Statistics

prof. Marek Kręglewski

Course content

Describing the data (types of data, graphical tools)

Probability, expectation values

Probability distributions

The binomial distribution

The Poisson distribution

The Gaussian distribution

Sampling distributions and estimation (central limit theorem, standard error of the mean)

Student's t distribution (confidence intervals, determining sample size)

Hypothesis testing. One-sample hypothesis tests of the mean (two-sided and one-sided tests)

Two-sample hypothesis tests of the mean

Hypothesis tests of variance (one-sample test and two-sample test)

The F distribution. Chi-square (χ^2) distribution.

The analysis of variance (ANOVA).

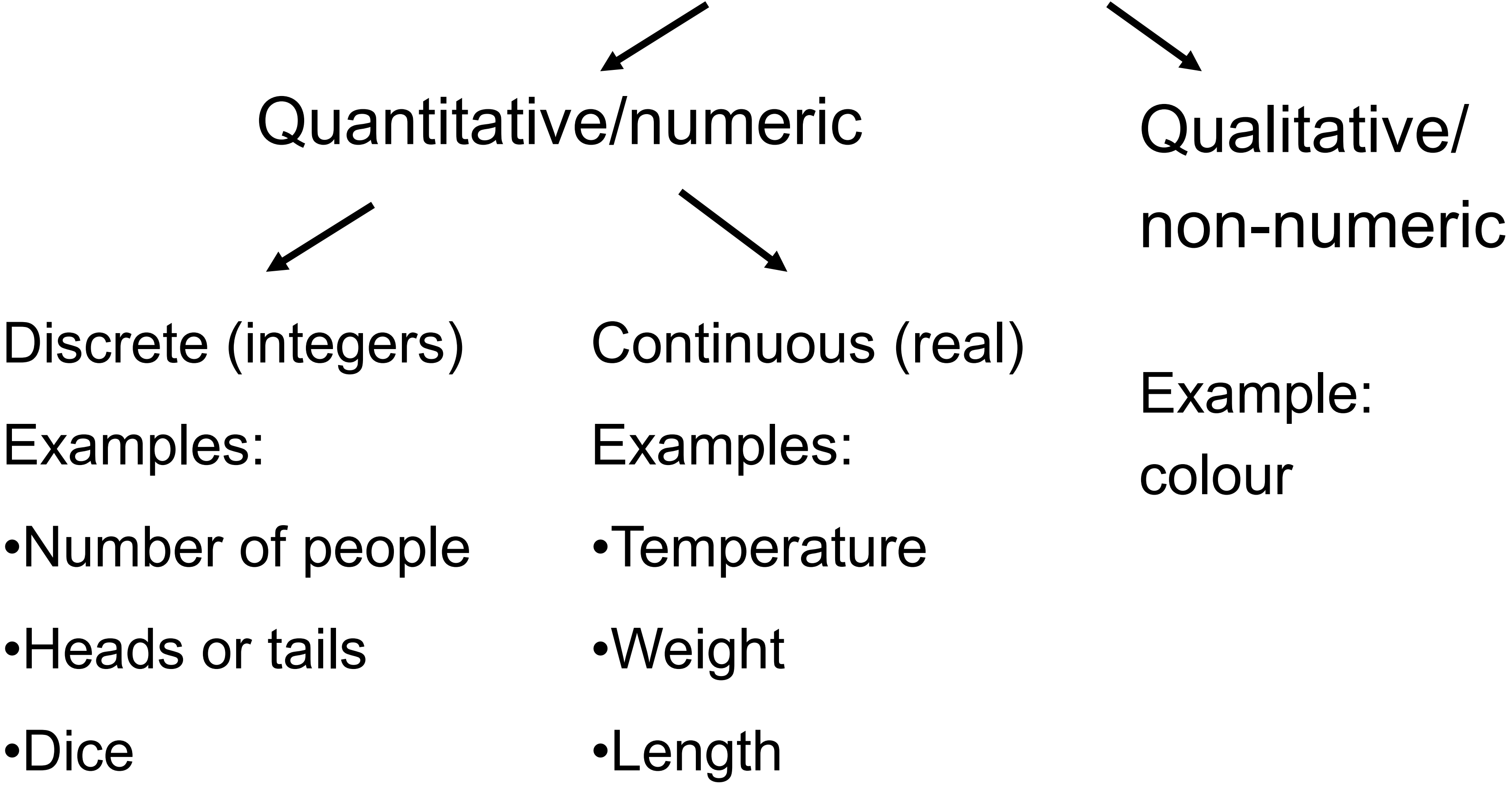
Linear regression analysis (the straight line fit, covariance, correlation)

Polynomial regression

Basic definitions

- Statistics – study of ensembles of data
- Object of statistical analysis – observation, event in relation value ↔ frequency (distribution)
- Population – all data
- Sample of the size n – n observations
- Goal of the statistical analysis – relation between sample and population

Types of Data



Quantitative/numeric

Discrete (integers)

Examples:

- Number of people
- Heads or tails
- Dice

Continuous (real)

Examples:

- Temperature
- Weight
- Length

Qualitative/
non-numeric

Example:
colour

Probability = P(A)

A, B are events from the population Ω

Properties of probability

- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$
- If A and B exclude each other, then

  $P(A \text{ or } B) = P(A) + P(B)$
- If A and B do not exclude each other, then

  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Calculation of the probability: $P(A) = n_A/n$

Where n_A – number of events A

n – total number of events

Simple distribution

- Throwing the dice

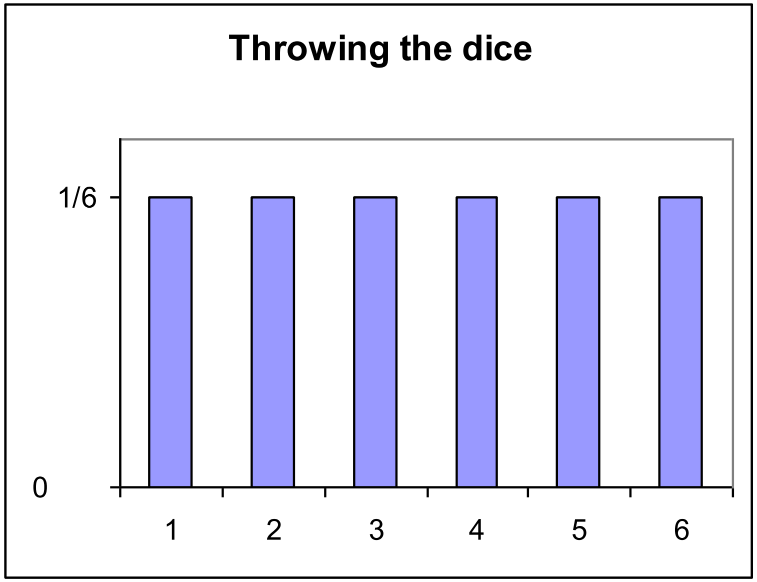
P(1)=P(2)=P(3)=P(4)=P(5)=P(6)= 1/6

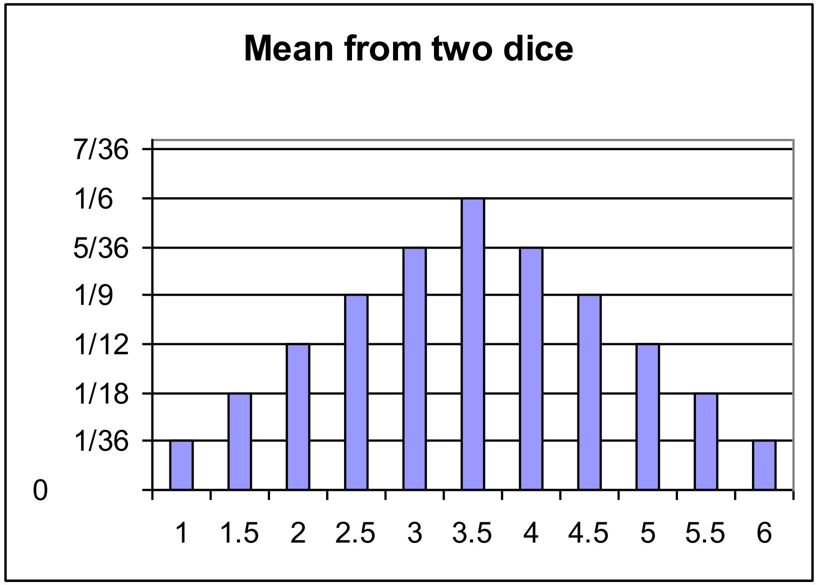
The Arithmetic Mean: $$\bar{x} = \sum_{k=1}^{n} P_k x_k$$

The Variance: $$\sigma^2 = \sum_{k=1}^{n} P_k \left(x_k - \bar{x}\right)^2$$

The Standard Deviation: $$\sigma = \sqrt{\sigma^2}$$

Experiments with dice





How do the arithmetic mean and variance differ for both distributions?

Dice - calculation

Single dice

$$\bar{x} = \sum_{k=1}^{n} P_k x_k \qquad \bar{x} = \tfrac{1}{6}1 + \tfrac{1}{6}2 + \tfrac{1}{6}3 + \tfrac{1}{6}4 + \tfrac{1}{6}5 + \tfrac{1}{6}6 = \tfrac{7}{2} = 3.5$$

$$\sigma^2 = \sum_{k=1}^{n} P_k \left(x_k - \bar{x}\right)^2$$

$$\sigma^2 = \tfrac{1}{6}\left(1 - \tfrac{7}{2}\right)^2 + \tfrac{1}{6}\left(2 - \tfrac{7}{2}\right)^2 + \tfrac{1}{6}\left(3 - \tfrac{7}{2}\right)^2 + \tfrac{1}{6}\left(4 - \tfrac{7}{2}\right)^2 + \tfrac{1}{6}\left(5 - \tfrac{7}{2}\right)^2 + \tfrac{1}{6}\left(6 - \tfrac{7}{2}\right)^2 =$$

$$= \tfrac{1}{6}\left[\left(-\tfrac{5}{2}\right)^2 + \left(-\tfrac{3}{2}\right)^2 + \left(-\tfrac{1}{2}\right)^2 + \left(\tfrac{1}{2}\right)^2 + \left(\tfrac{3}{2}\right)^2 + \left(\tfrac{5}{2}\right)^2\right] = \tfrac{35}{12}$$

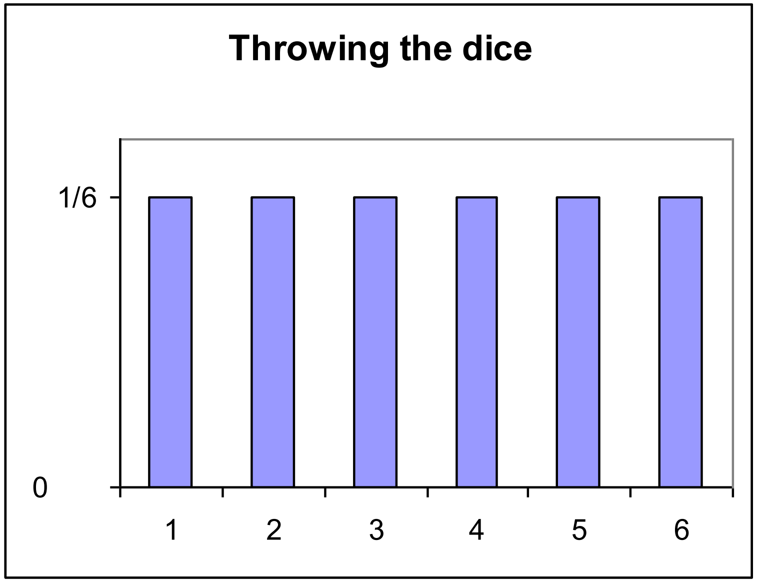
Two dice

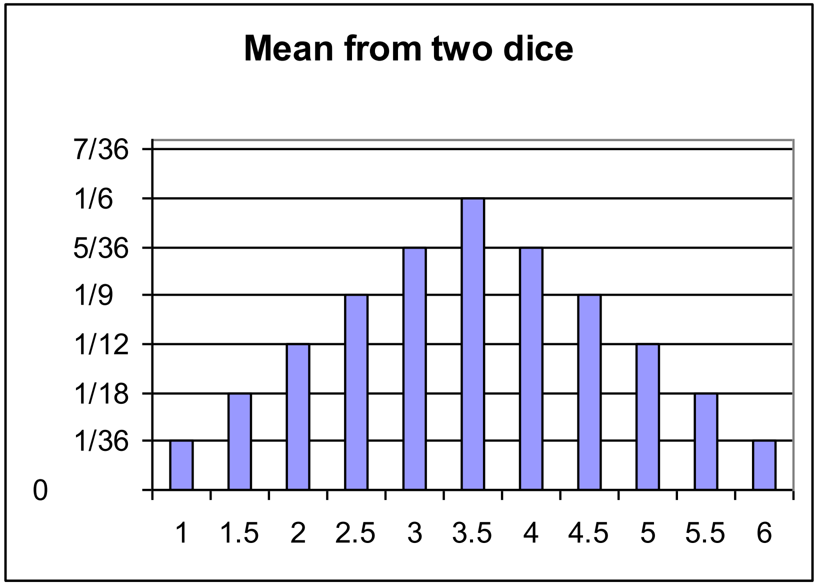
$$\bar{x} = \sum_{k=1}^{n} P_k x_k \qquad \bar{x} = \tfrac{1}{36}1 + \tfrac{2}{36}1.5 + \tfrac{3}{36}2 + \ldots + \tfrac{6}{36}3.5 + \ldots + \tfrac{1}{36}6 = \tfrac{7}{2} = 3.5$$

$$\sigma^2 = \sum_{k=1}^{n} P_k \left(x_k - \bar{x}\right)^2$$

$$\sigma^2 = \tfrac{1}{36}\left(1 - \tfrac{7}{2}\right)^2 + \tfrac{2}{36}\left(1.5 - \tfrac{7}{2}\right)^2 + \tfrac{3}{36}\left(2 - \tfrac{7}{2}\right)^2 + \ldots + \tfrac{6}{36}\left(3.5 - \tfrac{7}{2}\right)^2 + \ldots + \tfrac{1}{36}\left(6 - \tfrac{7}{2}\right)^2 = \tfrac{35}{24}$$

