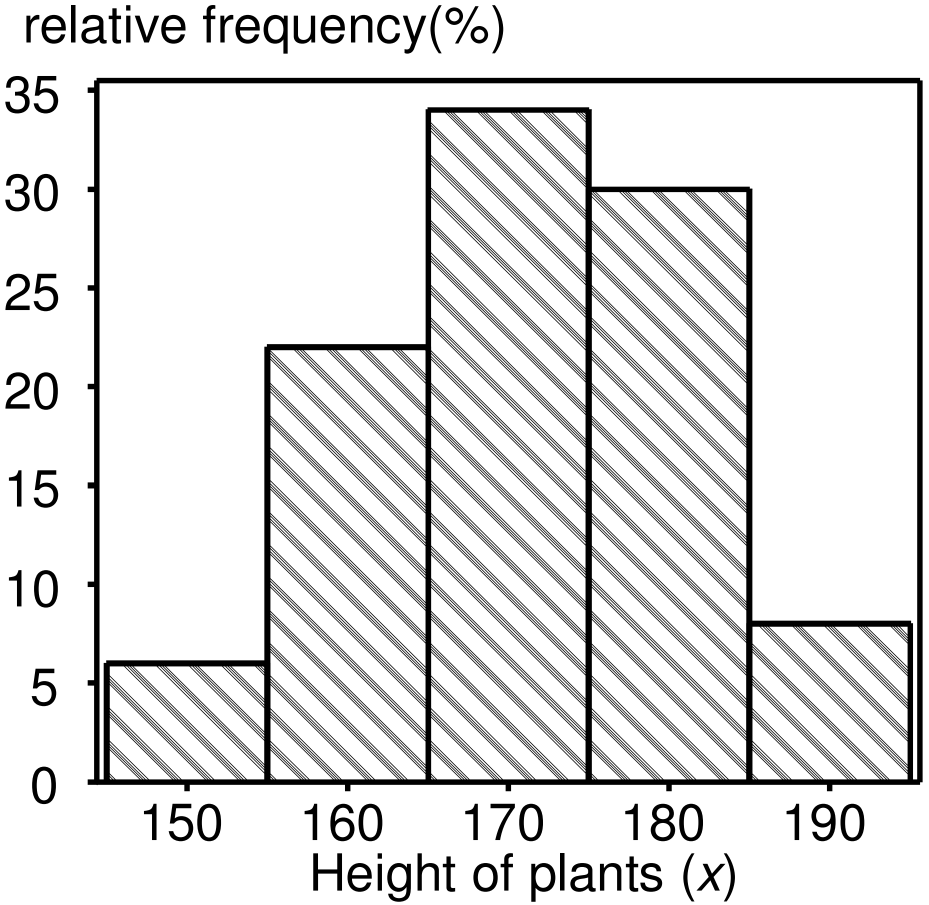
Precise mathematical notation of classes (without overlapping!):

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

Check list of frequencies

Height (cm)	plot	number of plants/	Percent/
	counts	absolute frequency	relative frequency
145 – 154.9	|||	3	6
155 – 164.9	卌 卌 |	11	22
165 – 174.9	卌 卌 卌 ||	17	34
175 – 184.9	卌 卌 卌	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

At least t% of values are less than or equal to Q_t

Median = Q_{50}= 50%-quantile

Example: Sample of corn *n* = 16, 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 $\underset{Q_{50}}{\Big|}$ 172 175 175 176 177 179 186 194

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

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
$\boldsymbol{x}_{[1]}$	$\boldsymbol{x}_{[2]}$	$\boldsymbol{x}_{[3]}$	$\boldsymbol{x}_{[4]}$	$\boldsymbol{x}_{[5]}$	$\boldsymbol{x}_{[6]}$	$\boldsymbol{x}_{[7]}$	$\boldsymbol{x}_{[8]}$	$\boldsymbol{x}_{[9]}$	$\boldsymbol{x}_{[10]}$	$\boldsymbol{x}_{[11]}$	$\boldsymbol{x}_{[12]}$	$\boldsymbol{x}_{[13]}$	$\boldsymbol{x}_{[14]}$	$\boldsymbol{x}_{[15]}$	$\boldsymbol{x}_{[16]}$

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

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

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

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

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

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

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:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{175 + 172 + 179 + \cdots + 194 + 176 + 162}{16} = 170.5625$$

