WHY DO WE NEED STATISTICS? – ILLUSTRATIVE EXAMPLES

1. Steam consumption of a hydrocarbon process

The goal is to estimate the steam demand of a particular process.
- *Is the estimated model „good enough”?*
- *In what extent can we trust in the estimate?*

Date	Outside Temp. °C	Feed, m3/h	Raw material IB content, m/m%	Steam consumption t/h	Estimated steam consumption, t/h
2012-01-01	1,2	17,0	27,0	3,83	3,36
2012-01-02	4,6	17,0	20,5	3,80	3,51
2012-01-03	3,7	17,0	20,9	3,79	3,50
2012-01-04	4,3	17,0	21,9	3,87	3,47
2012-01-05	3,9	16,7	20,0	3,91	3,50
2012-01-06	4,7	16,5	22,8	3,76	3,38
2012-01-07	3,0	16,5	22,5	3,75	3,41
2012-01-08	1,5	16,5	22,9	3,78	3,41

WHY DO WE NEED STATISTICS? – ILLUSTRATIVE EXAMPLES

2. Metal content of soil samples

A new analytical method based on x-ray fluorescence (XRF) spectroscopy is introduced to measure the metal content of soil samples. The XRF analyzer enables us to measure the metal content in situ, thus there is no need to take the samples into the laboratory.
- *Do we get the „same” results with the old technique in the laboratory and with the XRF analyzer on field?*

The conditions of the analysis with the XRF analyzer is chosen by the analyst. The conditions are: 1) humidity content of the samples 2) preparation of samples 3) amount of the analyzed sample 4) duration of the measurement.
- *In what sense can a setting be better than another?*

Experiments are made to find the „best conditions”.
- *How should we carry out the experiments and evaluate the results?*

MEASUREMENT SCALES

- Nominal, categorical
- Ordinal, ordered categorical
- Interval
- Proportional, ratio

GAS BOTTLE PROBLEM

There are 2.5 million gas bottles in Hungary. They want to fill more gas into them. The volume of the bottle limits the maximum amount of gas. (i.e. the smaller the bottle the higher the pressure in it – for a given amount of gas) Even though the nominal volume of the bottle is 25 l, some of them are smaller others are larger than 25 l.

How large is the smallest bottle?

GAS BOTTLE PROBLEM

500 bottles were selected and their volume were measured.

The year of production is displayed on the bottles.

	Date	Volume
1	1970	24,96
2	1970	25,37
3	1965	25,22
4	1960	25,41
5	1970	24,93
6	1965	24,82
7	1975	25,16
⋮	⋮	⋮
500		

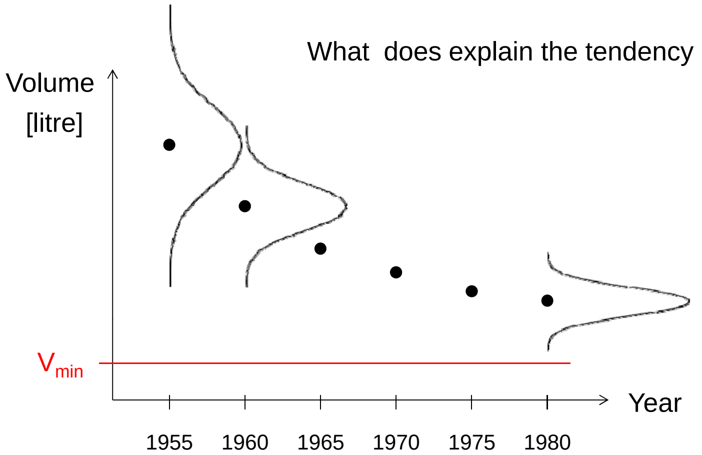
First step of the analysis:

Visualizing data

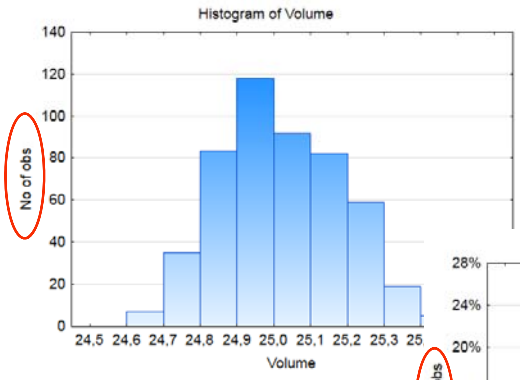
GAS BOTTLE PROBLEM

Why do we have bottles from every five years?

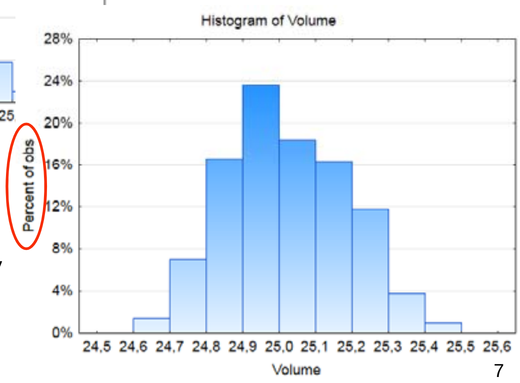
What does explain the tendency in the volume?