Experiments with dice





$$\bar{x} = {}^{7}\!/_{2} = 3.5$$

$$\sigma^2 = {}^{35}\!/_{12}$$

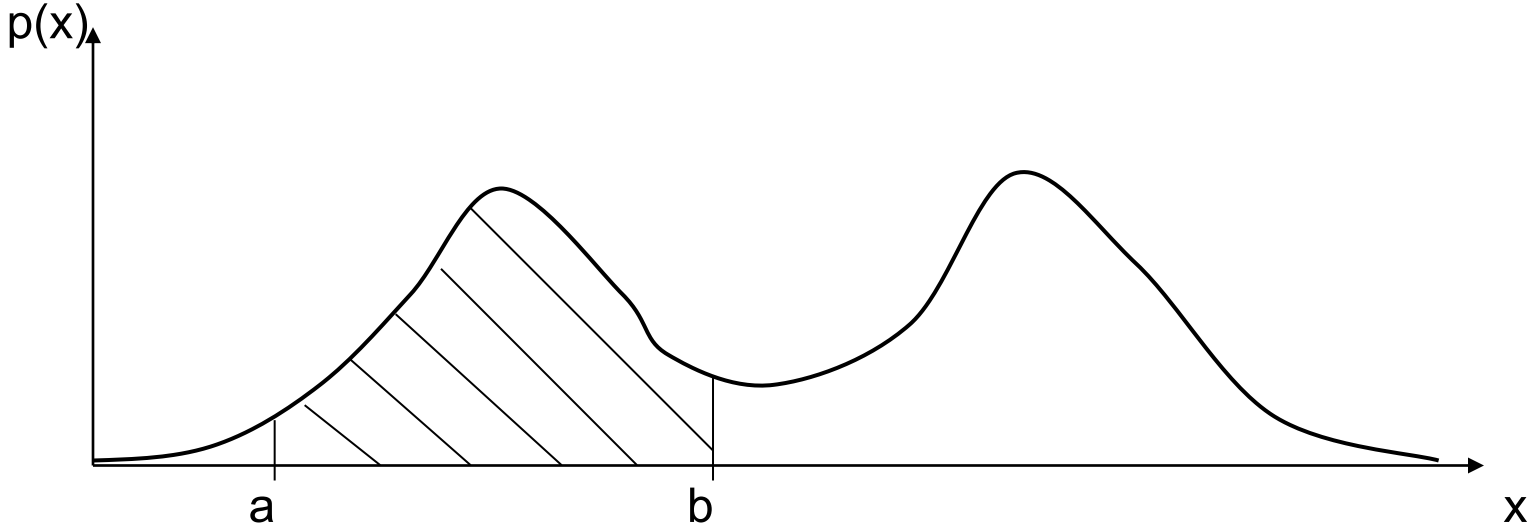
$$\sigma = 1.7078$$

$$\bar{x} = {}^{7}\!/_{2} = 3.5$$

$$\sigma^2 = {}^{35}\!/_{24}$$

$$\sigma = 1.2076$$

Probability distribution for continuous variable



$$p(x) - \textit{probability density distribution}$$

$$P(a \langle x \langle b) = \int_a^b p(x) dx$$

$$P(-\infty \langle x \langle \infty) = \int_{-\infty}^{\infty} p(x) dx = 1$$

Heads and tails

Definition: P(r)= probability of r heads (H)

1. Tossing one coin: P(0)=P(1)=½
2. Tossing four coins: P(0)=P(4)=(½)^4= $\frac{1}{16}$ TTTT or HHHH

TTTH, TTHT, THTT, HTTT P(1)=P(3)= $\frac{4}{16}=\frac{1}{4}$

TTHH,THHT,HHTT,THTH,HTHT,HTTH P(2)= $\frac{6}{16}=\frac{3}{8}$

$$\sum_r P(r)=P(0)+P(1)+P(2)+P(3)+P(4)=\frac{1}{16}+\frac{4}{16}+\frac{6}{16}+\frac{4}{16}+\frac{1}{16}=\frac{16}{16}=1$$

r=	0	1	2	3	4
16 tosses					
theory	1	4	6	4	1
data	1	4	2	7	2
160 tosses					
theory	10	40	60	40	10
data	13	36	61	40	10
1600 tosses					
theory	100	400	600	400	100
data	96	409	577	403	115

Expectation values

$$average(expected)\,number\,of\ heads \quad \langle r \rangle = \sum_r rP(r)$$

$$four\,coins \quad \langle r \rangle = 0 \times \tfrac{1}{16} + 1 \times \tfrac{4}{16} + 2 \times \tfrac{6}{16} + 3 \times \tfrac{4}{16} + 4 \times \tfrac{1}{16} = 2$$

Expectation value of a function f $\quad \langle f \rangle = \sum_r f(r)P(r)$

Law of large numbers

For a data sample of size N the mean over the sample

$$\bar{f} \xrightarrow{N \to \infty} \langle f \rangle$$

The Binomial Distribution

1. A process with two possible outcomes
2. *p* – probability of a success

   (1-*p*) – probability of failure

   n – number of trials

   r – number of successes in *n* trials

   (*n*-*r*) – number of failures in *n* trials
3. Probability of *r* consecutive successes and, then, of (*n*-*r*) failures

$$p^r\,(1-p)^{n-r}$$

4. Number of different sequences of *r* successes in *n* trials $\binom{n}{r} = \frac{n!}{r!\,(n-r)!}$

5. Probability of *r* successes in *n* trials in any order

$$P(r;p,n) = p^r\,(1-p)^{n-r}\,\frac{n!}{r!\,(n-r)!}$$

The Binomial Distribution

Probability of 0 to *n* successes in *n* trials

$$\sum_{r=0}^{n} p^r (1-p)^{n-r} \frac{n!}{r!\,(n-r)!} = \sum_{r=0}^{n} \binom{n}{r} p^r (1-p)^{n-r} = [p + (1-p)]^n = 1$$

The mean number of successes is

$$\langle r \rangle = \sum_{r=0}^{n} r\, P(r;p,n) = \sum_{r=0}^{n} r p^r (1-p)^{n-r} \frac{n!}{r!\,(n-r)!}$$

Take out a factor of *np* and drop the *r*=0 term

$$\langle r \rangle = np \sum_{r=1}^{n} p^{r-1} (1-p)^{n-r} \frac{(n-1)!}{(r-1)!\,(n-r)!}$$

Substituting *r’*=*r*-1, *n’*=*n*-1

$$\langle r \rangle = np \sum_{r'=0}^{n'} p^{r'} (1-p)^{n'-r'} \frac{n'!}{r'!\,(n'-r')!} = np$$

The Binomial Distribution

The variance is $V(r) = \sum_{r=0}^{n} (r - \langle r \rangle)^2 P(r;p,n) = \langle (r - \langle r \rangle)^2 \rangle = \langle r^2 \rangle - \langle r \rangle^2$

Start with the expression

$$\langle r(r-1) \rangle = \sum_{r=0}^{n} r(r-1) p^r (1-p)^{n-r} \frac{n!}{r!\,(n-r)!}$$

Substitute *r'*=*r*-2, *n'*=*n*-2

$$\langle r(r-1) \rangle = p^2 n(n-1) \sum_{r'=0}^{n'} p^{r'} (1-p)^{n'-r'} \frac{n'!}{r'!\,(n'-r')!} = n(n-1)p^2$$

$$\langle r^2 \rangle - \langle r \rangle^2 = \langle r(r-1) \rangle + \langle r \rangle - \langle r \rangle^2 = n(n-1)p^2 + np - (np)^2$$

Variance $V(r) = np(1-p)$

Standard deviation $\sigma = \sqrt{np(1-p)}$

The Binomial Distribution - example

Guessing cards: ♥A, ♥2, ♥3, ♥4, ♥5

What is a probability of guessing more than 3 times in 6 trials?

Probability of guessing a card in a single trial:

p=0.2

Four correct guesses:

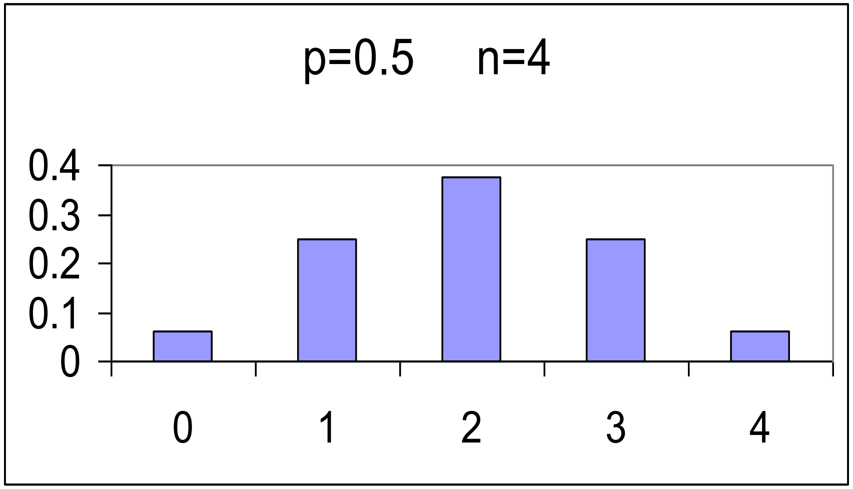
$$P(4; 0.2,6) = 0.2^4(1-0.2)^{6-4}\frac{6!}{4!\,(6-4)!} = 0.015360$$

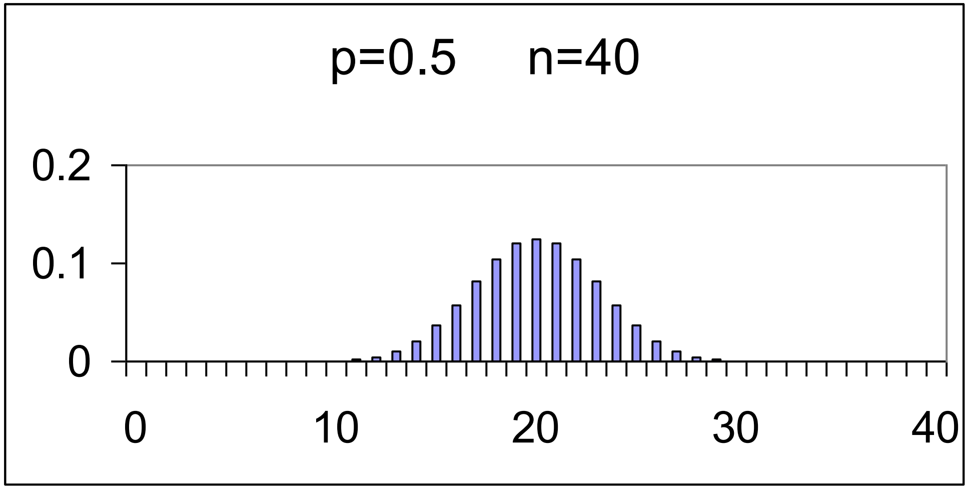
Final result:

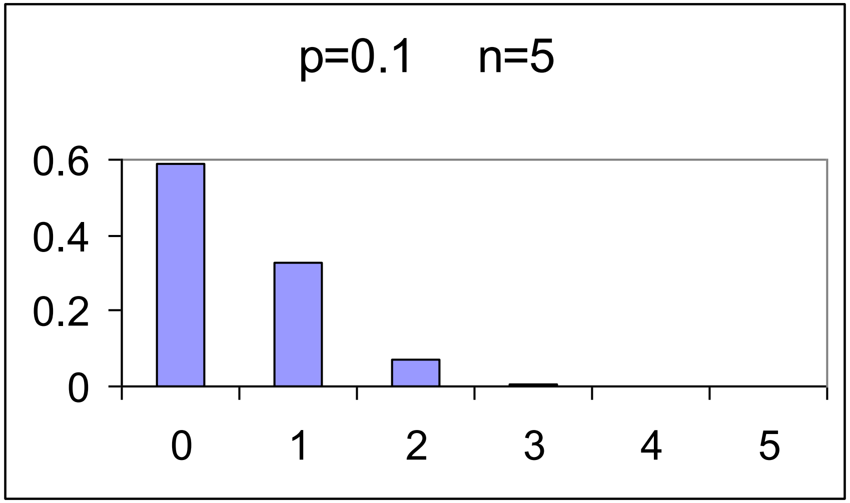
P(4; 0.2, 6) + P(5; 0.2, 6) + P(6; 0.2, 6)=0.015360+0.001536+0.000064=

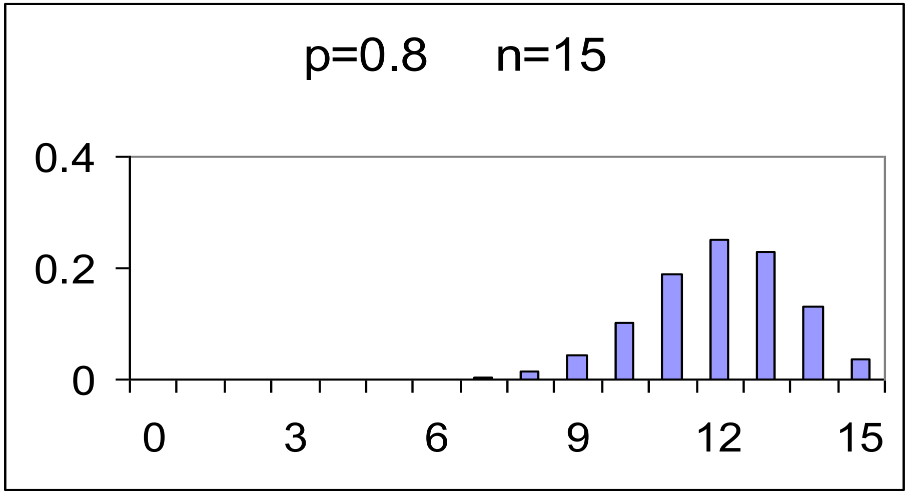
=0.016960=1.7%

Some binomial distributions


p=0.5 n=4


p=0.5 n=40


p=0.1 n=5


p=0.8 n=15

The Poisson Distribution

1. A process where particular outcomes occur in a certain number of trials,

   „sharp independent events occurring in a continuum” , e.g. flashes of lightning during the thunderstorm

2. *λ* – average number of events in some interval

   n – number of sections in the interval

   p = *λ/n* probability that a given section contains an event

   (probability that a given event contains 2 event must be zero)