$(Median\ =\ 169.5)$

Median: more robust against outliers than the mean

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

11175 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)

$x_{[i]}$	$[i]$	
0.30	1	
0.34	2	
0.39	3	
0.40	4	
0.40	5	
0.42	6	
0.48	7	$\boxed{Q_{25}}$
0.54	8	
0.56	9	
0.58	10	
0.62	11	
0.68	12	
0.74	13	
0.74	14	
0.78	15	
0.82	16	
0.96	17	
1.02	18	

1.06	19	
1.10	20	
1.44	21	$\boxed{Q_{75}}$
1.60	22	
1.68	23	
2.40	24	
2.40	25	
2.56	26	
3.60	27	
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.

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

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

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

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

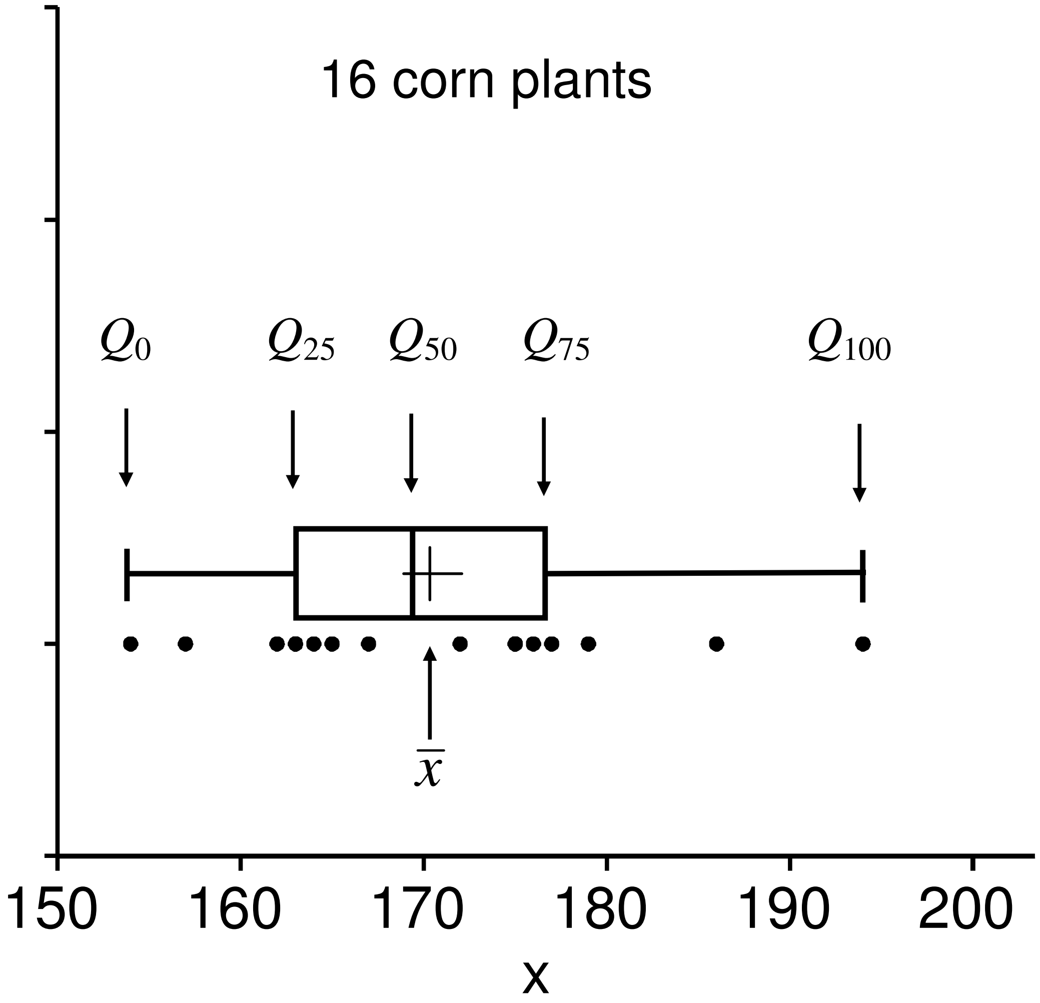
Example (Control)