GAS BOTTLE PROBLEM

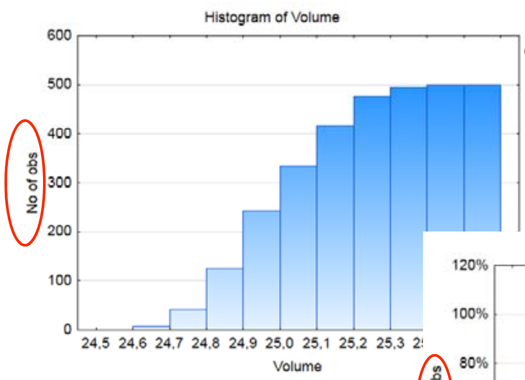


Frequency histogram

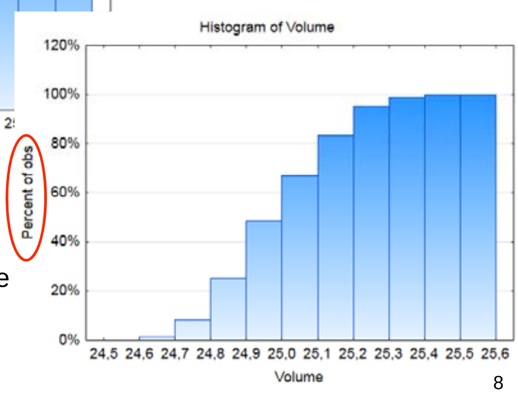


Relative frequency histogram

GAS BOTTLE PROBLEM

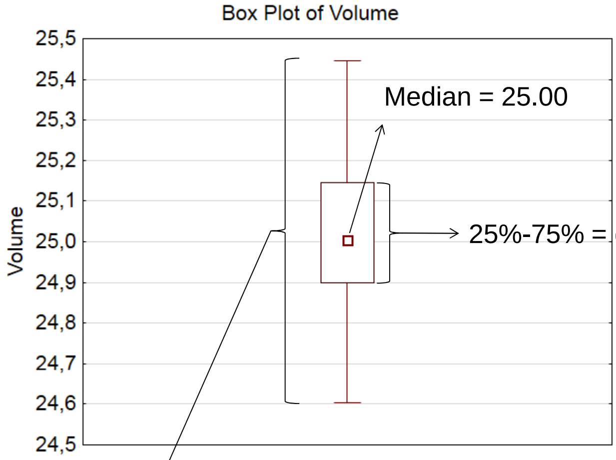


Cumulative frequency histogram



Cumulative relative frequency histogram

Gas Bottle Problem

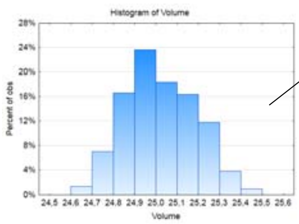


Median = 25.00

25%-75% = (24.90, 25.15)

Non-Outlier Range
= (24.60, 25.45)

Gas Bottle Problem – Population and Sample



This is the distribution of the volume in the sample.

What would you expect? Is the distribution of the 2.5 million bottle similar to this?

The importance of sampling

<u>What we have</u>

Sample
Histogram

Conclusion

<u>What we are interested in</u>

Population (Lot)
Density function

RANDOM VARIABLES

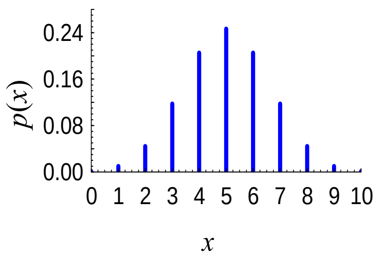
The variability, uncertainty is inevitable

- produced parts are different
- repeated measurements differ
- the ratio of nonconforming products differ from sample to sample
- two water samples from the same natural water can never be the same
- samples on different days are not identical

DISCRETE RANDOM VARIABLE

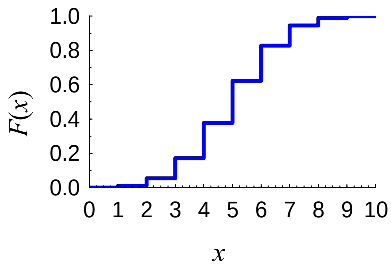
Random variable: It does not have a fixed value.
The discrete random variable can take on a set of possible different values, each with an associated probability.
Eg. tossing a fair dice 10 times. *x*: number of heads.

Probability density function



$$p(k) = P(x = k)$$

Cumulative distribution function

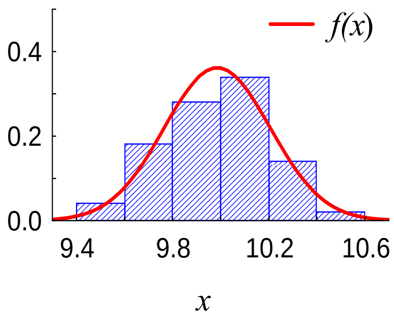


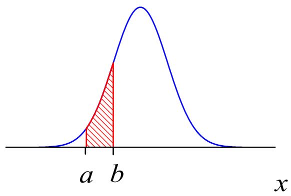
$$F(k) = P(x \le k) = \sum_{x_i \le k} p(x_i)$$

CONTINUOUS RANDOM VARIABLE

The continuous random variable can take any numerical value in a given interval with some probability.

Probability density function

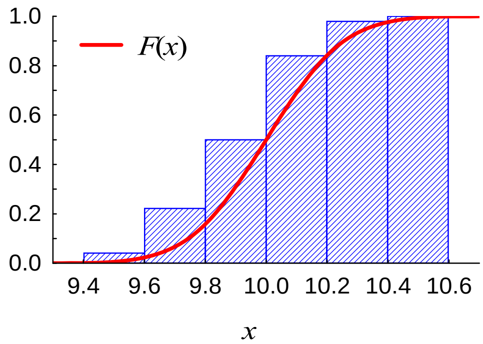


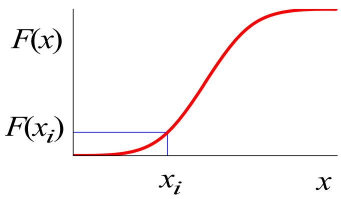


$$P(a < x \le b) = \int_a^b f(x)dx$$

CONTINUOUS RANDOM VARIABLE

Cumulative distribution function





$$F(x_i) = P(x \le x_i) = \int_{-\infty}^{x_i} f(x)dx$$

PARAMETER AND STATISTIC

Parameter
Characteristic of the **population**
Typically constant

expected value:

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx$$

variance

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

Statistic
Characteristic of the **sample**
random variable

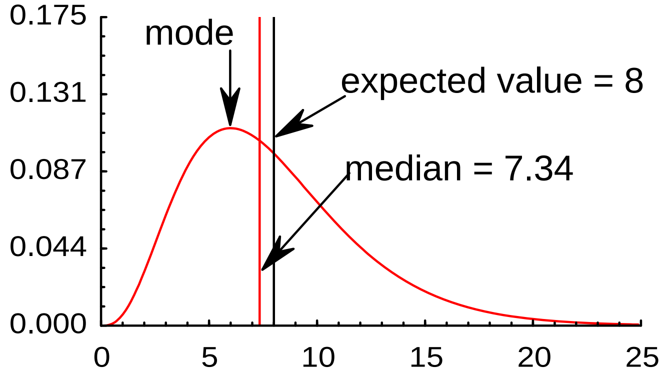
sample mean:

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

sample variance

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

ON AVERAGE…



Independent Measurements

The data in the samples are the results of measurements. These are subject to error.

Type of measurement errors:
- systematic
- random

Why is it important to do independent repetitions?

Two measurements are independent if their errors are independent.

Functions of Random Variables

$$E(cx) = cE(x)$$

The expected value of the volume of gas bottles is 25.8 dm^3. What is the expected value in cm^3?

$$Var(cx) = c^2 Var(x)$$

The variance of the volume of gas bottles is 0.25 $(dm^3)^2$. What is the variance in $(cm^3)^2$?

Functions of Random Variables

$$E(x_1 + x_2 + ... + x_n) = E(x_1) + E(x_2) + ... + E(x_n)$$

If x_1, x_2, …, x_n are independent:

$$Var(x_1 + x_2 + ... + x_n) = Var(x_1) + Var(x_2) + ... + Var(x_n)$$

If x_1, x_2, …, x_n are independent and has the same distribution, with E(x) expected value and Var(x) variance.

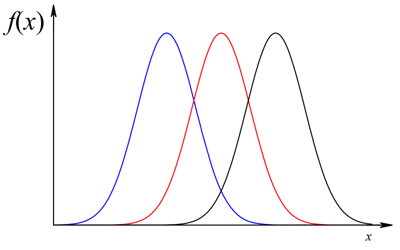
$$E(x_1 + x_2 + ... + x_n) = nE(x)$$

$$Var(x_1 + x_2 + ... + x_n) = nVar(x)$$

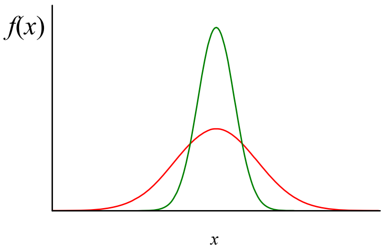
Gauss (Normal) Distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right]$$

Two parameters: μ and σ^2



μ is different



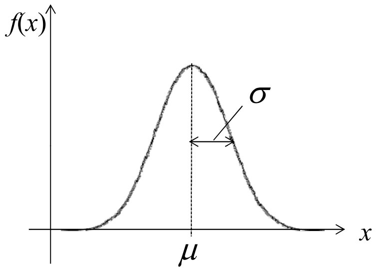
σ^2 is different

GAUSS (NORMAL) DISTRIBUTION

Short notation $N(\mu,\sigma^2)$ e.g. $N(0,1)$

Expected value $E(x)=\mu$

Variance $Var(x)=\sigma^2$



Standardisation

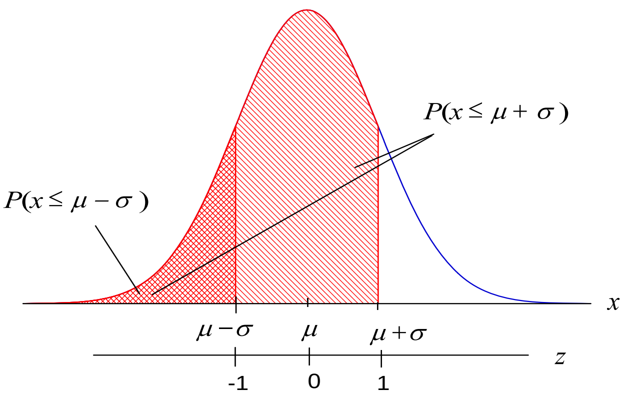
$$z=\frac{x-\mu}{\sigma} \qquad f(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)$$

$$\mu=E(z)=0 \qquad \sigma^2=Var(z)=1$$

GAUSS (NORMAL) DISTRIBUTION

What is the probability of finding the *x* Gauss d. random variable in the $(\mu-\sigma,\mu+\sigma)$ range?

$$P(\mu-\sigma<x\le\mu+\sigma)=F(\mu+\sigma)-F(\mu-\sigma)$$



$$z=\frac{x-\mu}{\sigma}$$

Gauss (Normal) Distribution

$$z_{lower} = \frac{\mu - \sigma - \mu}{\sigma} = -1 \qquad z_{upper} = \frac{\mu + \sigma - \mu}{\sigma} = 1$$

Width of the interval	$\pm\sigma$	$\pm 2\sigma$	$\pm 3\sigma$
P	0.68268	0.9545	0.9973

Interpretation of the results

The variance of a measurement system is 0.09(mg/l)^2.

The measured concentration of the same sample vary in a ±2*0.3mg/l= ±0.6mg/l range with ~96% probability.

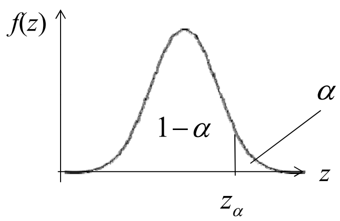
The maximum bite force of a spotted hyena is 4500N, the sigma is 45N.