   Probability of *r* events in n sections:

$$P(r;\lambda/n\,,n) = \frac{\lambda^r}{n^r}\left(1-\frac{\lambda}{n}\right)^{n-r}\frac{n!}{r!\,(n-r)!}$$

$$\frac{n!}{(n-r)!} = n(n-1)(n-2)\ldots(n-r+1)\xrightarrow{n\to\infty} n^r$$

$$\left(1-\frac{\lambda}{n}\right)^{n-r}\xrightarrow{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n}\to e^{-\lambda}$$

The Poisson Distribution

Probability of *r* events in an interval if the mean expected number is λ:

$$P(r;\lambda) = \frac{e^{-\lambda}\lambda^r}{r!}$$

Important properties:

The total probability is $$\sum_{r=0}^{\infty} P(r;\lambda) = e^{-\lambda}\sum_{r=0}^{\infty}\frac{\lambda^r}{r!} = e^{-\lambda}e^{\lambda} = 1$$

The mean number of events $$\langle r\rangle = \lambda$$

The variance $$V(r) = \lambda$$

The Poisson Distribution – a proof of properties

The mean number of events

$$\langle r\rangle = \sum_{r=0}^{\infty} r e^{-\lambda}\frac{\lambda^r}{r!} = \lambda e^{-\lambda}\sum_{r=1}^{\infty}\frac{\lambda^{r-1}}{(r-1)!}$$

Set *r'*=*r*-1

$$\langle r\rangle = \lambda e^{-\lambda}\sum_{r'=0}^{\infty}\frac{\lambda^{r'}}{r'!} = \lambda e^{-\lambda}e^{\lambda} = \lambda$$

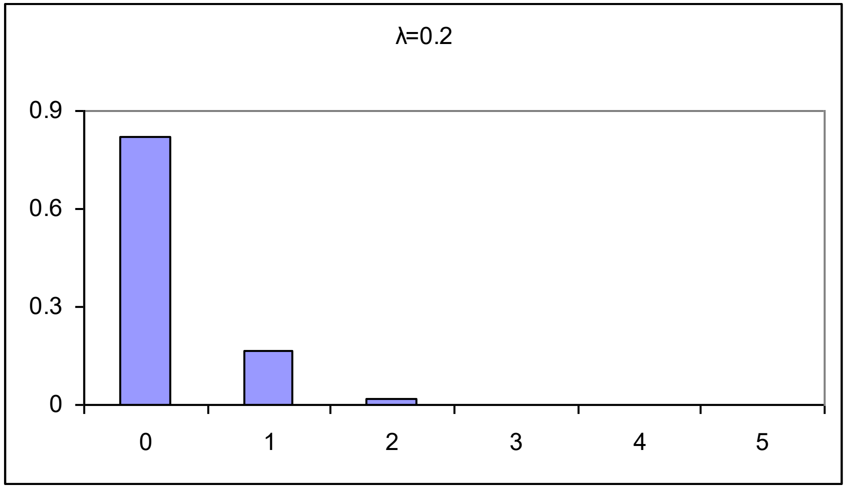
The variance

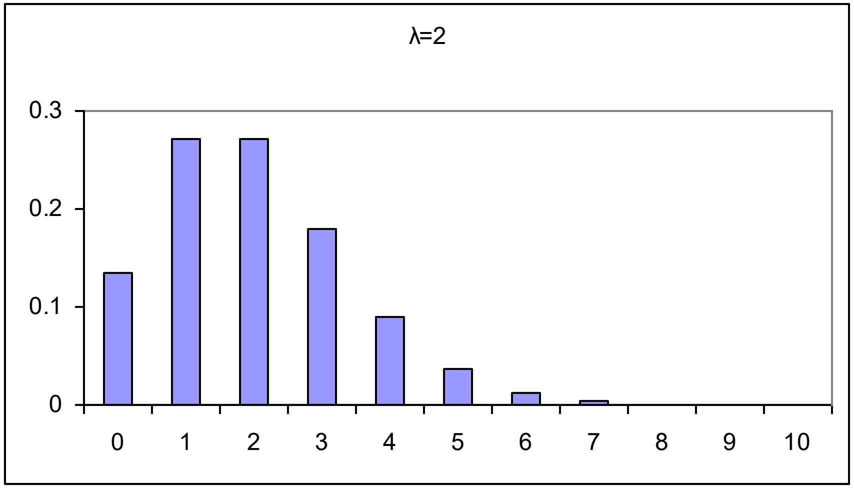
$$V(r) = \sum_{r=0}^{\infty}(r-\langle r\rangle)^2 P(r;\lambda) = \langle (r-\langle r\rangle)^2\rangle = \langle r^2\rangle - \langle r\rangle^2$$

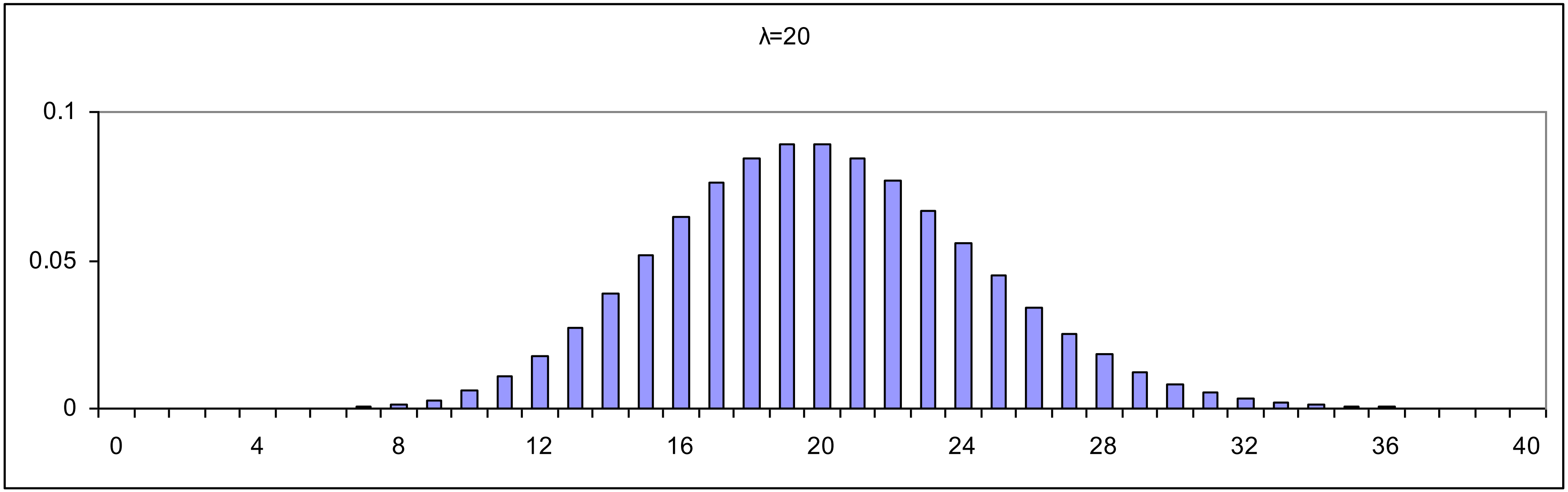
$$\langle r(r-1)\rangle = \sum_{r=0}^{\infty} r(r-1)e^{-\lambda}\frac{\lambda^r}{r!} = \lambda^2 e^{-\lambda}\sum_{r=2}^{\infty}\frac{\lambda^{r-2}}{(r-2)!} = \lambda^2 e^{-\lambda}\sum_{r'=0}^{\infty}\frac{\lambda^{r'}}{r'!} = \lambda^2$$

$$V(r) = \langle r^2\rangle - \langle r\rangle^2 = \langle r^2\rangle - \langle r\rangle + \langle r\rangle - \langle r\rangle^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

The Poisson Distribution







The Poisson Distribution - example

Number of Prussian soldiers kicked to death by horses in 20 years of the XIXth century. In 10 army corps there were 122 death cases. Thus:

λ=122/(10*20)=0.610 death/(corps*years)

Probability of no death occurring in a given corps for a given year is

$P(0;0.610) = e^{-0.61}\, 0.61^0/0! = 0.5434$

Number of events „0 fatalities” = 0.5434*200 = 108.7

Summary of

the results

Number of deaths in 1 corps in 1 year	Actual number of such cases	Poisson prediction
0	109	108.7
1	65	66.3
2	22	20.2
3	3	4.1
4	1	0.6
Sum	200	199.9

The Binomial and Poisson Distributions - comparison

Example:

A student is trying to hitch a lift. Cars pass at random intervals, at an average rate of 2 per minute. The probability of car giving a lift is 1%. What is the probability that student will be waiting:

a) After 60 cars have passed

Binomial distribution

$p=0.01$ $r=0$ $n=60$ $P(0;0.01,60) = 0.01^0*0.99^{60} = 0.547 = 54.7\%$

b) After 30 minutes

Poisson distribution

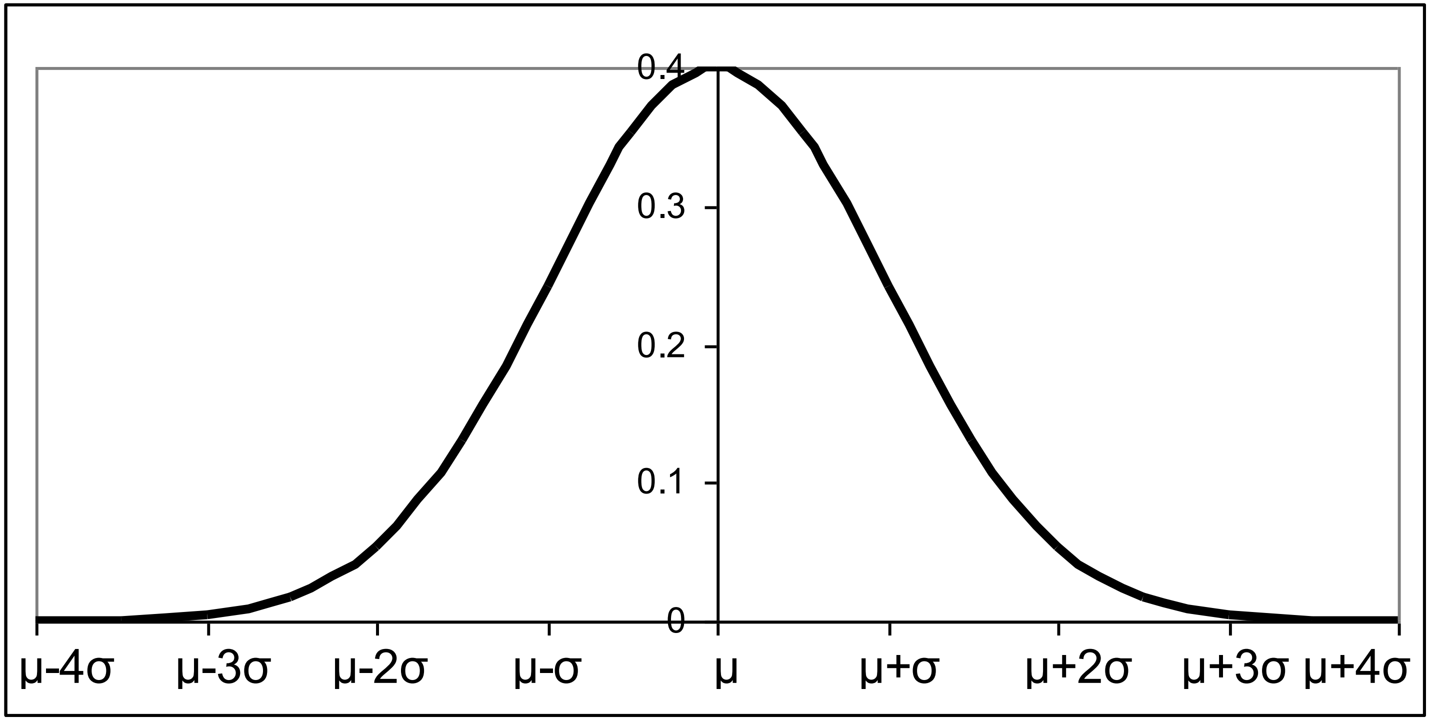
The mean number of lift-giving cars in 30 minutes is: $\lambda=0.01*30*2=0.6$

$r=0$ $P(0;0.6) = e^{-0.6} = 0.549 = 54.9\%$

The Gaussian Distribution

The Gaussian probability density distribution function

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



Basic properties of the Gaussian distribution

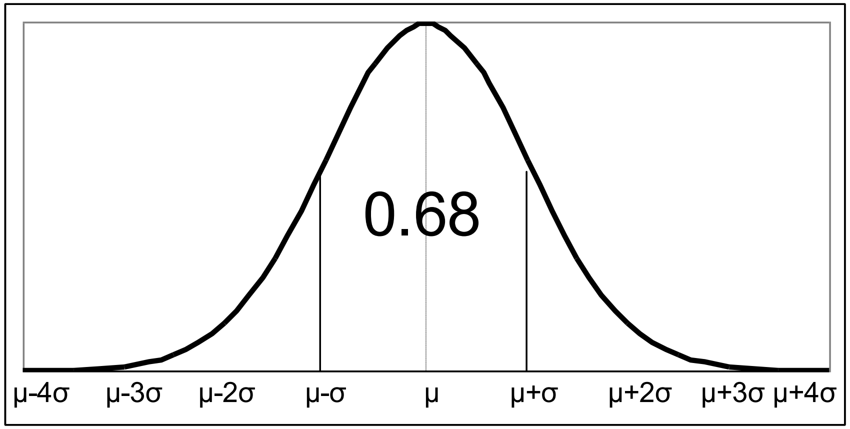
$$\int_{-\infty}^{+\infty} p(x)dx = P(-\infty \leq x \leq \infty) = 1$$

$$\int_{-\infty}^{\mu} p(x)dx = P(-\infty \leq x \leq \mu) = \tfrac{1}{2}$$

$$P(\mu - \sigma \leq x \leq \mu + \sigma) = 0.6827$$

$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 0.9545$$

$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) = 0.9973$$



If round numbers are required:

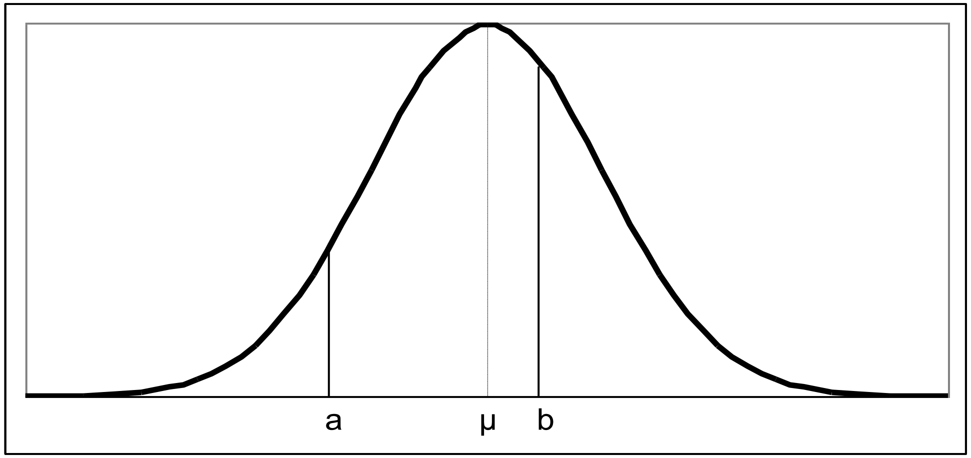
$$P(\mu - 1.645\sigma \leq x \leq \mu + 1.645\sigma) = 0.90 = 90\%$$

$$P(\mu - 1.96\sigma \leq x \leq \mu + 1.96\sigma) = 0.95 = 95\%$$

$$P(\mu - 2.576\sigma \leq x \leq \mu + 2.576\sigma) = 0.99 = 99\%$$

$$P(\mu - 3.290\sigma \leq x \leq \mu + 3.290\sigma) = 0.999 = 99.9\%$$

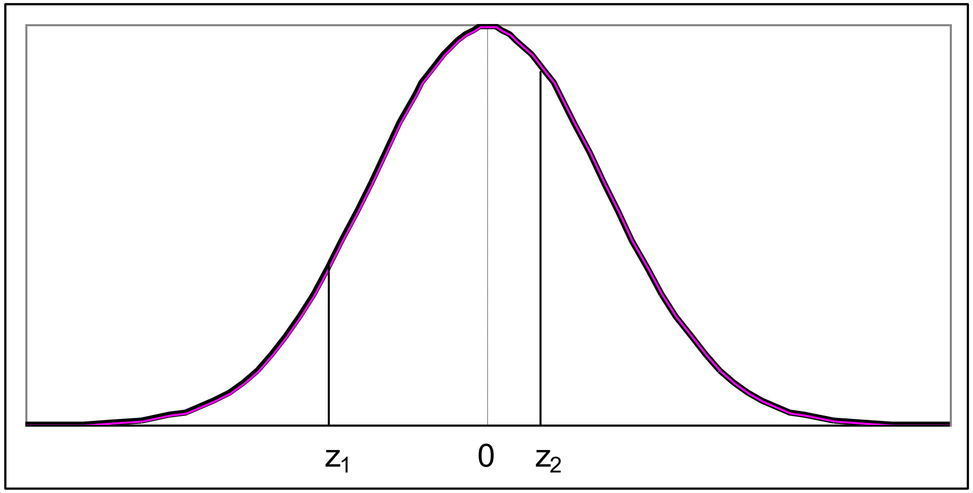
The Gaussian distribution



How to calculate

$$P\left(a \le x \le b\right) = \int_a^b p\left(x\right)dx\ ?$$

The unit Gaussian distribution



where:

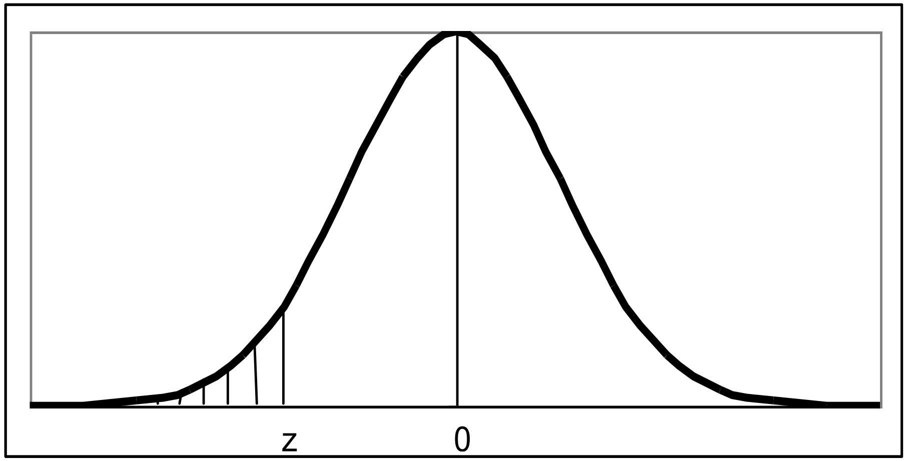
$$P\left(z_1 \le z \le z_2\right) = \int_{z_1}^{z_2} p\left(z\right)dz$$

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

z is the reduced variable

Normal curve areas for the reduced variable *z*

z	0.00	-0.01	-0.02	-0.03	-0.04	-0.05	-0.06	-0.07	-0.08	-0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379



$$if \quad z = -0.58, \quad then$$

$$P(-\infty < z \le -0.58) =$$

$$= P(z \le -0.58) = 0.2810$$

Example

The monthly salaries in a factory follow the Gaussian distribution with the mean $\mu=3280$ zł and standard deviation $\sigma=360$ zł. What is a probability that an employee chosen at random earns:

a) Less then 2800 zł

b) More then 3800 zł

c) Between 2800 zł and 3800 zł

$\mu=$	3280		
$\sigma=$	360		
$z1=(2800-3280)/360=$	-1.33333	$P(z<-1.3333)=$	0.0912
$z2=(3800-3280)/360=$	1.444444	$P(z>1.4444)=P(z<-1.4444)=$	0.0743

$$P(2800<x<3800)=P(-1.3333<z<1.4444)=1-P(z<-1.3333)-P(z<-1.4444)=$$

$$=1-0.0912-0.0743=0.8345$$

The Central Limit Theorem

If you take an average $\bar{x}$ of N independent variables, x_i, where i=1,2,3,...,N, each taken from a distribution of mean μ and variance σ^2, the distribution for $\bar{x}$

(a) has an expectation value $< \bar{x} > = \mu$,

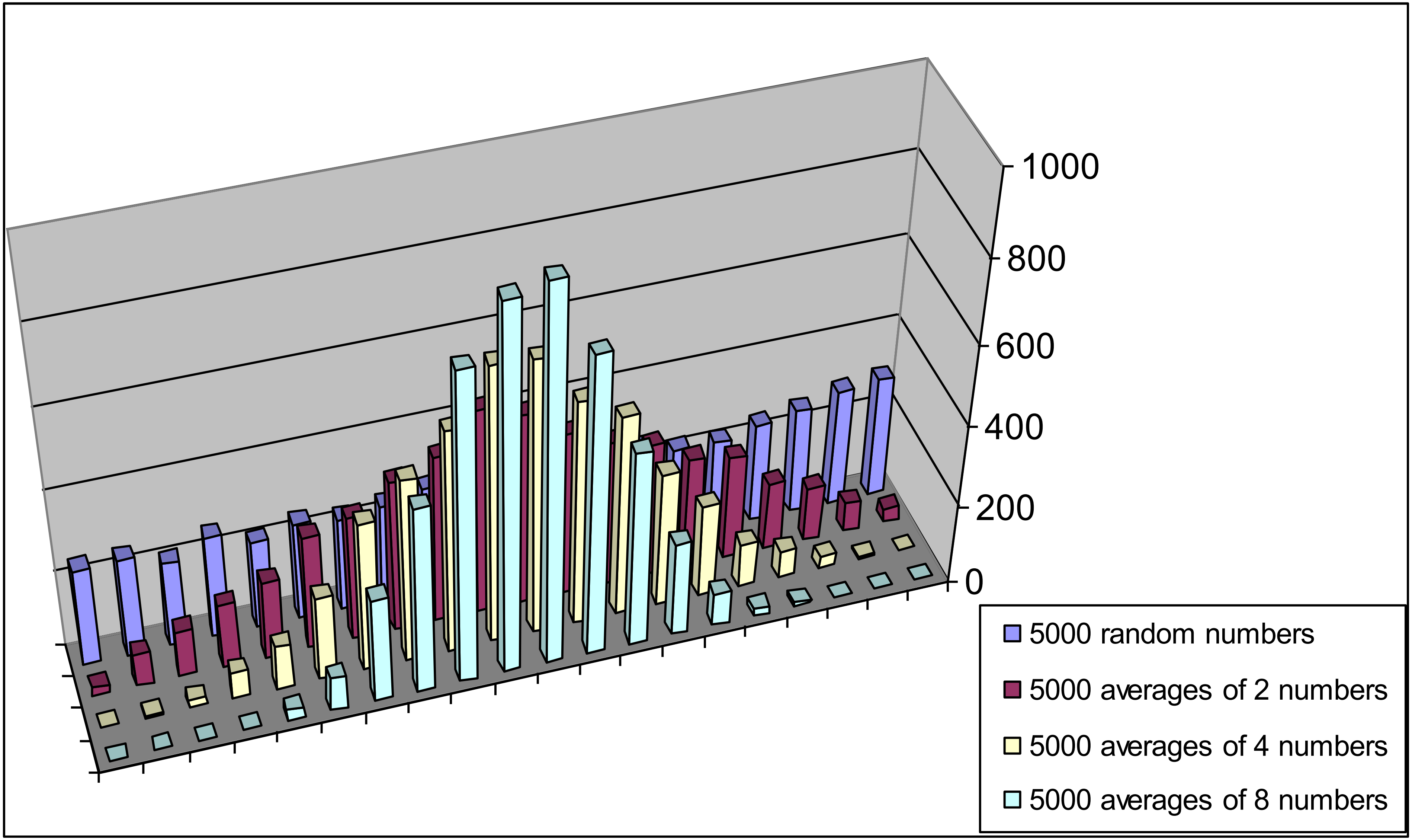
(b) has variance $V(\bar{x}) = \sigma^2/N$

(c) becomes Gaussian as N → ∞

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}{N-1} \qquad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{N}$$

Conclusion: The standard deviation for the average is smaller than for individual measurements

Distribution of random numbers and their averages



Interval for the mean

In the series of n=144 measurements the average is $\overline{x} = 60$ and the estimation of the standard deviation $s_x = 9$. Find the interval where the mean of the distribution can be determined with the probability of 0.95.

Solution:

$$s_x = \frac{s_x}{\sqrt{n}} = \frac{9}{\sqrt{144}} = 0.75$$

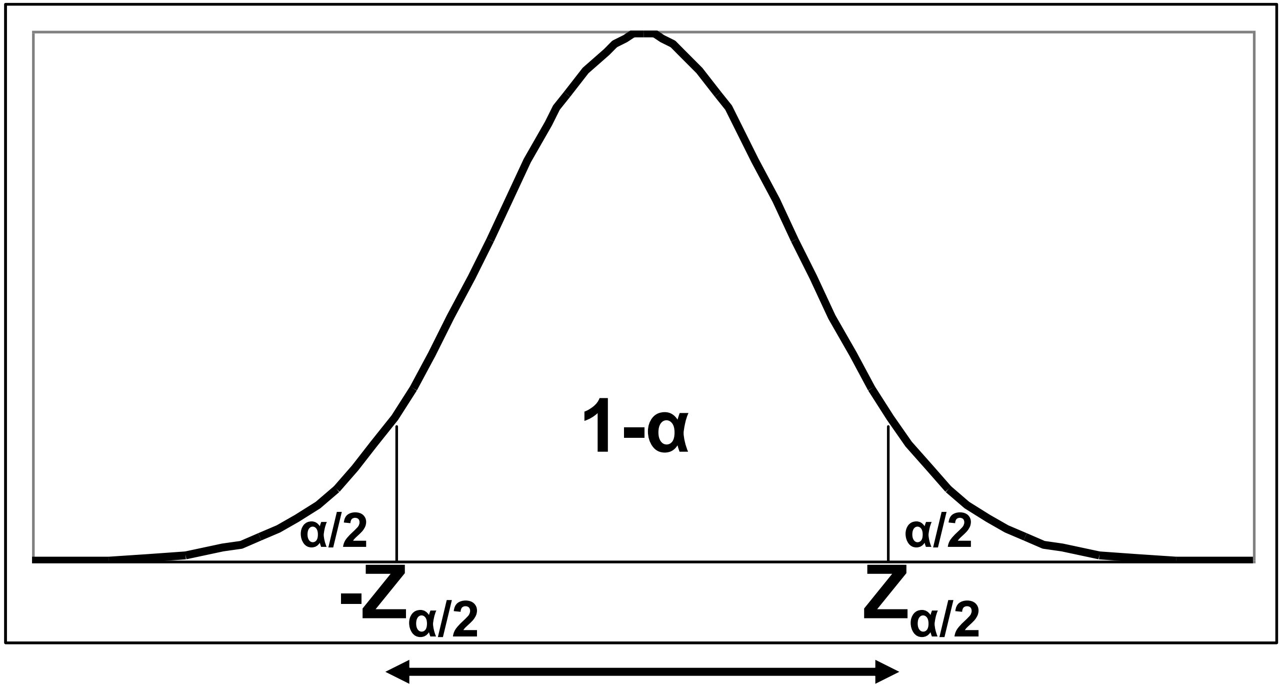
For P=0.95 $z_{critical}$ = 1.96

$$P\left(\overline{x} - 1.96 s_x \le \mu \le \overline{x} + 1.96 s_x\right) = 0.95$$

$$P\left(60 - 1.96 * 0.75 \le \mu \le 60 + 1.96 * 0.75\right) = P\left(60 - 1.5 \le \mu \le 60 + 1.5\right) =$$

$$P\left(58.5 \le \mu \le 61.5\right) = 0.95$$

Confidence and significance level



Central confidence interval = $\mu \pm Z_{\alpha/2} * \sigma$

α – significance level

(1-α) – confidence level

$Z_{\alpha/2}$ – critical value

Number of trials

Goal: determine central confidence interval for the mean ($\bar{x}$ ± d), where d is given, at a confidence level (1-α):

$$\bar{x} \pm d \quad \Rightarrow \quad \bar{x} \pm Z_{\alpha/2} * \sigma_{\bar{x}}$$

$$d = Z_{\alpha/2} * \sigma_{\bar{x}}$$

$$d = Z_{\alpha/2} * \frac{\sigma_x}{\sqrt{n}} \qquad \Rightarrow \qquad n = \frac{\left(Z_{\alpha/2}\right)^2 * \sigma_x^2}{d^2}$$

Number of trials - example

Suppose packets of cereals are produced according to Gaussian distribution of mean 350 g and standard deviation 3 g. How many packets should be selected at random to determine their average weight with the precision ±2 g at a confidence level (1-α) = 0.99 ?

$$\alpha/2 = (1-0.99)/2 = 0.005$$

$$If \quad P(Z < Z_{\alpha/2}) = 0.005, \quad then\ Z_{\alpha/2} = 2.58$$

$$n = \frac{2.58^2 * 3^2}{2^2} = 15$$

A statistical test for μ – hypothesis testing

„Is the population mean equal to a specific value μ_0 ?”