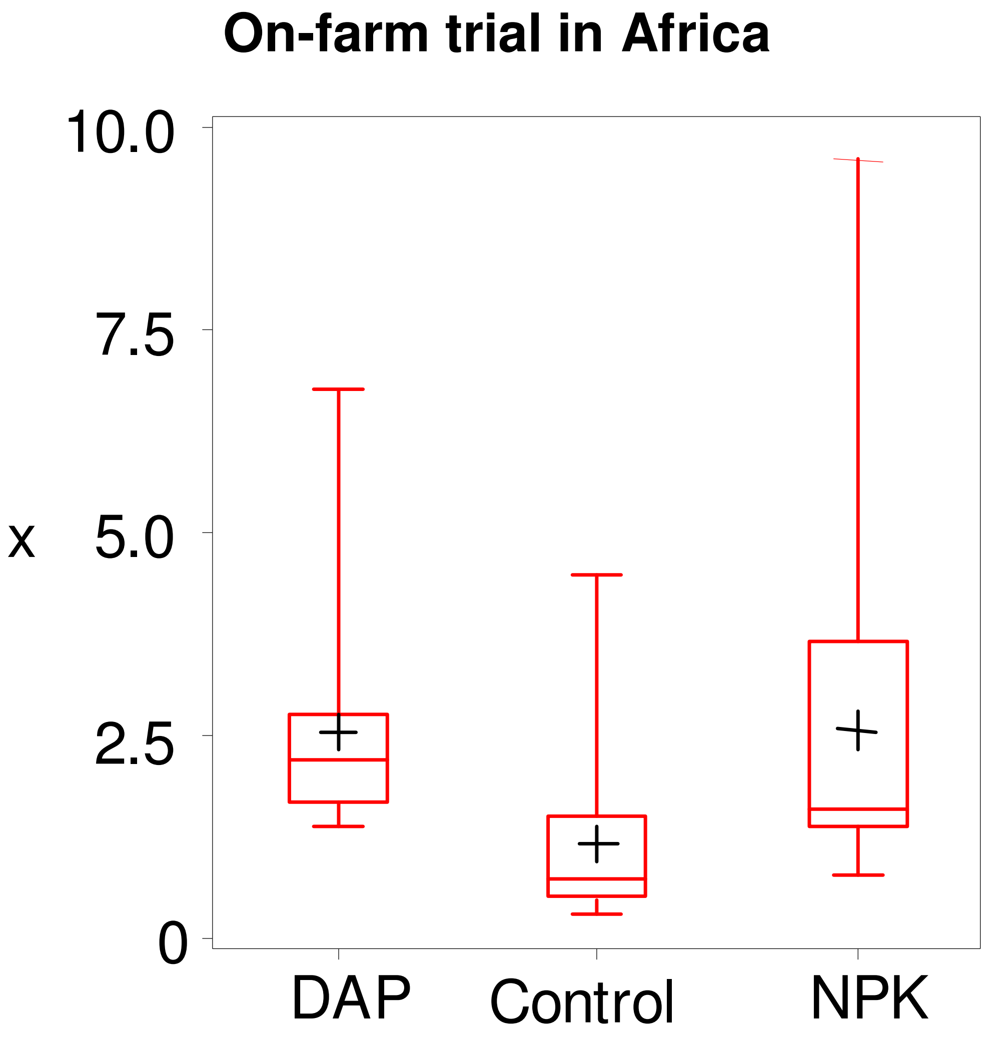
$$\bar{x} = 1.183;\ s_x = 1.039;\ CV = \frac{1.039}{1.183} = 0.8784 = 87.84\%$$

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 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 of an agricultural trial

- Typically, the population is very large, making a census or 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 trial 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	μ	$\bar{x}$
variance	σ_x^2	s_x^2

Basics of sampling

Sampling

- Selecting a random (or a wishfully "*representative*") subset of a population.
- **Sampling** is connected with the selection of a subset of individuals within a population to estimate characteristics of the whole population.
- Researchers rarely survey the entire population because the cost of a census is too high.

Sampling methods

- **Simple random sampling:** each object of the population has the same chance of being chosen
- **Systematic sampling:** the selection of elements from an ordered sampling frame. Systematic sample units are uniformly distributed over the population
- **Stratified sampling:** Where the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is sampled as an independent sub-population, out of which individual elements can be randomly selected

Simple random sampling

- Each individual is chosen randomly and entirely by chance
- Each individual has the same probability of being chosen at any stage during the sampling process
- Each subset of k individuals has the same probability of being chosen for the sample as any other subset of k individuals

Systematic sampling

- Is to be applied only if the given population is logically homogeneous
- The most common form of systematic sampling is an equal-probability method, in which every k^{th} element in the frame is selected
- A random starting point
- The sampling interval k (sometimes known as the *skip*), is calculated as:

$$k = \frac{N}{n}$$

where n is the sample size, and N is the population size.