The maximum bite force of a spotted hyena is between 4365-4635N with ~99.7% probability.

Notation

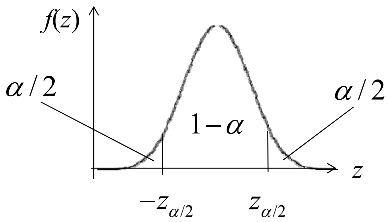
One sided

$$P\left(z \le z_{\alpha}\right) = 1 - \alpha$$



Two sided

$$P\left(-z_{\alpha/2} < z \le z_{\alpha/2}\right) = 1 - \alpha$$



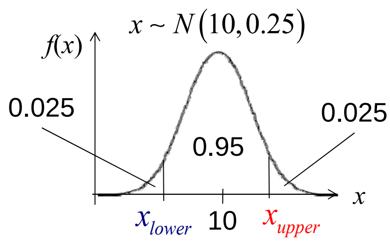
Example 1

The variance of a measurement is $\sigma^2 = 0.25 g^2$. The measurement is unbiased. We measure a 10 g etalon weight. In which range will be the measured weight with 95% probability?

The question „translated to maths”:

$$P\left(x_{lower} < x \le x_{upper}\right) = 0.95$$

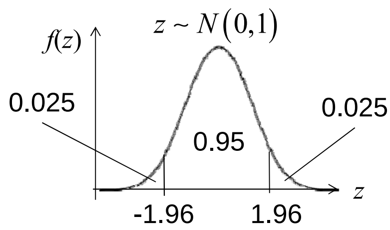
$$x_{lower} = \text{?} \qquad x_{upper} = \text{?}$$



This known from the distribution function (table):

$$P\left(-z_{\alpha/2} < z \le z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-1.96 < z \le 1.96\right) = 0.95$$



Question:

$$P\left(x_{lower} < x \le x_{upper}\right) = 0.95$$

Known:

$$P\left(-1.96 < z \le 1.96\right) = 0.95$$

Connection:

$$z = \frac{x - \mu}{\sigma}$$

$$\mu = 10$$

$$\sigma = \sqrt{0.25}$$

$$P\left(-1.96 < \frac{x - 10}{\sqrt{0.25}} \le 1.96\right) = 0.95$$

$$P\left(\overbrace{10 - 1.96 \cdot 0.5}^{x_{lower}} < x \le \overbrace{10 + 1.96 \cdot 0.5}^{x_{upper}}\right) = 0.95$$

$$\boxed{P\left(9.02 < x \le 10.98\right) = 0.95}$$

Interpreting the Results

Example 1

The variance of a measurement is σ^2 =0.25g ^{2}. The measurement is unbiased. We measure a 10 g etalon weight. In which range will be the measured weight with 95% probability?

$$\boxed{P\left(9.02 < x \le 10.98\right) = 0.95}$$

The measured weight will be between 9.02 and 10.98 with 95% probability.

Can we expect that the outcome will be between these limits 95 times out of 100? Exactly 95 times?

Generalization of Example 1

Question: $\boxed{P\left(x_{lower} < x \le x_{upper}\right) = 1-\alpha}$ $\boxed{x_{lower} = \textbf{?} \quad x_{upper} = \textbf{?}}$

Known: $P\left(-z_{\alpha/2} < z \le z_{\alpha/2}\right) = 1-\alpha$

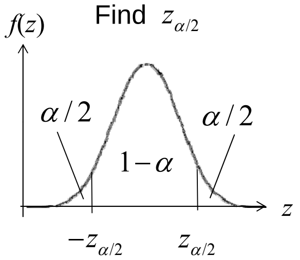
$1-\alpha$ is given

$$z = \frac{x-\mu}{\sigma}$$

$$P\left(-z_{\alpha/2} < \frac{x-\mu}{\sigma} \le z_{\alpha/2}\right) = 1-\alpha$$

Find $z_{\alpha/2}$

$$\boxed{P\left(\mu - z_{\alpha/2}\sigma < x \le \mu + z_{\alpha/2}\sigma\right) = 1-\alpha}$$



In which range will be the measured weight in Example 1 with 99% probability?

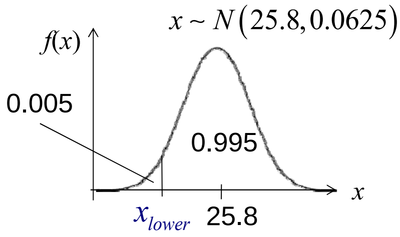
Example 2

The volume of gas bottles are normally distributed with 25.8 dm^3 expected value and 0.0625 (dm^3)^2 variance. What is that minimum volume, that 99.5% of the bottles exceed?

The question „translated to maths":

$$P\left(x_{lower} < x\right) = 0.995$$

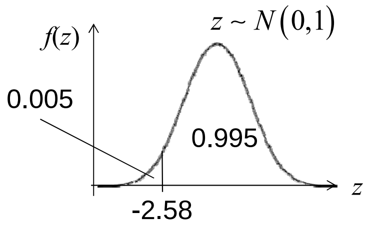
$$x_{lower} = \textbf{?}$$



This known from the distribution function (table):

$$P\left(-z_{\alpha} < z\right) = 1-\alpha$$

$$P\left(-2.58 < z\right) = 0.995$$



Question:

$$P\left(x_{lower} < x\right) = 0.95$$

Known:

$$P\left(-2.58 < z\right) = 0.995$$

Connection:

$$z = \frac{x-\mu}{\sigma}$$

$$\mu = 25.8$$

$$\sigma = \sqrt{0.0625}$$

$$P\left(-2.58 < \frac{x-25.8}{\sqrt{0.0625}}\right) = 0.995$$

$$P\left(\overbrace{25.8-2.58\cdot 0.25}^{x_{lower}} < x\right) = 0.995$$

$$\boxed{P\left(25.16 < x\right) = 0.995}$$

GENERALIZATION OF EXAMPLE 2

Question: $\boxed{P\left(x_{lower} < x\right) = 1-\alpha}$ $\boxed{x_{lower} = \textbf{?}}$ $1-\alpha$ is given