A statistical test is based on the concept of proof by contradiction and is composed of the five parts:

1. Null hypothesis, denoted by H_0.
2. Alternative hypothesis, denoted by H_a.
3. Test statistic, denoted by T.S.
4. Rejection region, denoted by R.R.
5. Conclusion

Example

The test of gas consumption for 100 cars:

$$\bar{x} = 6.28 l/100km \qquad s_x = 0.80 l/100km$$

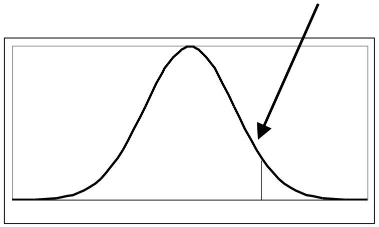
Can we accept the value of the mean gas consumption of *6.1 l/100km* given by the producer at significance level α=0.05?

$$H_0 : \mu = \mu_0$$

$$H_a : \mu > \mu_0$$

T.S. Gaussian distribution

R.R. $Z_\alpha = 1.65$



$$Z = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$$

$$Z = \frac{6.28 - 6.10}{0.80 / \sqrt{100}} = 2.25$$

$$Z > 1.65$$

Conclusion: We reject the null hypothesis

Example

In 49 rooms of the castle the average measured temperature is:

$$\bar{t} = 20.80^{\circ} C \quad with \quad s_t = 0.35^{\circ} C$$

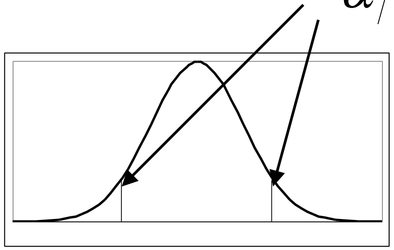
On automatic gauges the temperature was set at 21ºC. Can we say at significance level α=0.05 that the gauges are working correctly?

$$H_0 : \mu = \mu_0$$

$$H_a : \mu \neq \mu_0$$

T.S. Gaussian distribution

R.R. $Z_{\alpha/2} = 1.96$



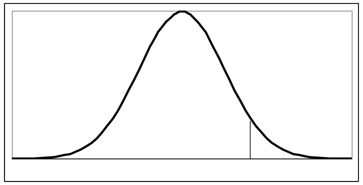
$$Z = \frac{\bar{t} - \mu_0}{s_t / \sqrt{n}}$$

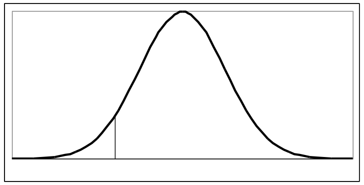
$$Z = \frac{20.80 - 21.00}{0.35 / \sqrt{49}} = -4.0$$

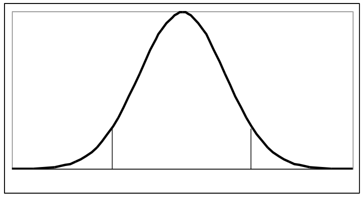
$$|Z| > 1.96$$

Conclusion: We reject the null hypothesis

Summary







$$H_0 : \mu = \mu_0 \qquad (\,\mu_0 \ given)$$

$$H_a : \begin{array}{l} \left.\begin{array}{l} 1)\ \mu > \mu_0 \\ 2)\ \mu < \mu_0 \end{array}\right\} one-tailed\ tests \\ 3)\ \mu \neq \mu_0 \quad two-tailed\ test \end{array}$$

$$Z = \frac{\bar{x} - \mu_0}{s_x \big/ \sqrt{n}}$$

R.R. at a significance level α. H_0 rejected, if:

$$1)\, Z > Z_\alpha$$

$$2)\, Z < Z_\alpha$$

$$3)\, |Z| > Z_{\alpha/2}$$

Type I and Type II Errors

Rules of decision taking in hypothesis testing

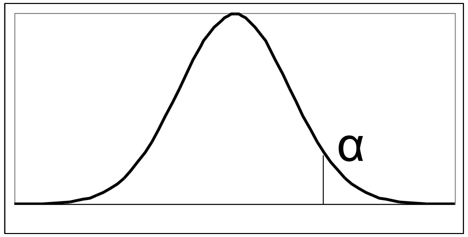
	State of nature	
Decision	H_0 true	H_0 false
H_0 rejected	Type I error α	Correct: $P=1-\beta$
H_0 not rejected	Correct: $1-\alpha$	Type II error β

α – significance level

$1-\beta$ – power of the test

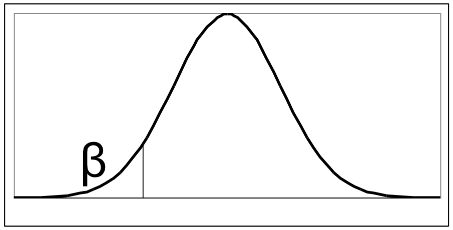
β – probability of not rejecting H0 when it is false

Hypothesis H_0



ACCEPT x REJECT

Alternative H_a



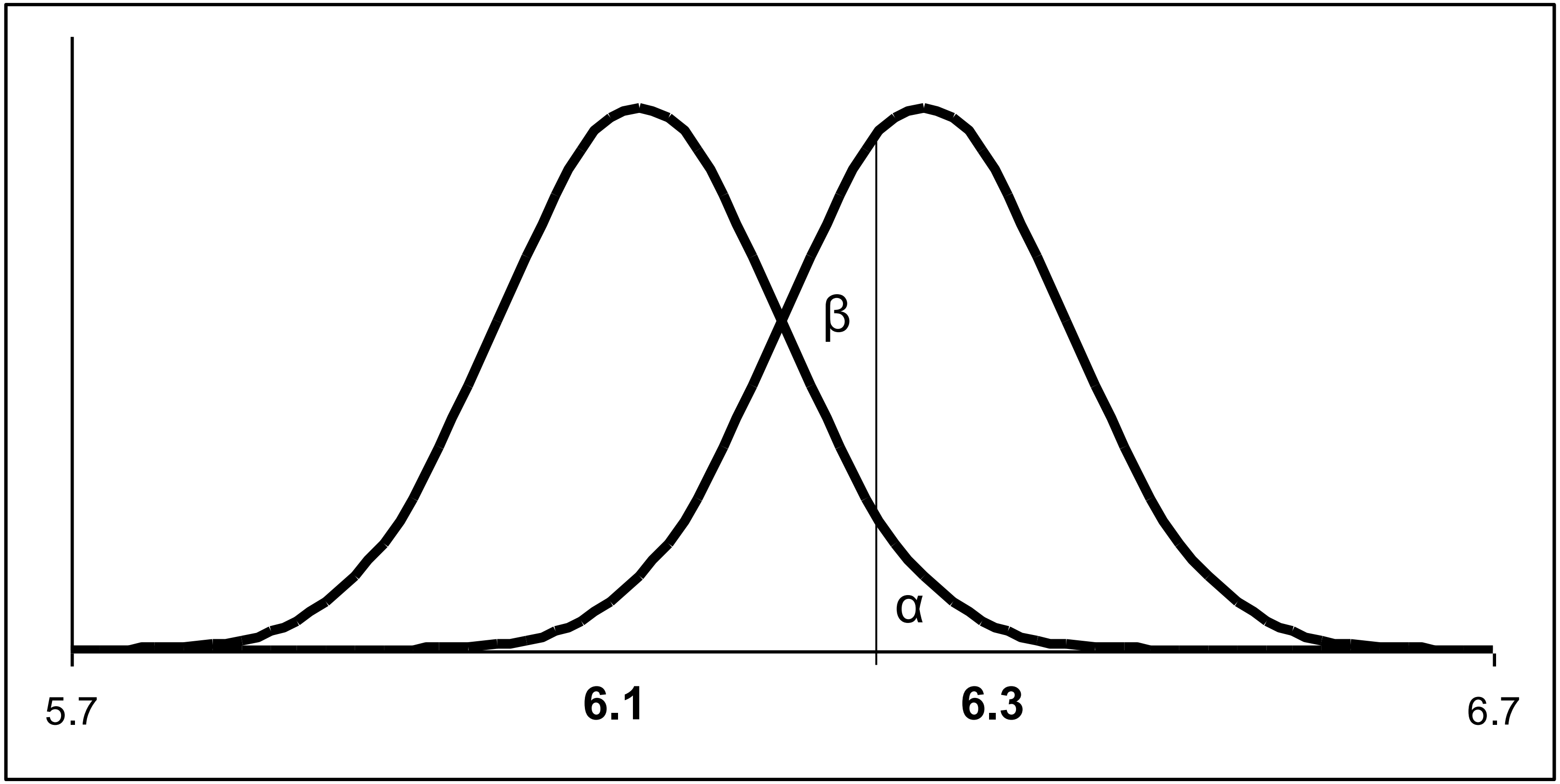
ACCEPT REJECT

Type I and Type II Errors

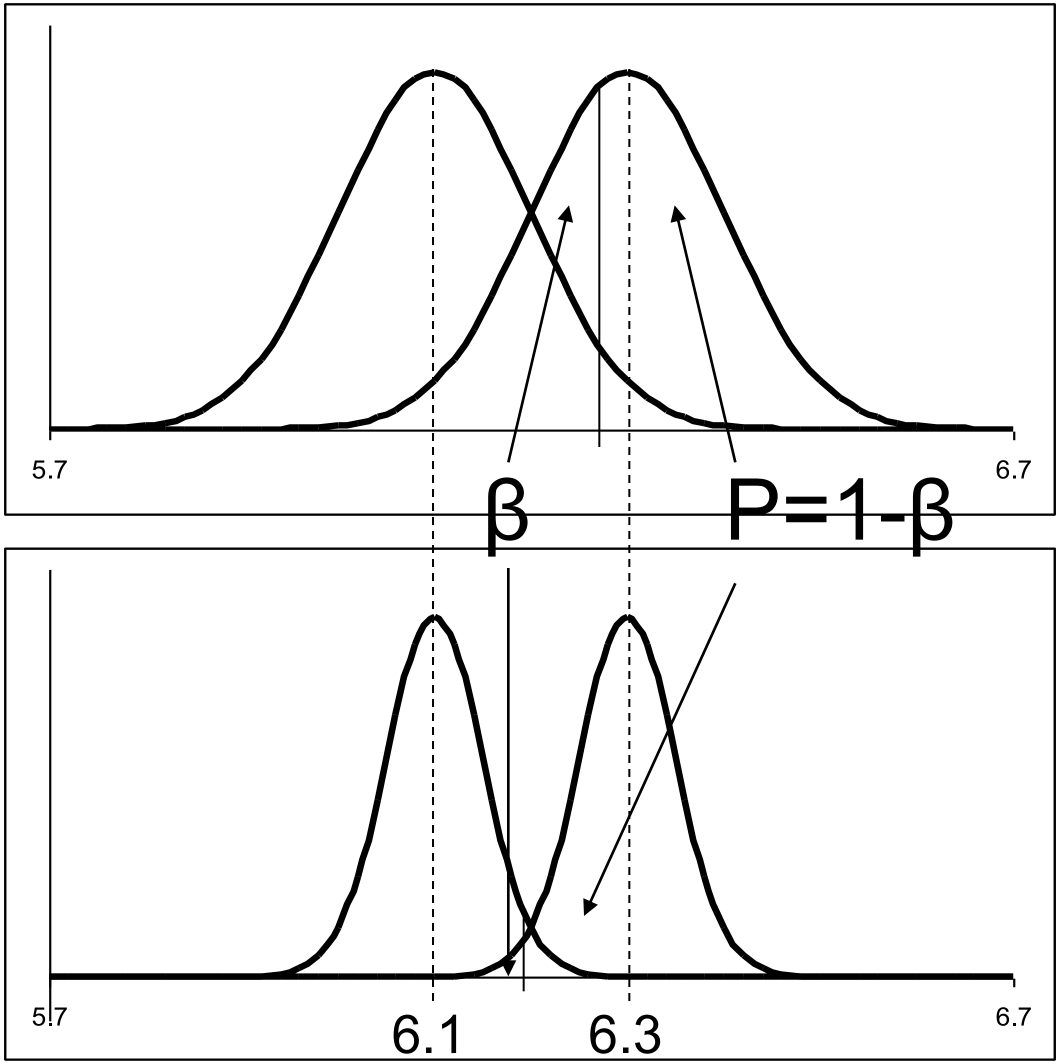
H_0: μ=6.1 H_a: μ=6.3

Standard deviation of the mean $\sigma_{\bar{x}}$ = 0.1

How to discriminate between two hypotheses?



How to increase the power of a test?



$\sigma = 1$

$n = 100 \qquad \sigma_{\bar{x}} = 0.1$

$n = 400 \qquad \sigma_{\bar{x}} = 0.05$

Inferences about μ$_1$- μ$_2$: independent samples

1) Two distributions have equal variances σ^2
2) Two samples are compared
3) Are the mean values of the distributions equal?