- Each element in the population has a known and equal probability to be selected.
- The chosen sampling interval must not hide a pattern

Stratified sampling

- Is advantageous, when subpopulations within an overall population vary
- Members of the population are divided into homogeneous subgroups (strata) before sampling **(Stratification)**
- The strata should be mutually exclusive: every element in the population must be assigned to only one stratum
- The strata should also be collectively exhaustive: no population element can be excluded
- Each subpopulation (stratum) is sampled independently
- Within each stratum the random or systematic sampling is applied

Random Variable Y

[is always written with 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 variable - 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 cows 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 $\boldsymbol{y}$ [is always written with lowercase letter]

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

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 by:

Probability distribution: Is a function that describes the probability of a random variable taking certain values.

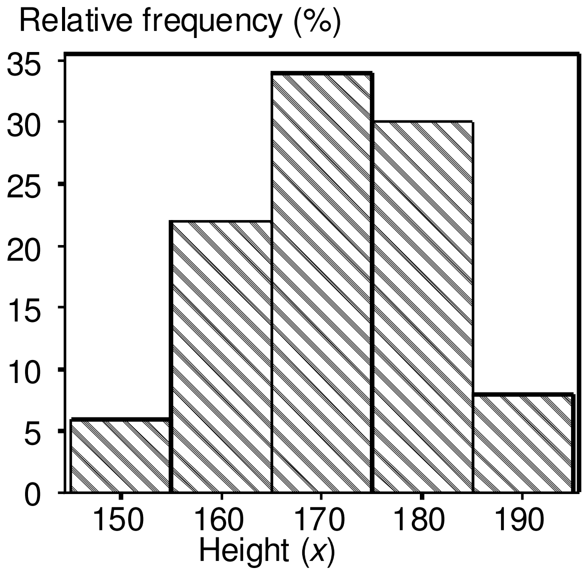
The **expected value** (or **expectation**, or **mathematical expectation**, or **mean**) of a random variable

- 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 densities of probability density function (f).