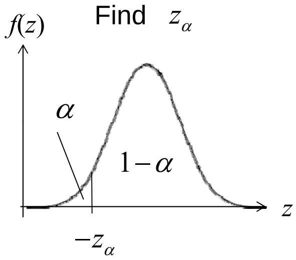
Known: $P\left(-z_{\alpha} < z\right) = 1-\alpha$

$$z = \frac{x-\mu}{\sigma}$$

$$P\left(-z_{\alpha} < \frac{x-\mu}{\sigma}\right) = 1-\alpha$$

$$\boxed{P\left(\mu - z_{\alpha}\sigma < x\right) = 1-\alpha}$$

Find z_{α}



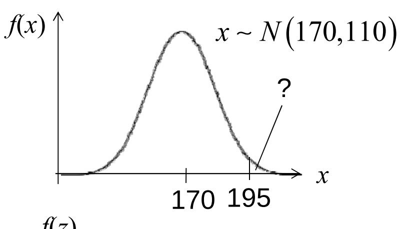
Give a 99% upper limit for the volume of the bottles in Example 2!

Example 3

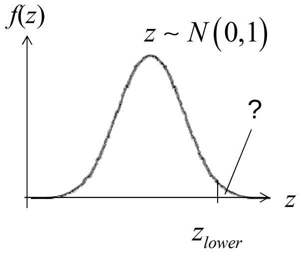
Harry is 204 cm tall and he is fond of tall girls (preferably above 195 cm). The average height of girls at his age is 170 cm with a variance of 110 cm². A random girl sits next to him in the library at the university. What is the probability that the girl meets his standard i.e. she is above 195 cm?

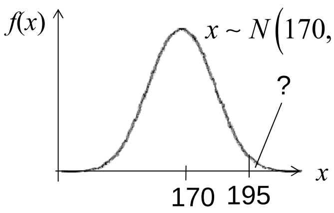
The question „translated to maths":

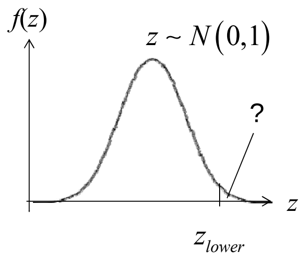
$$P\left(195 < x\right) = ?$$



If we knew z_{lower} we could read $P\left(z_{lower} < z\right)$ from the distribution function (table)







$$z = \frac{x - \mu}{\sigma}$$

$$z_{lower} = \frac{x_{lower} - \mu}{\sigma} = \frac{195 - 170}{\sqrt{110}} = 2.34$$

$$P(2.34 < z) = 0.00964 = P(195 < x)$$

Interpreting the Results

Example 3

Harry is 204 cm tall and he is fond of tall girls (preferably above 195 cm). The average height of girls at his age is 170 cm with a variance of 110 cm^2. A random girl sits next to him in the library at the university. What is the probability that the girl meets his standard i.e. she is above 195 cm?

$$\boxed{P(195 < x) = 0.00964}$$

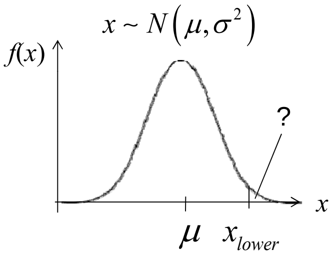
Harry doesn't know a thing about statistics and probability. How would you explain this result to him?

What assumptions were made during this calculation? What if they are not true?

Generalization of Example 3

Question: $\boxed{P\left(x_{lower} < x\right) = ?}$

x_{lower} is given



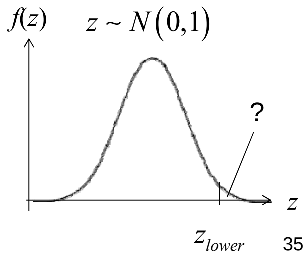
The question is transformed to z-scale:

$$P\left(\frac{x_{lower}-\mu}{\sigma} < \frac{x-\mu}{\sigma}\right) = ?$$

$$z = \frac{x-\mu}{\sigma}$$

$$P\left(z_{lower} < z\right) = ?$$

$$\boxed{P\left(x_{lower} < x\right) = P\left(\frac{x_{lower}-\mu}{\sigma} < z\right)}$$



How many percent of he girls are between 170 and 185 cm?

Examples

1. The variance of a measurement is $\sigma^2 = 4g^2$. The bias of the measurement is 2 g. We measure an object, its weight is 200 g.
a) In which range will be the outcome of the measurement with 99% probability?
b) What is the probability that the measured weight is above 205g?
c) What is the probability that the measured weight is below 200g?
d) Give a 90% upper limit for the measured weight! (The value that the measured weight will not achieve with 90% probability.)

2. The volume of gas bottles are normally distributed with 25.8 dm^3 expected value and 0.0625 (dm^3)^2 variance.
a) How many percent of the bottles will be in the 25.8±0.3 dm^3 interval?
b) In what interval will be 99% of the bottles?