$\bar{x}_1$	n_1	s_1^2
$\bar{x}_2$	n_2	s_2^2

H$_0$: μ$_1$- μ$_2$ = 0

H$_a$: μ$_1$- μ$_2$ $\neq$ 0 significance level = α

Z$_{\alpha/2}$ for df = n$_1$+n$_2$-2

$$s_{x_1-x_2} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Inferences about μ$_1$- μ$_2$: independent samples

The results from two independent laboratories:

$$\bar{x}_1 = 90.3 \qquad n_1 = 92 \qquad s_1 = 9.8$$

$$\bar{x}_2 = 87.7 \qquad n_2 = 112 \qquad s_2 = 7.3$$

H_0: $\mu_1 - \mu_2 = 0$

H_a: $\mu_1 - \mu_2 \neq 0$ significance level $\alpha=0.05$

$Z_{\alpha/2}=1.96$ for df = $n_1+n_2-2=92+112-2=202$

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{(92-1)9.8^2+(112-1)7.3^2}{92+112-2}}\sqrt{\frac{1}{92}+\frac{1}{112}} = 1.198$$

$$Z = \frac{90.3-87.7}{1.198} = 2.16$$

$Z > Z_{\alpha/2}$ H_0 rejected (two laboratories present different results)

Inferences about μ$_1$- μ$_2$: independent samples

Two distributions have **different** variances σ^2

$$\bar{x}_1 \qquad n_1 \qquad s_1^2$$

$$\bar{x}_2 \qquad n_2 \qquad s_2^2$$

H$_0$: μ$_1$- μ$_2$ = 0

H$_a$: μ$_1$- μ$_2$ ≠ 0

Significance level = α

Z$_{α/2}$ for *df*

$$c = \frac{s_1^2 / n_1}{s_1^2 / n_1 + s_2^2 / n_2}$$

$$df = \frac{(n_1 - 1)(n_2 - 1)}{(1-c)^2(n_1 - 1) + c^2(n_2 - 1)}$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_{\bar{x}_1 - \bar{x}_2}}$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

If Z > Z$_{α/2}$

H$_0$ rejected (two laboratories present different results)

Student’s *t* distribution

How to determine variance from a small sample?

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

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{1}{n} \sum_{i=1}^{n} (\bar{x} - \mu)^2 + 0 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sigma_{\bar{x}}^2$$

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

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

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

Estimation of the standard deviation: $s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

Student's *t* distribution

1) Used for small samples – a rough estimation of variance can be calculated
2) For large samples *t* distribution approaches Gaussian
3) Shape of the distribution depends on df
4) Introduced by William Gosset in 1900

Testing of a hypothesis:
H_0
H_a
Significance level α
Sample: $x_1, x_2, \ldots, x_n$
Estimation of the std.dev. s
Reduced variable:

$$t = \frac{x - \overline{x}}{s} \qquad or \qquad t = \frac{\overline{x} - \mu_0}{s_{\overline{x}}}$$

Critical value t_α or $t_{\alpha/2}$

If $|t| > t_\alpha$ or $t_{\alpha/2}$, then H_0 rejected

Student’s *t* distribution

one-tailed			0.8	0.4	0.2	0.1	0.02	0.002	
two-tailed	0.8	0.6	0.4	0.2	0.1	0.05	0.01	0.001	alpha
df									df
1	0.325	0.727	1.376	3.078	6.314	12.706	63.657	636.619	1
2	0.289	0.617	1.061	1.886	2.920	4.303	9.925	31.599	2
3	0.277	0.584	0.978	1.638	2.353	3.182	5.841	12.924	3
4	0.271	0.569	0.941	1.533	2.132	2.776	4.604	8.610	4
5	0.267	0.559	0.920	1.476	2.015	2.571	4.032	6.869	5
6	0.265	0.553	0.906	1.440	1.943	2.447	3.707	5.959	6
7	0.263	0.549	0.896	1.415	1.895	2.365	3.499	5.408	7
8	0.262	0.546	0.889	1.397	1.860	2.306	3.355	5.041	8
9	0.261	0.543	0.883	1.383	1.833	2.262	3.250	4.781	9
10	0.260	0.542	0.879	1.372	1.812	2.228	3.169	4.587	10
200	0.254	0.525	0.843	1.286	1.653	1.972	2.601	3.340	200
infinit.	0.253	0.524	0.842	1.282	1.645	1.960	2.576	3.291	infinit

Student’s distribution - example

A test of 9 professors shows an average IQ of 128, with an s of 15. What are the 95% confidence limits on the true value of the average IQ of all professors?

$n = 9 \qquad\qquad df = 9 - 1 = 8$

$$s_x = \frac{15}{\sqrt{9}} = 5$$

$If\ this\ were\ Gaussian, the\ lim\,its\ would\ be 128 \pm 1.96\sigma_x,$

$i.e.\ \langle 118.2\,; 137.8 \rangle.$

$For\ Student's: the\ critical\ t_{\alpha/2}\ for\ df = 8\ is\ 2.306.$

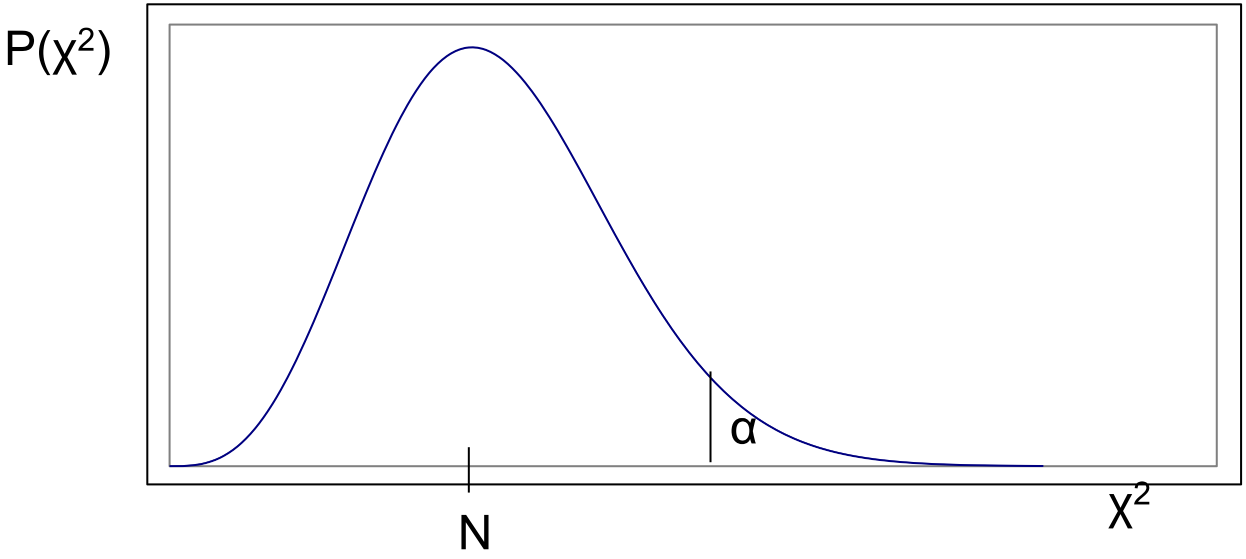
$The\ lim\,its\ are\ broader\ 128 \pm 2.306 s_x\ i.e. \langle 116{,}5\,; 139{,}5 \rangle.$

The χ^2 test for goodness of the fit

The data consist of a set of independent measurements of x and y, where the x values are exact and each y is measured with error σ. The function f(x) claims to give the ideal value of y for a given x. Then χ^2 is:

$$\chi^2 = \sum_{i=1}^{N} \frac{\left[y_i - f\left(x_i\right)\right]^2}{\sigma_i^2}$$

df = N-1



The χ^2 test for goodness of the fit

The test applied to the number of events in the i-th category. The events are subject to Poisson distribution.

$$\chi^2 = \sum_{i=1}^{N} \frac{\left[n_i - E_i\right]^2}{E_i}$$

n_i – the number of events in the i-th category

E_i – the theoretical number of events in the i-th category

Example: testing the quality of a die in 300 attempts at α=0.1

Result	1	2	3	4	5	6
n_i	52	46	59	44	48	51
E_i	50	50	50	50	50	50

H_0: $\chi^2=0$

H_a: $\chi^2 > 0$

$X^2=(52-50)^2/50 + (46-50)^2/50 + \ldots + (51-50)^2/50 = 2.84$

α=0.1 df=6-1=5 $\chi_\alpha^2=9.24$

Lotto

Probability of having a „six” in a single drawing of Lotto is equal:

$$P = \frac{1}{\begin{pmatrix} 49 \\ 6 \end{pmatrix}} = \frac{6!*(49-6)!}{49!} = 7.15112*10^{-8}$$

The number of coupons in each drawing = 20 million

The results of 200 successive drawings are given in a table at the next slide.

Are the results really random?

Execute the test at the significance level α=0.01

Comment: the results of drawings are subject to Poisson distribution, the goodness of the hypothesis is tested using the χ^2 test.

H_0: χ^2=0

H_a: χ^2 > 0

Lotto

„sixes”	n_i	P(Poisson)	E_i	$(n_i-E_i)^2/E_i$
0	52	0.2393	47.85	0.3597
1	72	0.3422	68.44	0.1854
2	45	0.2447	48.94	0.3173
3	20	0.1167	23.33	0.4759
4	6	0.0417	8.34	0.6578
5	2	0.0119	2.39	0.0625
6	3	0.0028	0.57	10.3907
				12.4493

For $\alpha=0.01$ and df=6 $\chi_\alpha^2=16.81$

Conclusion: the hypothesis about the random results of drawings cannot be rejected.

Tests for a population variance

Variability of a population is sometimes more important than its mean.

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

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

can be used for inferences concerning a population variance σ^2.

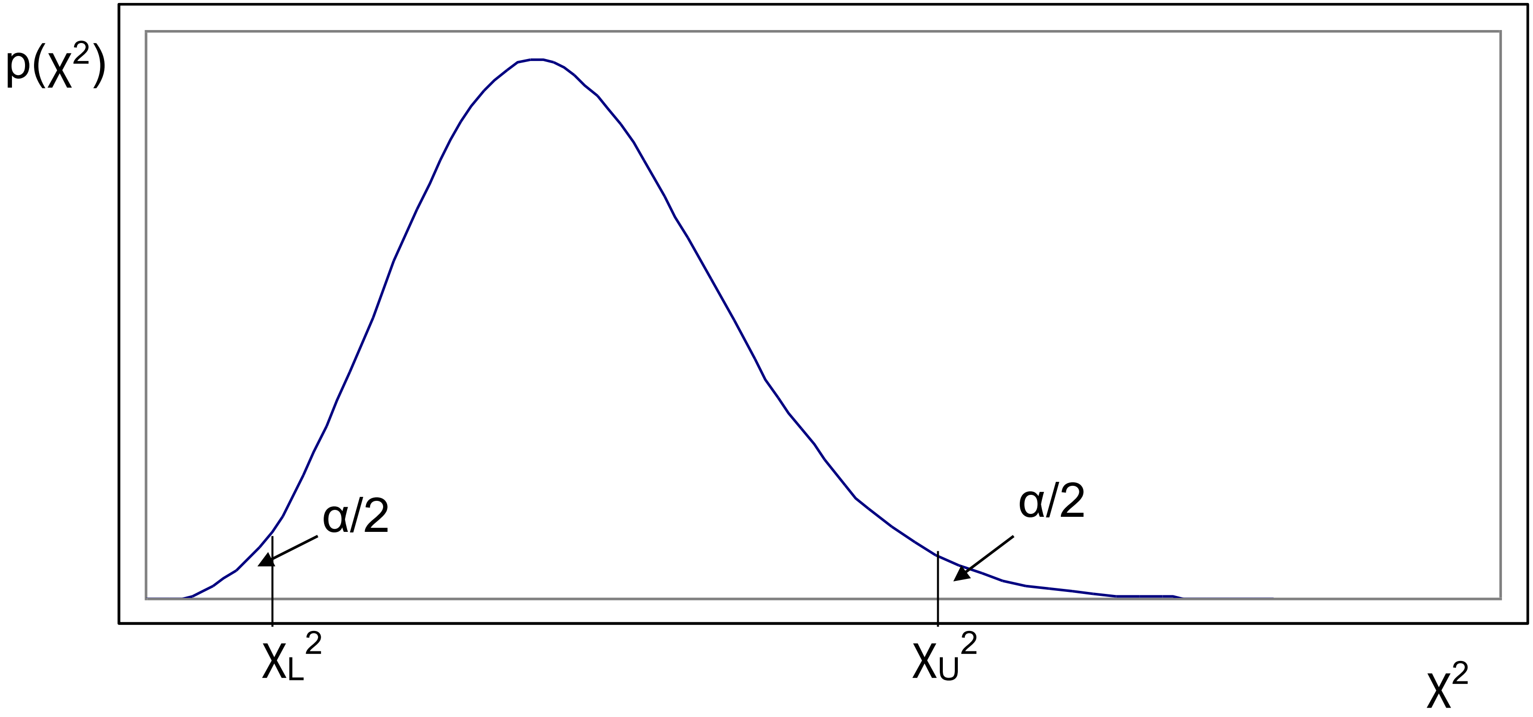
The quantity $(n-1)s^2/\sigma^2$ follows a chi-square distribution with df=n-1.

Confidence interval for σ^2:

$$\frac{(n-1)s^2}{\chi_U^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}$$

where

Upper-tail and lower-tail values of χ^2



df=n-1

Example: reaction time of drivers

The variability of reaction time was tested on a group of 7 drivers and the results in ms are the following:

120, 102, 135, 115, 118, 112 124

Estimate the population variance σ^2 for the reaction time at the confidence level 1-α = 0.90

$$\bar{x} = 118 \qquad df = 7 - 1 = 6 \qquad \alpha/2 = 0.05$$

$$s^2 = 105 \qquad \chi_L^2 = 1.6354 \qquad \chi_U^2 = 12.5916$$

$$\frac{6*105}{12.5916} < \sigma^2 < \frac{6*105}{1.6354}$$

$$50.033 < \sigma^2 < 385.23$$

$$\boxed{7.1 < \sigma < 19.6}$$

Tests for comparing two population variances

Are the variances σ_1^2 and σ_2^2 for two populations equal?

The knowledge of the variances comes from two independent samples, which are used to calculate the estimations of variances s_1^2 and s_2^2 .