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

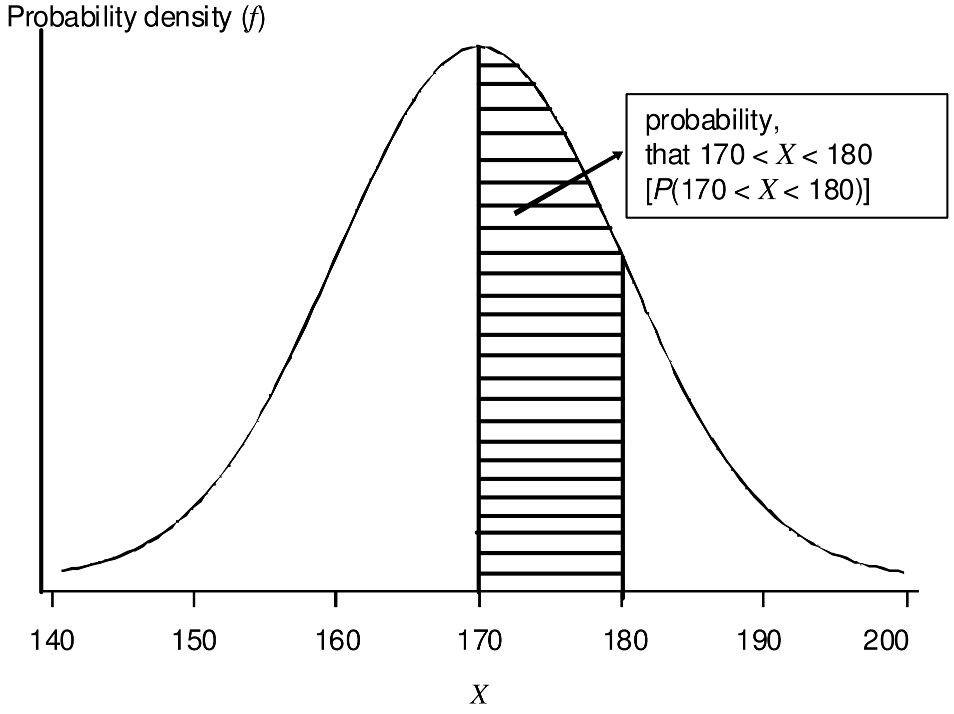
Histogram and Probabilities

Histogram for corn data



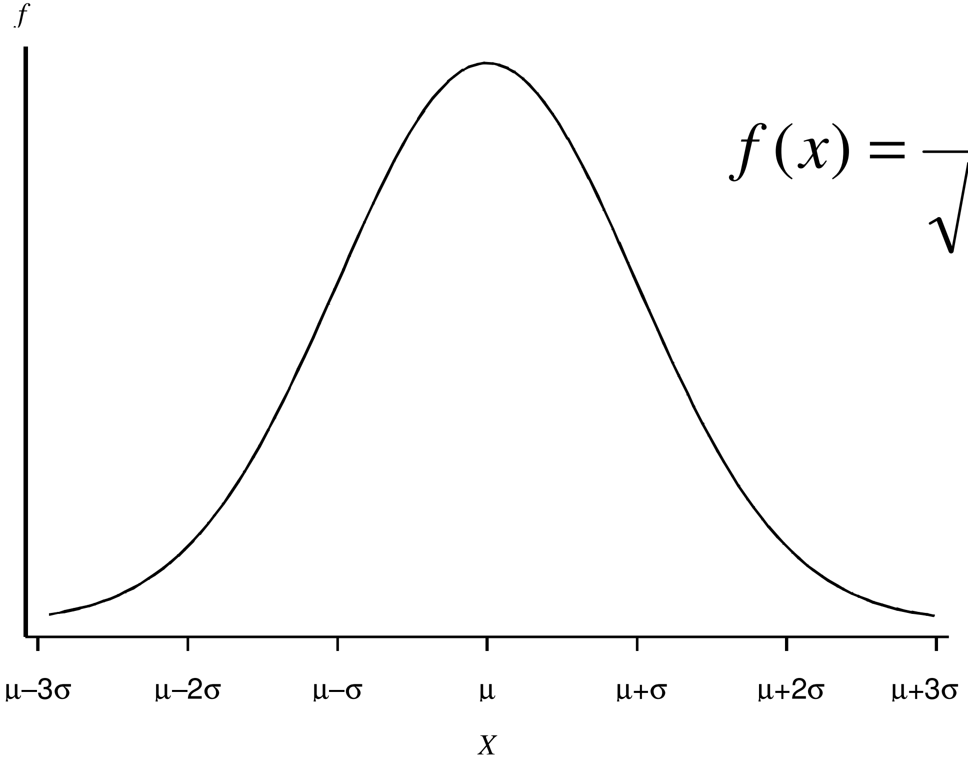
Probability density function (*f*) for corn example

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.



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

This probability density function is a symmetrical, bell-shaped curve, centred at its expected value μ. The variance is σ^2.