THE SAMPLE MEAN

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum x_i$$

$$E(\bar{x}) = \frac{1}{n}[nE(x)] = E(x) = \mu \qquad \sigma_{\bar{x}}^2 = Var(\bar{x}) = \frac{Var(x)}{n} = \frac{\sigma_x^2}{n}$$

$$x \sim N\left(\mu, \sigma^2\right) \qquad z = \frac{x-\mu}{\sigma}$$

$$\bar{x} \sim N\left(\mu, \sigma^2 / n\right) \qquad z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

Sampling distribution: probability distribution of a statistic (e.g. sample mean or sample variance)

$N\left(\mu, \sigma^2 / n\right)$ is the sampling distribution of the sample mean

CENTRAL LIMIT THEOREM

The mean of sample elements taken from any distribution approximately follows Gauss distribution around the expected value of the original distribution with variance σ^2/n. Where n is the sample size.

Sum as well $\sum_{i=1}^{N} x_i \sim N\left(n\mu, n\sigma^2\right)$

Based on the Central Limit Theorem: $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

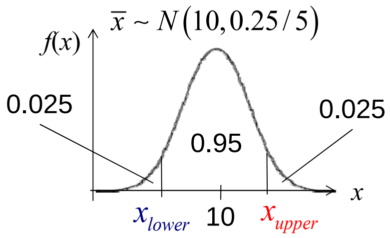
Example 4

The variance of a measurement is σ^2 =0.25g^2. The measurement is unbiased. We measure a 10 g etalon weight with five repetition. In which range will be the measured weight with 95% probability?

The question „translated to maths”:

$$P\left(\bar{x}_{lower} < \bar{x} \le \bar{x}_{upper}\right) = 0.95$$

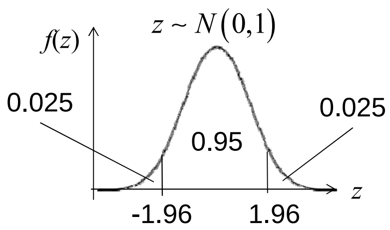
x_{lower} = **?** x_{upper} = **?**



This known from the distribution function (table):

$$P\left(-z_{\alpha/2} < z \le z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-1.96 < z \le 1.96\right) = 0.95$$



Question:

$$P\left(\bar{x}_{lower} < \bar{x} \le \bar{x}_{upper}\right) = 0.95$$

Known:

$$P\left(-1.96 < z \le 1.96\right) = 0.95$$

Connection:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$\mu = 10$$
$$\sigma = \sqrt{0.25}$$
$$n = 5$$

$$P\left(-1.96 < \frac{\bar{x} - 10}{\sqrt{0.25} / \sqrt{5}} \le 1.96\right) = 0.95$$

$$P\left(\overbrace{10 - 1.96 \cdot \sqrt{0.5}}^{\bar{x}_{lower}} < \bar{x} \le \overbrace{10 + 1.96 \cdot \sqrt{0.5}}^{\bar{x}_{upper}}\right) = 0.95$$

$$\boxed{P\left(9.56 < \bar{x} \le 10.44\right) = 0.95}$$

INTERPRETING THE RESULTS

Example 1
The variance of a measurement is $\sigma^2 = 0.25g^2$. The measurement is unbiased. We measure a 10 g etalon weight. In which range will be the measured weight with 95% probability?

$$P(9.02 < x \le 10.98) = 0.95$$

Example 4
The variance of a measurement is $\sigma^2 = 0.25g^2$. The measurement is unbiased. We measure a 10 g etalon weight with five repetition. In which range will be the measured weight with 95% probability?

$$P(9.56 < \bar{x} \le 10.44) = 0.95$$

In general: $\boxed{P\left(\mu - z_{\alpha/2}\sigma/\sqrt{n} < \bar{x} \le \mu + z_{\alpha/2}\sigma/\sqrt{n}\right) = 1-\alpha}$

Example 5
The $CaNO_3$ content of a water sample is measured with four repetitions. The results: 12.3, 13.5, 13.4, 12.9 mg/l. The variance of a measurement is $\sigma^2 = 0.3\ (mg/l)^2$.
a) Give an estimation for the true $CaNO_3$ concentration.
b) In which range could be the true concentration with 95% probability?

a) $\hat{\mu} = \bar{x} = 13.03$

b) Question: $P\left(\mu_{lower} < \mu \le \mu_{upper}\right) = 1-\alpha$

Known: $P\left(-z_{\alpha/2} < z \le z_{\alpha/2}\right) = 1-\alpha$

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

$$P\left(-z_{\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha$$

$$\boxed{P\left(\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} < \mu \le \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}\right) = 1-\alpha}$$

Example 5

$$\bar{x} = 13.03 \quad \sigma^2 = 0.3 \quad n = 4 \quad z_{0.025} = 1.96 \quad \alpha = 0.05$$

$$\boxed{P\left(\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} < \mu \le \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}\right) = 1-\alpha}$$

$$P\left(13.03 - 1.96\sqrt{0.3}/\sqrt{4} < \mu \le 13.03 + 1.96\sqrt{0.3}/\sqrt{4}\right) = 0.95$$

$$P\left(12.49 < \mu \le 13.56\right) = 0.95$$

95% confidence interval for the expected value

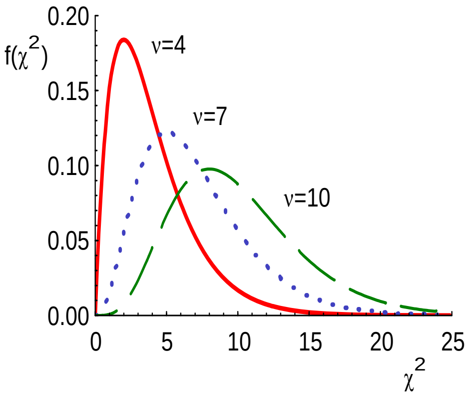
c) Give a 95% upper confidence limit for the true $CaNO_3$ concentration!

A confidence interval bounds population or distribution parameters.

Interpretation

If the 95% confidence interval is calculated for the expected value from 100 different sample, than approximately 95 interval contains the true expected value out of the 100.