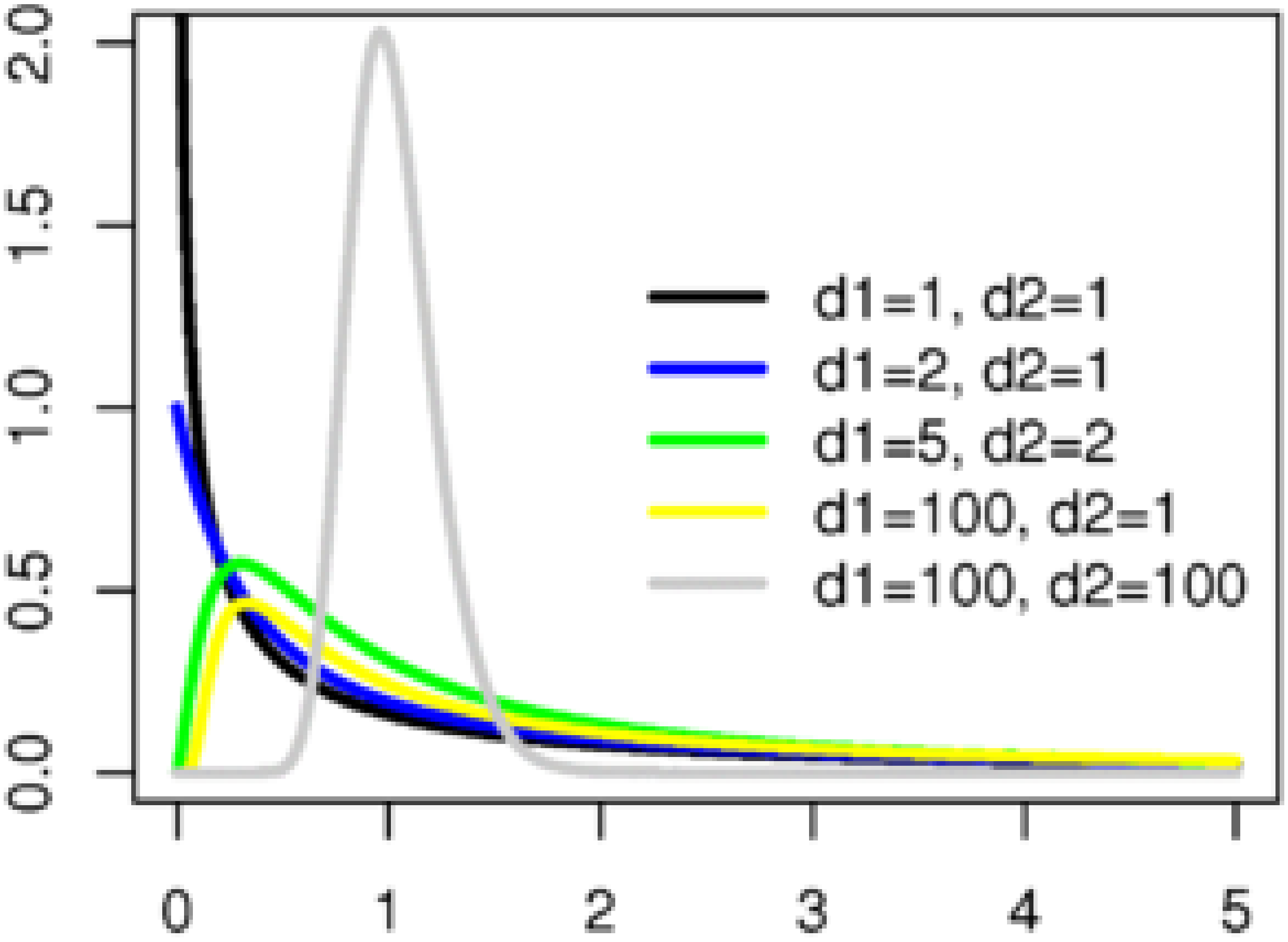
Properties of the F distribution:

1. F assume only positive values
2. F is nonsymmetrical
3. There are many F distributions associated with degrees of freedom of s_1^2 and s_2^2 , df_1 and df_2, respectively.
4. For null hypothesis $\sigma_1^2=\sigma_2^2$, the F distribution assumes the form:
5. The tables are built for $s_1^2>s_2^2$

$$F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

$$F = \frac{s_1^2}{s_2^2}$$

The F distribution



The F distribution table for $\alpha=0.05$ (one-tailed test)

df2/df1	1	2	3	4	5	6	7	8	9	10
1	161.4476	199.5	215.7073	224.5832	230.1619	233.986	236.7684	238.8827	240.5433	241.8817
2	18.51282	19	19.16429	19.24679	19.29641	19.32953	19.35322	19.37099	19.38483	19.3959
3	10.12796	9.552094	9.276628	9.117182	9.013455	8.940645	8.886743	8.845238	8.8123	8.785525
4	7.708647	6.944272	6.591382	6.388233	6.256057	6.163132	6.094211	6.041044	5.998779	5.964371
5	6.607891	5.786135	5.409451	5.192168	5.050329	4.950288	4.875872	4.81832	4.772466	4.735063
6	5.987378	5.143253	4.757063	4.533677	4.387374	4.283866	4.206658	4.146804	4.099016	4.059963
7	5.591448	4.737414	4.346831	4.120312	3.971523	3.865969	3.787044	3.725725	3.676675	3.636523
8	5.317655	4.45897	4.066181	3.837853	3.687499	3.58058	3.500464	3.438101	3.38813	3.347163
9	5.117355	4.256495	3.862548	3.633089	3.481659	3.373754	3.292746	3.229583	3.178893	3.13728
10	4.964603	4.102821	3.708265	3.47805	3.325835	3.217175	3.135465	3.071658	3.020383	2.978237
11	4.844336	3.982298	3.587434	3.35669	3.203874	3.094613	3.01233	2.94799	2.896223	2.853625
12	4.747225	3.885294	3.490295	3.259167	3.105875	2.99612	2.913358	2.848565	2.796375	2.753387
13	4.667193	3.805565	3.410534	3.179117	3.025438	2.915269	2.832098	2.766913	2.714356	2.671024
14	4.60011	3.738892	3.343889	3.11225	2.958249	2.847726	2.764199	2.698672	2.645791	2.602155
15	4.543077	3.68232	3.287382	3.055568	2.901295	2.790465	2.706627	2.640797	2.587626	2.543719

Example: testing of the drug potency

Potency of the drug after one year. Comparison of a sample taken from the production line and another sample after one year.

Sample 1: $n_1 = 10 \quad \bar{x}_1 = 10.37 \quad s_1^2 = 0.058$

Sample 2: $n_2 = 10 \quad \bar{x}_2 = 9.83 \quad s_2^2 = 0.105$

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2$$

$$T.S.: \qquad F = {}^{0.105}\!\!\diagup_{0.058} = 1.81$$

$$R.R.\ \ for\ \ \alpha = 0.01 \qquad F_{0.01,9,9} = 5.35$$

The H_0 hypothesis can not be rejected ($F<F_{critical}$)

ANalysis Of VAriance - ANOVA

Comparison of two populations 1 and 2:

H_0: $\mu_1 - \mu_2 = 0$

H_a: $\mu_1 - \mu_2 \neq 0$ significance level = α

$t_{\alpha/2}$ for df = n_1+n_2-2

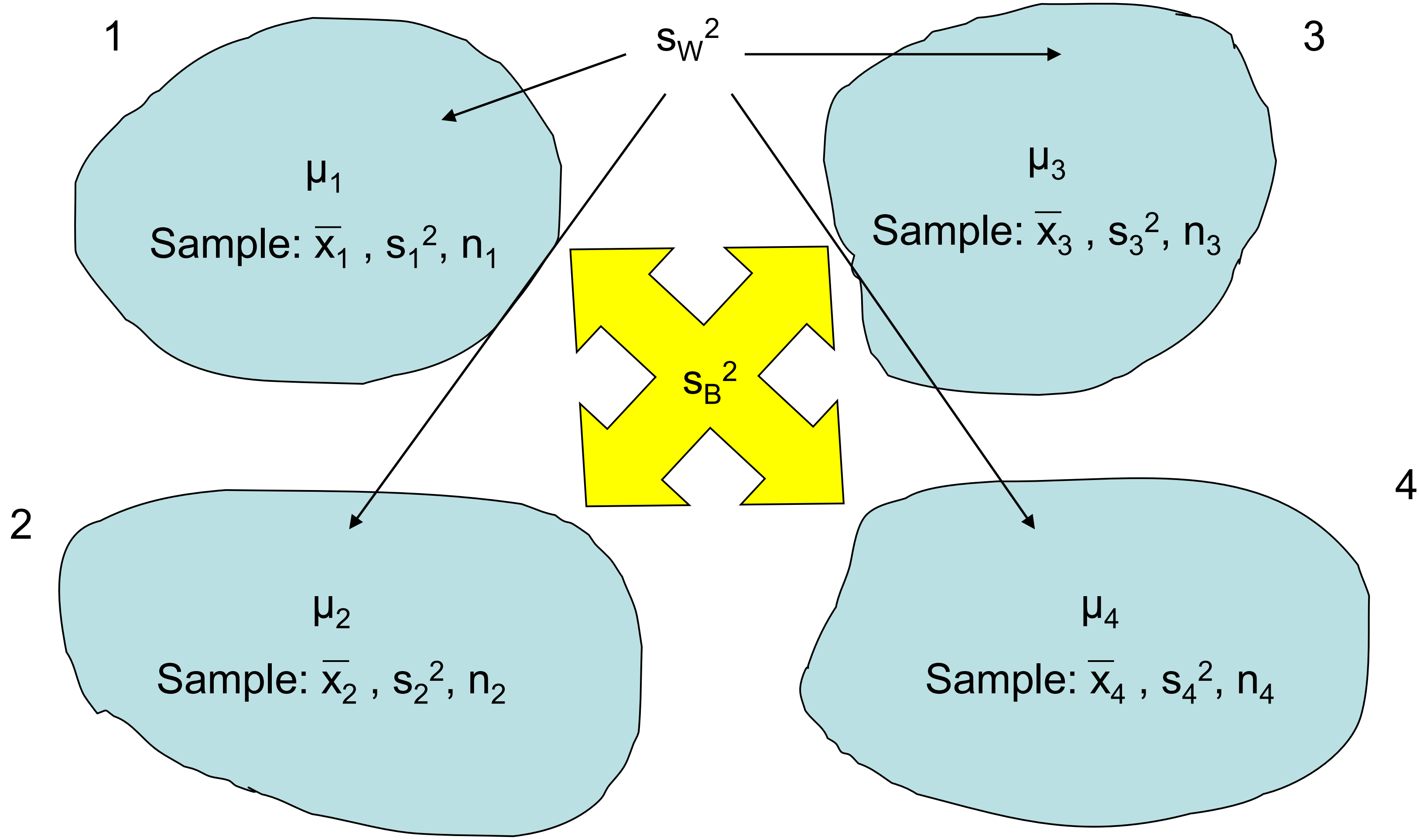
$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

H_0 rejected if $t > t_{\alpha/2}$.

How can we test the equality of more than two populations means?

ANOVA – 4 populations of equal variance σ^2



s_W^2

1

μ_1

Sample: $\bar{x}_1$, s_1^2, n_1

3

μ_3

Sample: $\bar{x}_3$, s_3^2, n_3

s_B^2

2

μ_2

Sample: $\bar{x}_2$, s_2^2, n_2

4

μ_4

Sample: $\bar{x}_4$, s_4^2, n_4

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: at least one μ_i is different

ANOVA

Calculation of variance within samples s_W^2:

$$s_W^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + (n_4 - 1)s_4^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1) + (n_4 - 1)}$$

Calculation of variance between samples s_B^2:

$$s_B^2 = \frac{\sum_{i=1}^{4}\left(\overline{x}_i - \overline{x}\right)^2}{4 - 1} \quad , \quad where \quad \overline{x} = \frac{\sum_{j=1}^{4}\overline{x}_j}{4}$$

Test statistics:

$$F = \frac{s_B^2}{s_W^2}$$

$df_1=4-1=3$

$df_2=n_1+n_2+n_3+n_4-4$

H_0 rejected if $F > F_{\alpha,\ df1,\ df2}$

ANOVA – two-way table, one-way classification

Summary of sample data for a one-way classification

Sample	Data					Total	Mean
1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	T_1	$\bar{x}_1$
2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	T_2	$\bar{x}_2$
3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	T_3	$\bar{x}_3$

Notation: x_{ij} - the jth sample observation selected from population i

n_i - the number of sample observations

n - the total sample size

T_i - the (total) sum of sample measurements from population i

G - the sum of all observations; $G = \Sigma T_i$

$\bar{x}_i$ - the average of n_i sample observations = T_i/n_i

$\bar{x}$ - the average of all sample observations = G/n

ANOVA – one-way classification

Total sum of squares: $TSS = \sum_{i,j}\left(x_{ij} - \bar{x}\right)^2 = \sum_{i,j} x_{ij}^2 - \frac{G^2}{n}$

$$\sum_{i,j}\left(x_{ij} - \bar{x}\right)^2 = \sum_{i,j}\left(x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x}\right)^2 = \sum_{i,j}\left(x_{ij} - \bar{x}_i\right)^2 + \sum_i n_i\left(\bar{x}_i - \bar{x}\right)^2$$

df: n-1 n-p p-1 ,

where p = the number of populations.

Within sample sum of squares $SSW = \sum_{i,j}\left(x_{ij} - \bar{x}_i\right)^2 = TSS - SSB$

Between-sample sum of squares $SSB = \sum_i n_i\left(\bar{x}_i - \bar{x}\right)^2 = \sum_i\left(\frac{T_i^2}{n_i}\right) - \frac{G^2}{n}$

$$s_B^2 = \frac{SSB}{p-1} \qquad s_W^2 = \frac{SSW}{n-p} \qquad F = \frac{s_B^2}{s_W^2}$$

ANOVA table

Source	Sum of squares	Degrees of freedom	Mean square	F test
Between samples	SSB	p-1	s_B^2	
Within samples	SSW	n-p	s_W^2	s_B^2/s_W^2
Totals	TSS	n-1		

ANOVA - example

Analysis of phosphorus content of tree leaves from 3 different varieties of apple trees (1, 2, and 3) at significance level α = 0.05

Variety	Phosphorus content					Totals
1	.35	.40	.58	.50	.47	2.30
2	.65	.70	.90	.84	.79	3.88
3	.60	.80	.75	.73	.66	3.54
Total						9.72

TSS = $.35^2+.40^2+\ldots+.66^2-9.72^2/15 = 6.673 - 6.299 = .374$

SSB = $(2.30^2/5+3.88^2/5+3.54^2/5) - 6.299 = .276$

SSW = $.374 - .276 = 0.098$

ANOVA – example (continued)

Source	Sum of squares	Degrees of freedom	Mean square	F test
Between samples	.276	2	.276/2=.138	
Within samples	.098	12	.098/12=.008	.138/.008 = 17.25
Totals	.374	14		

The critical value of F_α at α=0.05, df_1=2, and df_2=12 is 3.89.