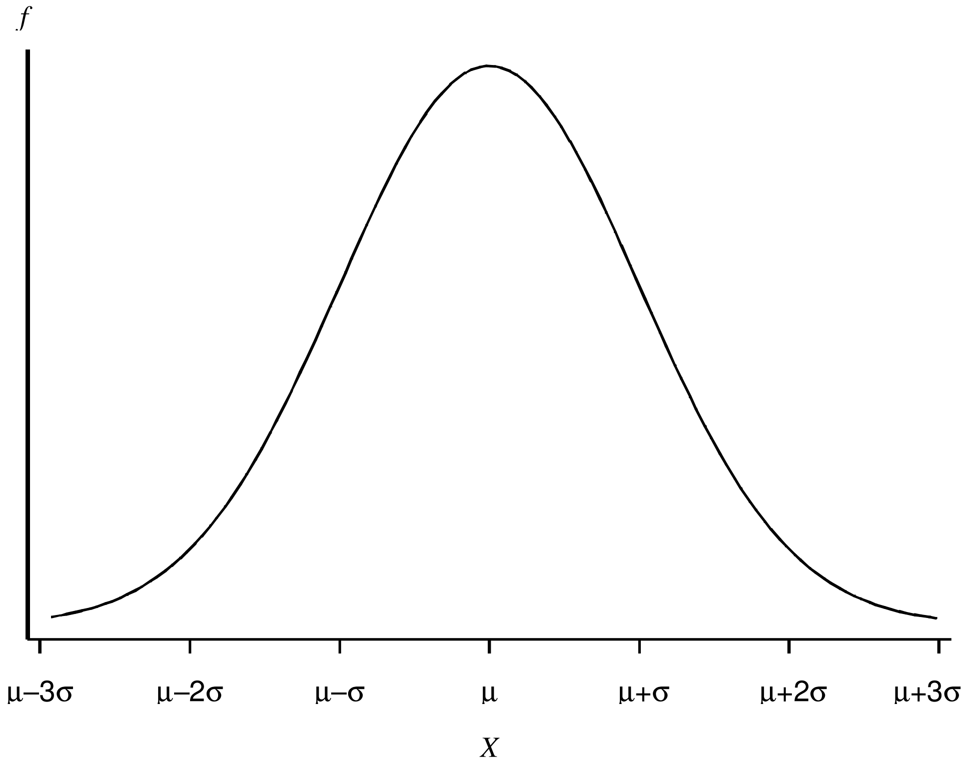


$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(x-\mu)^2}{2\sigma_x^2}}$$

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

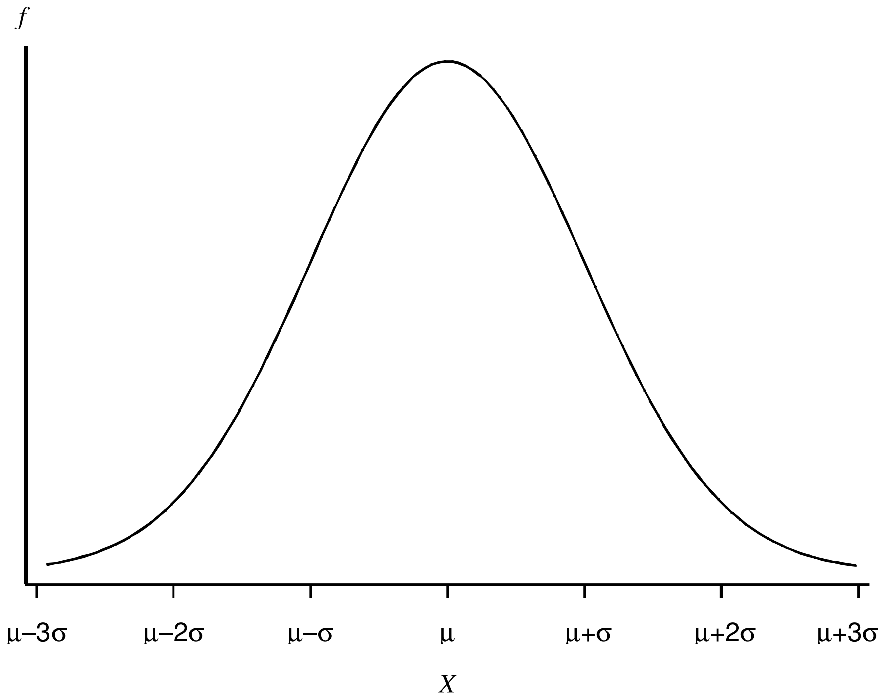
Some properties of the normal distribution function:

- 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 $x = \mu + \sigma$
- $\int_{-\infty}^{+\infty} f(x)dx = 1$
- $P(X = \boldsymbol{x}) = 0$



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

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



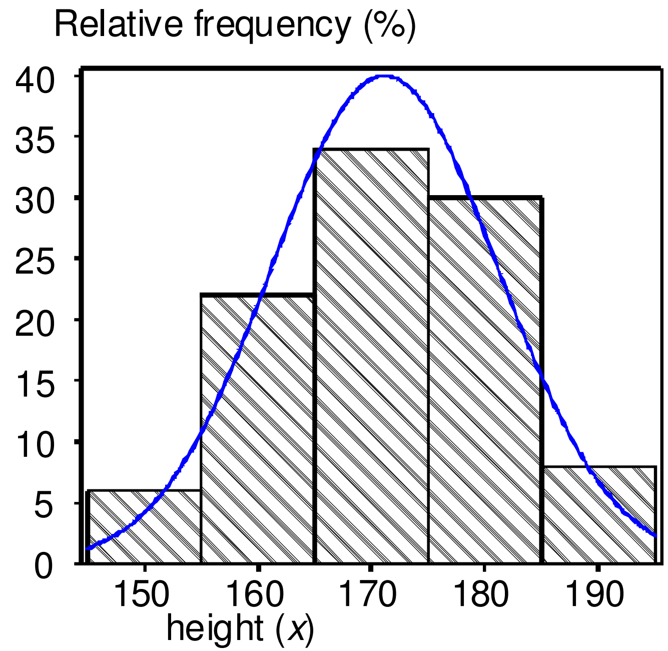
(1) about 68% of the values under the curve lie between ±σ_x around the mean μ

(2) about 95% of the values under the curve lie between ±2σ_x around the mean μ (more precisely 95% lie inside the boundaries ± 1.96σ_x).