CHI-SQUARE DISTRIBUTION



$$\chi^2 = \sum_{i=1}^{\nu} z_i^2$$

$$E\left(\chi^2\right) = \nu$$

$$Var\left(\chi^2\right) = 2\nu$$

ν is the degrees of freedom

$$s^2 = \frac{\chi^2\sigma^2}{\nu}$$

Fischer-Cochran Theorem

$$\chi^2 = \sum_{i=1}^{\nu} z_i^2 = Q_1 + Q_2 + \ldots + Q_k$$

$\chi^2 = \sum_{i=1}^{\nu} z_i^2$ is chi-squared distributed, *df*=ν

χ^2 equals to the sum of Q_i-s, where each Q_i is a sum of squares of linear combinations of the z-s. The df of each Q_i is ν_i.

If, and only if $\nu = \nu_1 + \nu_2 + \ldots + \nu_k$, than each Q_i is chi-squared distributed, with *df*=ν_i

$$\nu = \nu_1 + \nu_2 + \ldots + \nu_k \iff Q_i \sim \chi^2, \quad df = \nu_i \quad \text{for each } i$$

(Sampling) Distribution of Sample Variance

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2 \qquad\qquad z = \frac{x - \mu}{\sigma}$$

$$\chi^2 \sigma^2 = \sum_{i=1}^{n} \left(x_i - \mu \right)^2 = \sum_{i=1}^{n} \left((x_i - \bar{x}) + (\bar{x} - \mu) \right)^2$$

$$= \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 + \underbrace{2 \sum_{i=1}^{n} (x_i - \bar{x})(\bar{x} - \mu)}_{2(\bar{x} - \mu) \sum_{i=1}^{n} (x_i - \bar{x}) = 0} + n \left(\bar{x} - \mu \right)^2$$

$$\chi^2 \sigma^2 = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 + n \left(\bar{x} - \mu \right)^2$$

n $\qquad$ $n-1$ $\qquad$ 1 $\qquad$ ν (degrees of freedom)

(Sampling) Distribution of Sample Variance

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

$$\nu:\quad n \quad = \quad n-1 \quad + \quad 1 \quad \longrightarrow \quad \sum_{i=1}^{n}(x_i-\bar{x})^2 \sim \chi^2\sigma^2,\quad \nu = n-1$$

Fisher-Cochran theorem

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \sim \frac{\chi^2\sigma^2}{n-1},\quad \nu = n-1$$

$$\boxed{s^2 = \frac{\chi^2\sigma^2}{\nu},\quad \nu = n-1}$$

$$E\left[s^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \frac{1}{n-1}E\left[\chi^2\sigma^2\right]$$

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

$$\nu = n-1$$

$$\boxed{E\left[s^2\right] = \sigma^2}$$

Example 6
The variance of a scale is σ^2 =0.25g^2. In which range will be the sample variance of five repetition?

The question „translated to maths”:

$$P\left(s^2_{lower} < s^2 \le s^2_{upper}\right) = 0.95$$

$$s^2_{lower} = \textbf{?} \qquad s^2_{upper} = \textbf{?}$$

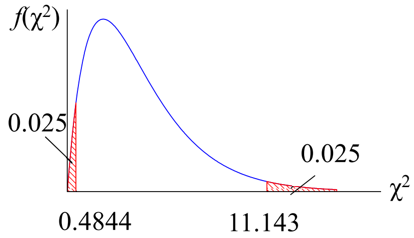
This known from the distribution function (table):

$$P\left(\chi^2_{\text{lower}} < \chi^2 \le \chi^2_{\text{upper}}\right) = 0.95$$

$$P\left(\chi^2_{\text{lower}} < \frac{s^2\nu}{\sigma^2} \le \chi^2_{\text{upper}}\right) = 0.95$$

$$s^2 = \frac{\chi^2\sigma^2}{\nu} \longrightarrow \frac{s^2\nu}{\sigma^2} = \chi^2$$

$$P\left(\frac{\chi^2_{\text{lower}}\sigma^2}{\nu} < s^2 \le \frac{\chi^2_{\text{upper}}\sigma^2}{\nu}\right) = 0.95$$



$$\chi^2_{\text{lower}} = 0.4844$$

$$\chi^2_{\text{upper}} = 11.143$$

$$\nu = 4$$

$$\boxed{P\left(\frac{\chi^2_{\text{lower}}\sigma^2}{\nu} < s^2 \le \frac{\chi^2_{\text{upper}}\sigma^2}{\nu}\right) = 0.95}$$

$$P\left(0.033 < s^2 \le 0.696\right) = 0.95$$

Give a 95% lower limit for the sample variance!

Example 7
Results of 10 measurements:
24.46; 23.93; 25.79; 25.17; 23.82; 25.39; 26.54; 23.85; 24.19; 25.50.
Give the 99% confidence interval for the variance!

The question „translated to maths":

$$P\left(\sigma^2_{lower} < s^2 \le \sigma^2_{upper}\right) = 1-\alpha$$

$$\sigma^2_{lower} = \textbf{?} \qquad \sigma^2_{upper} = \textbf{?}$$

$$s^2 = \frac{\chi^2\sigma^2}{\nu} \longrightarrow \frac{s^2\nu}{\sigma^2} = \chi^2$$

This known from the distribution function (table):

$$P\left(\chi^2_{\text{lower}} < \chi^2 \le \chi^2_{\text{upper}}\right) = 1-\alpha$$

$$P\left(\chi^2_{\text{lower}} < \frac{s^2\nu}{\sigma^2} \le \chi^2_{\text{upper}}\right) = 1-\alpha$$

$$P\left(\frac{s^2\nu}{\chi^2_{\text{lower}}} > \sigma^2 \ge \frac{s^2\nu}{\chi^2_{\text{upper}}}\right) = 1-\alpha$$

! Inequality sign is reversed !