(3) about 99.7% of the values under the curve lie between ±3$\sigma_x$around the mean μ.

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

For the corn data:

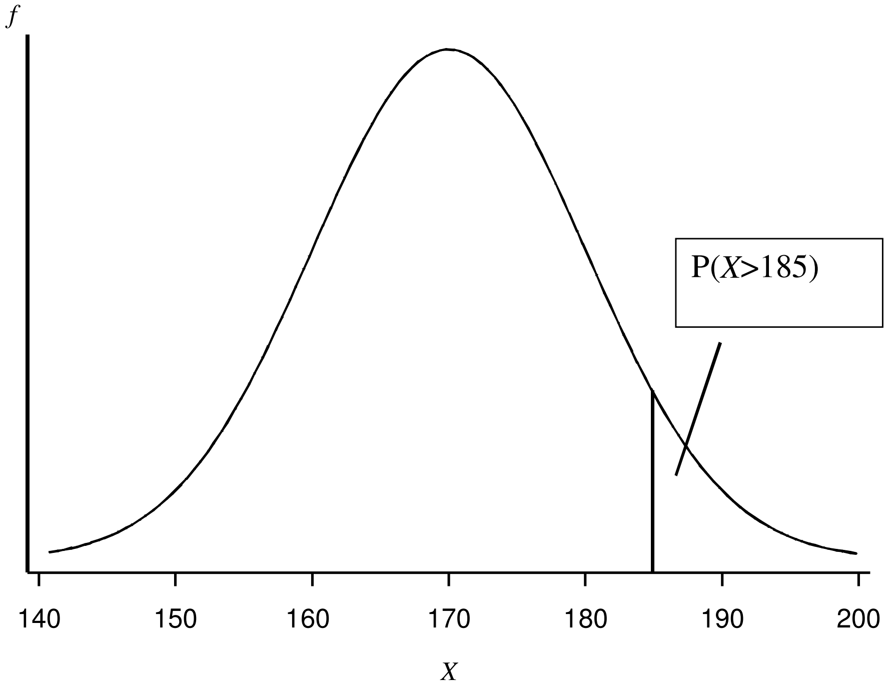


(1) about 68% of the corn plants have a height between 170 – 10 and 170 + 10 cm, so between 160 and 180 cm.

(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.

Question 1: Compute a probability, that one randomly taken plant of the field has a height of $X = 185$ cm or longer



$$P(X > 185) = \int_{185}^{\infty} \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(x-\mu)^2}{2\sigma_x^2}} dx = \int_{185}^{\infty} f(x)dx = 0.0668$$

The **standard normal distribution** is a distribution with $\mu = 0$ and $\sigma^2 = 1$

Notation: $\boldsymbol{N(0,1)}$

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

Z-Transformation and Standard Normal Distribution

The simplest case of the normal distribution, known as the Standard Normal Distribution, has expected value zero and variance one. This is written as $N(0,1)$.

- The Standardization of arbitrary normal distribution

- We need a table only for one distribution!

Z-Transformation:

$$Z = \frac{(X - \mu)}{\sigma_x}$$

Z-Transformation for height of $X = 185$: $\mu = 170$; $\sigma = 10$

$$z = \frac{(185 - 170)}{10} = 1.5$$

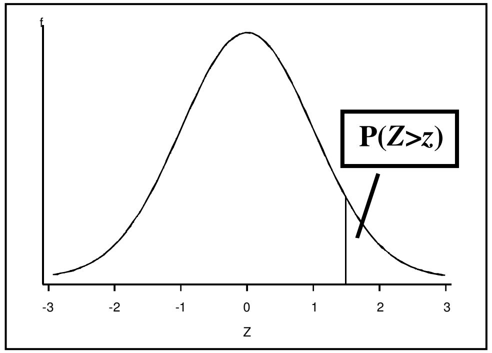
$$P(X > 1.5) = \int_{1.5}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \int_{1.5}^{\infty} f(z) dz = 0.0668$$

The area under the standard normal distribution to the left of $z = 1{,}5$ is equal to a probability, that $X > 185$.

In General:

The probability $P(Z > z)$ is called **exceedance probability.**

It represents a the probability that a standard normal random variable (Z) is greater than a given value (z)



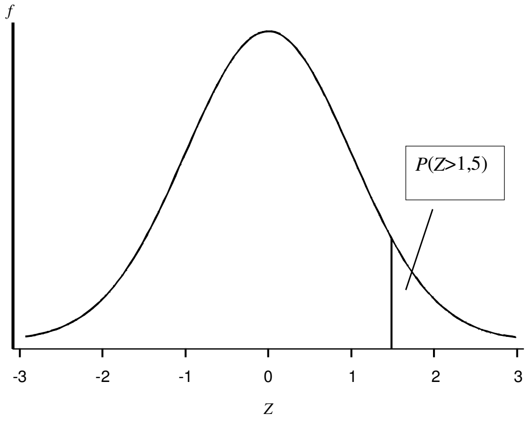
Exceedance probability $P(Z > z)$ for standard normal distribution.

Example: $P(Z > 1.96) = 0.025$

z	0.00	0.01	0.02	0.03	0.04
0.0	0.5000	0.4960	0.4920	0.4880	0.4840
0.1	0.4602	0.4562	0.4522	0.4483	0.4443
0.2	0.4207	0.4168	0.4129	0.4090	0.4052
0.3	0.3821	0.3783	0.3745	**0.3707**	0.3669
0.4	0.3446	0.3409	0.3372	0.3336	0.3300
0.5	0.3085	0.3050	0.3015	0.2981	0.2946

$$P(Z > 0.33) = 0.3707$$

Example:



$$\bar{x} = 185;\ \sigma_x = 10$$

Then

$$z = \frac{(x-\mu)}{\sigma_x} = \frac{185-170}{10} = 1.5$$

$$P(X > 185) = P(Z > 1.5) = 0.0668$$

In General:

If X is normal distributed with mean μ and variance σ_x^2

then

$$Z = \frac{X - \mu}{\sigma_x}$$

is standard normal distributed with mean 0 and variance 1

Question 2: Compute a probability, that the corn plants have a height between 175 and 185 cm.

$$\mu = 170;\ \sigma_x = 10$$

Then

$$z_1 = \frac{(x-\mu)}{\sigma_x} = \frac{175-170}{10} = 0.5$$

$$z_2 = \frac{(x-\mu)}{\sigma_x} = \frac{185-170}{10} = 1.5$$

From the table:

$P(Z > 1.5) = 0.0668$ and $P(Z > 0.5) = 0.3085$

So the area between 175 and 185 is equal to $0.3085 - 0.0668 = 0.2417$

The probability, that the corn plants have a height between 175 and 185 cm is 0.2417 or 24.17%

Question 3: Compute a probability, that the corn plants have a height smaller than 145 cm.

$$z_1 = \frac{(x - \mu)}{\sigma_x} = \frac{145 - 170}{10} = -2.5$$

$$P(Z < -2.5) = P(Z > 2.5) = 0.0062$$

Distribution symmetry

The probability, that the corn plants have a height smaller than 145 cm is 0.0062 or 0.62%

Question 4: Compute a probability, that the corn plants have a height between 145 and 175cm.

$$P(145 < Z < 175)-?$$

$$P(145 < Z < 175) = P(-2.5 < Z < 0.5) = \overset{\boxed{\text{Full area under the standard normal distribution curve}}}{1} - \underset{\boxed{\text{Area to the left of } -2.5}}{0.0062} - \underset{\boxed{\text{Area to the right of } 0.5}}{0.3085} = 0.6853$$

The probability, that the corn plants have a height between 145 and 175 cm is 0.6853 or 68.53%

Question 4: Compute a probability, that the corn plants are higher than 170 cm

$$z = \frac{170 - 170}{10} = 0$$

$$P(Z > 0) = 0.5$$

The probability, that the corn plants have a height greater than 170 cm is 0.5 or 50%

Important:

By continuous distributions $P(X = x) = 0$!

So

$$P(X > x) = P(X \geq x)$$

and

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