Example 7

$$\bar{x} = 24.878 \quad s^2 = 0.884 \quad n = 10 \quad \alpha = 0.01$$

$$\chi^2_{\text{lower}} = \chi^2_{0.995}(9) = 1.73 \quad \chi^2_{\text{upper}} = \chi^2_{0.005}(9) = 23.59 \quad \nu = 9$$

$$P\left(\frac{s^2\nu}{\chi^2_{\text{lower}}} > \sigma^2 \geq \frac{s^2\nu}{\chi^2_{\text{upper}}}\right) = 1 - \alpha$$

$$P\left(\frac{0.884 \cdot 9}{1.73} > \sigma^2 \geq \frac{0.884 \cdot 9}{23.59}\right) = 0.99$$

$$P\left(0.337 > \sigma^2 \geq 4.598\right) = 0.99$$

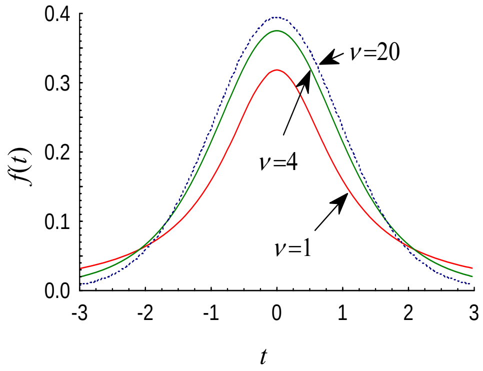
Give a 99% upper confidence limit for the variance!

STUDENT'S *t* DISTRIBUTION

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

$$E(t) = 0$$



Example 8

The $CaNO_3$ content of a water sample is measured with four repetitions. The results: 12.3, 13.5, 13.4, 12.9 mg/l.

a) Give an estimation for the true $CaNO_3$ concentration.

b) In which range could be the true concentration with 95% probability?

a) $\hat{\mu} = \overline{x} = 13.03$

b) Question: $P\left(\mu_{lower} < \mu \le \mu_{upper}\right) = 1 - \alpha$

Known: $P\left(-t_{\alpha/2} < t \le t_{\alpha/2}\right) = 1 - \alpha$

$$t = \frac{\overline{x} - \mu}{s / \sqrt{n}}$$

$$P\left(-t_{\alpha/2} < \frac{\overline{x} - \mu}{s / \sqrt{n}} \le t_{\alpha/2}\right) = 1 - \alpha$$

$$\boxed{P\left(\overline{x} - t_{\alpha/2} s / \sqrt{n} < \mu \le \overline{x} + t_{\alpha/2} s / \sqrt{n}\right) = 1 - \alpha}$$

Example 8

$\overline{x} = 13.03 \quad s^2 = 0.3 \quad n = 4 \quad z_{0.025} = 3.182 \quad \alpha = 0.05 \quad \nu = 3$

$$\boxed{P\left(\overline{x} - t_{\alpha/2} s / \sqrt{n} < \mu \le \overline{x} + t_{\alpha/2} s / \sqrt{n}\right) = 1 - \alpha}$$

$$P\left(13.03 - 3.182\sqrt{0.3}/\sqrt{4} < \mu \le 13.03 + 3.182\sqrt{0.3}/\sqrt{4}\right) = 0.95$$

$P\left(12.15 < \mu \le 13.90\right) = 0.95$ 95% confidence interval for the expected value

c) Give a 95% upper confidence limit for the true $CaNO_3$ concentration!

Compare the results with hose of Example 5!

F DISTRIBUTION

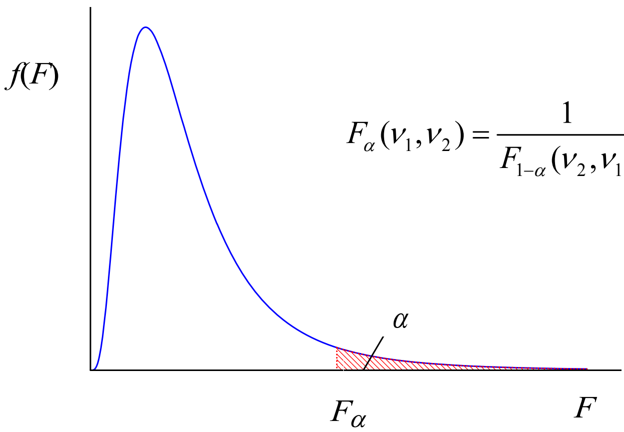
$$F = \frac{\chi_1^2}{\nu_1} \bigg/ \frac{\chi_2^2}{\nu_2}$$

Two parameters:
ν_1 is the degrees of freedom for the numerator, ν_2 is for the denominator

$$s^2 = \frac{\chi^2 \sigma^2}{\nu} \longrightarrow F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

$$\boxed{\begin{array}{l} \text{if} \quad \sigma_1^2 = \sigma_2^2 \\ F = s_1^2 / s_2^2 \end{array}}$$

CRITICAL VALUES FOR THE *F* DISTRIBUTION



$$F_\alpha(\nu_1, \nu_2) = \frac{1}{F_{1-\alpha}(\nu_2, \nu_1)}$$

Example 9

Two sets of measurements (4 and 7 repetitions) were performed using the same method and device.

Give the 90% probability interval for the ratio of sample variances.

$$\nu_1 = 3 \quad \nu_2 = 6$$

The population variances are equal: $\sigma_1^2 = \sigma_2^2$

$$P\left(F_{\text{lower}} < s_1^2 / s_2^2 \le F_{\text{upper}}\right) = P\left(F_{0.95} < s_1^2 / s_2^2 \le F_{0.05}\right) = 0.90$$

$$P\left(0.164 < s_1^2 / s_2^2 \le 4.12\right) = 0.90$$