Thus, we reject the null hypothesis of equality of the mean phosphorus content for the three varieties.

ANOVA – two-way classification

Two criteria A and B
x_{ijk} - belongs to class A_i (i=1,I) and to class B_j (j=1,J), k - the data number (k=1,K)

Which part of the x_{ijk} value comes from A (α_i) , B (β_j) and interaction $(\alpha\beta)_{ij}$ between both classes A and B?

$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$

α_i - effect of Factor A
β_j - effect of Factor B
$(\alpha\beta)_{ij}$ - effect of interaction between Factors A and B
ε_{ijk} – random error

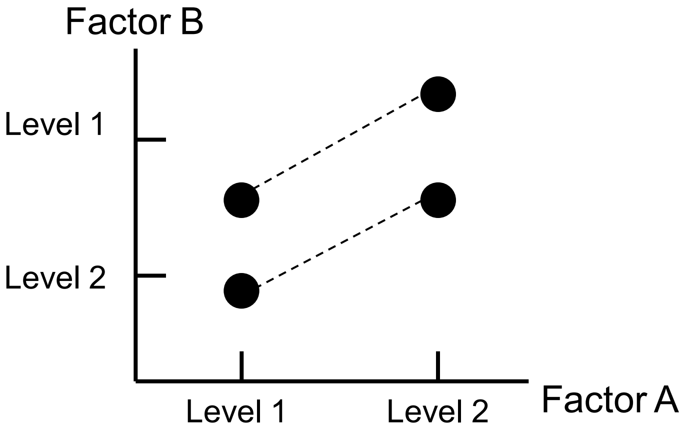
I - number of levels of Factor A
J – number of levels of Factor B

Null hypothesis: H_0 ($\mu_{ij} = \mu$) or ($\alpha_i=0$, $\beta_j=0$, $(\alpha\beta)_{ij}=0$)

ANOVA – profile plots

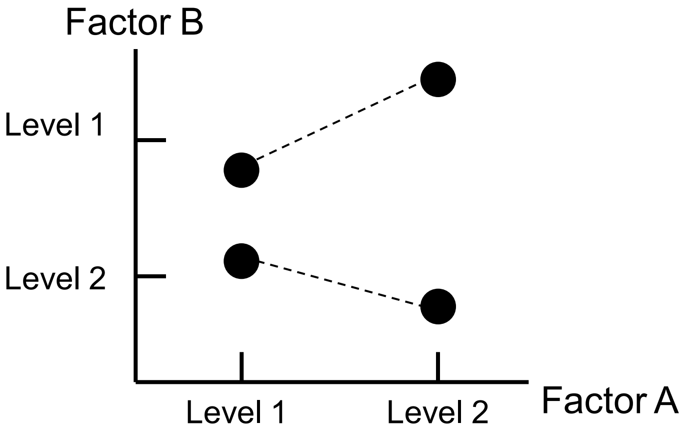
No interaction between Factors

Factor A	Factor B	
	Level 1	Level 2
Level 1	$\mu+\alpha_1+\beta_1$	$\mu+\alpha_1+\beta_2$
Level 2	$\mu+\alpha_2+\beta_1$	$\mu+\alpha_2+\beta_2$



What is interaction?

Factor A	Factor B	
	Level 1	Level 2
Level 1	$\mu+\alpha_1+\beta_1+\alpha\beta_{11}$	$\mu+\alpha_1+\beta_2+\alpha\beta_{12}$
Level 2	$\mu+\alpha_2+\beta_1+\alpha\beta_{21}$	$\mu+\alpha_2+\beta_2+\alpha\beta_{22}$



ANOVA – sum of squares

Factor A	Factor B	
	Level 1	Level 2
Level 1	$\mu+\alpha_1+\beta_1+\alpha\beta_{11}+\varepsilon_{11k}$	$\mu+\alpha_1+\beta_2+\alpha\beta_{12}+\varepsilon_{12k}$
Level 2	$\mu+\alpha_2+\beta_1+\alpha\beta_{21}+\varepsilon_{21k}$	$\mu+\alpha_2+\beta_2+\alpha\beta_{22}+\varepsilon_{22k}$

For each combination of Factors A and B k=1,K experimental data.

$$SSA = n_A K \sum_{i=1}^{I} \left(\overline{A}_i - \overline{x}\right)^2$$

$$SSB = n_B K \sum_{j=1}^{J} \left(\overline{B}_j - \overline{x}\right)^2$$

$$SSAB = K \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\overline{AB}_{ij} + \overline{x} - \overline{A}_i - \overline{B}_j\right)^2$$

$$SSW = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(x_{ijk} - \overline{AB}_{ij}\right)^2$$

ANOVA – table for two-way classification

Source	Sum of squares	Degrees of freedom	Mean square	F test
Classification A	SSA	I-1	s_A^2	$F^{(A)}=s_A^2/s_W^2$
Classification B	SSB	J-1	s_B^2	$F^{(B)}=s_B^2/s_W^2$
Interaction AB	SSAB	(I-1)(J-1)	s_{AB}^2	$F^{(AB)}=s_{AB}^2/s_W^2$
Error	SSW	IJ(K-1)	s_W^2	
Totals	TSS	IJK-1		

ANOVA – two-way classification example

Effects of 4 different pesticides on the yield of fruits for 3 varieties of citrus trees:

	Pesticide (B)			
Variety (A)	1	2	3	4
a	49	50	43	53
a	39	55	38	48
b	55	67	53	85
b	41	58	42	73
c	66	85	69	85
c	68	92	62	99

ANOVA

Source	*SS*	*df*	*MS*	*F*	*Value-p*	*Test F*
Sample (A)	3996.083	2	1998.042	47.24433	2.05E-06	3.885294
Columns (B)	2227.458	3	742.4861	17.55632	0.00011	3.490295
Interaction (AB)	456.9167	6	76.15278	1.800657	0.181684	2.99612
Error	507.5	12	42.29167			
Total	7187.958	23				

No interaction between factors A and B: the effects of the pesticide on the yield of fruits do not depend on the variety of the citrus tree ($F=1.8 < F_{test}$)
Different pesticides give different effects ($F=17 > F_{test}$)

ANOVA – one-way classification

As in previous example but we neglect differences between varieties and test the effects of different pesticides.

Analiza wariancji: jednoczynnikowa

PODSUMOWANIE=SUMMARY

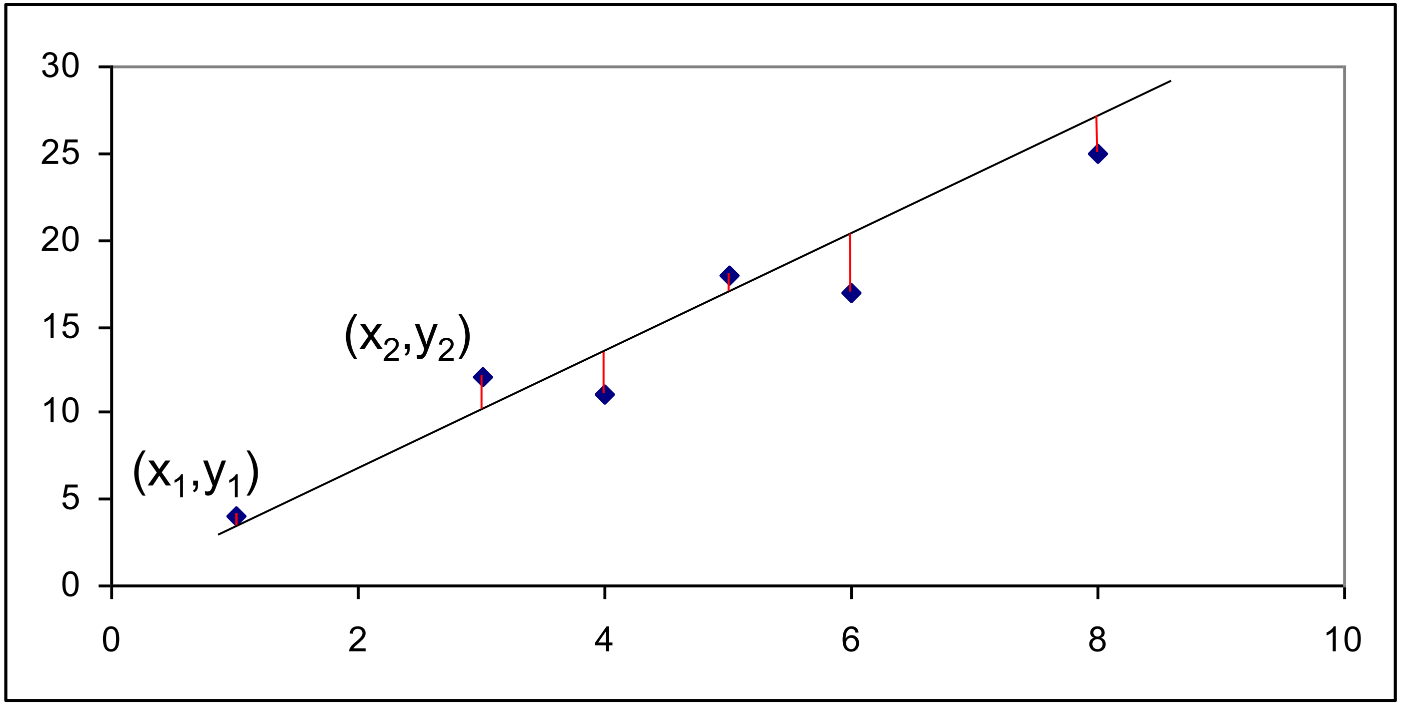
Grupy	*Licznik*	*Suma*	*Średnia*	*Wariancja*
Kolumna 1	6	318	53	150.8
Kolumna 2	6	407	67.83333	291.7667
Kolumna 3	6	307	51.16667	152.5667
Kolumna 4	6	443	73.83333	396.9667

ANALIZA WARIANCJI

Źródło wariancji	*SS*	*df*	*MS*	*F*	*Wartość-p*	*Test F*
Pomiędzy grupami	2227.458	3	742.4861	2.993594	0.055192	3.098391
W obrębie grup	4960.5	20	248.025			
Razem	7187.958	23				

No difference between pesticides (F=2.99 < F_{test})

Linear regression

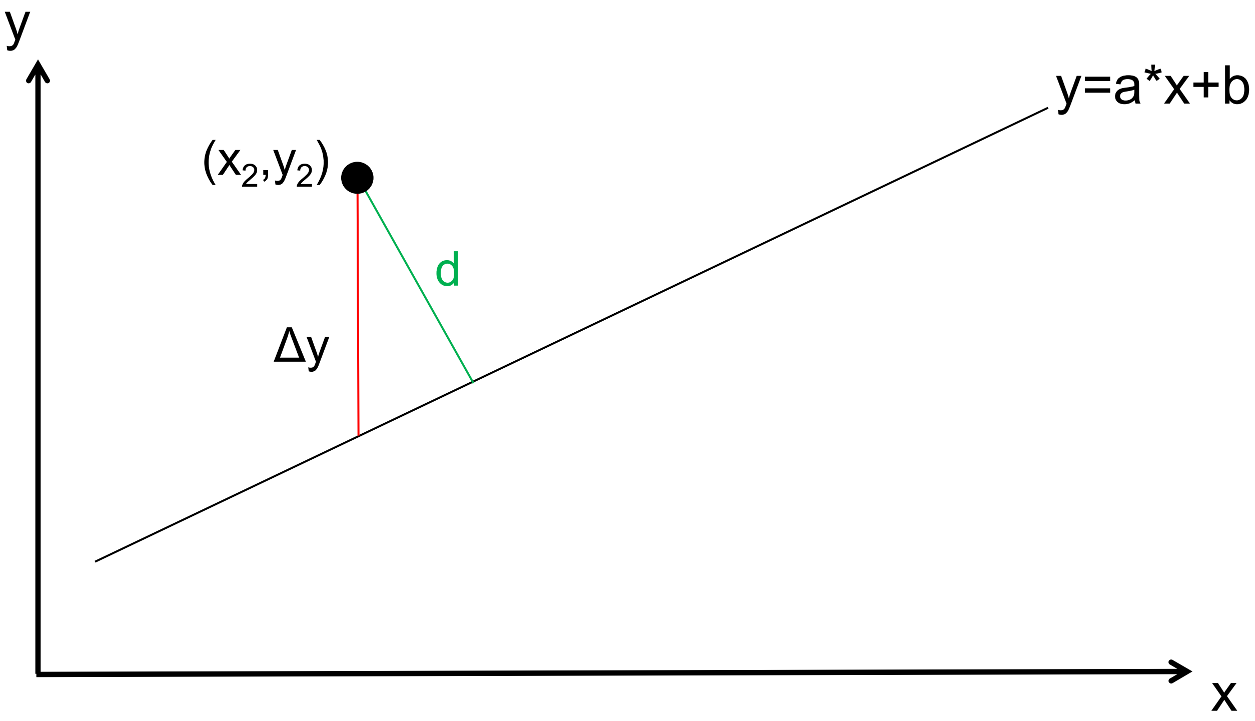


Linear regression:

y=a*x+b

Find the best values of a and b.

Linear regression



$\Delta y = y_2 - ax_2 - b$

$$d = \frac{|a * x_2 - y_2 + b|}{\sqrt{a^2 + 1}}$$

Linear regression

Basic assumptions:

1) Random distribution of y_i around the straight line
2) The variation σ_y^2 independent of x

Least squares method:

$$\Phi(a,b) = \sum_{i=1}^{n} \left[y_i - (a\,x_i + b) \right]^2$$

Determination of min Φ(a,b) with respect to a and b:

$$\frac{\partial \Phi(a,b)}{\partial a} = -2 \sum_{i=1}^{n} \left[y_i - (a\,x_i + b) \right] (x_i) = 0$$

$$\frac{\partial \Phi(a,b)}{\partial b} = -2 \sum_{i=1}^{n} \left[y_i - (a\,x_i + b) \right] = 0$$

Linear regression

$$\begin{cases} \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - bn = 0 \end{cases}$$

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases}$$

Solution of the equations system with respect to a, b:

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

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

Linear regression

The estimation of variance for y_i's:

$$s^2 = \frac{\sum_{i=1}^{n} \left(y_i - a\,x_i - b\right)^2}{n-2}$$

Estimations of variances of parameters a and b:

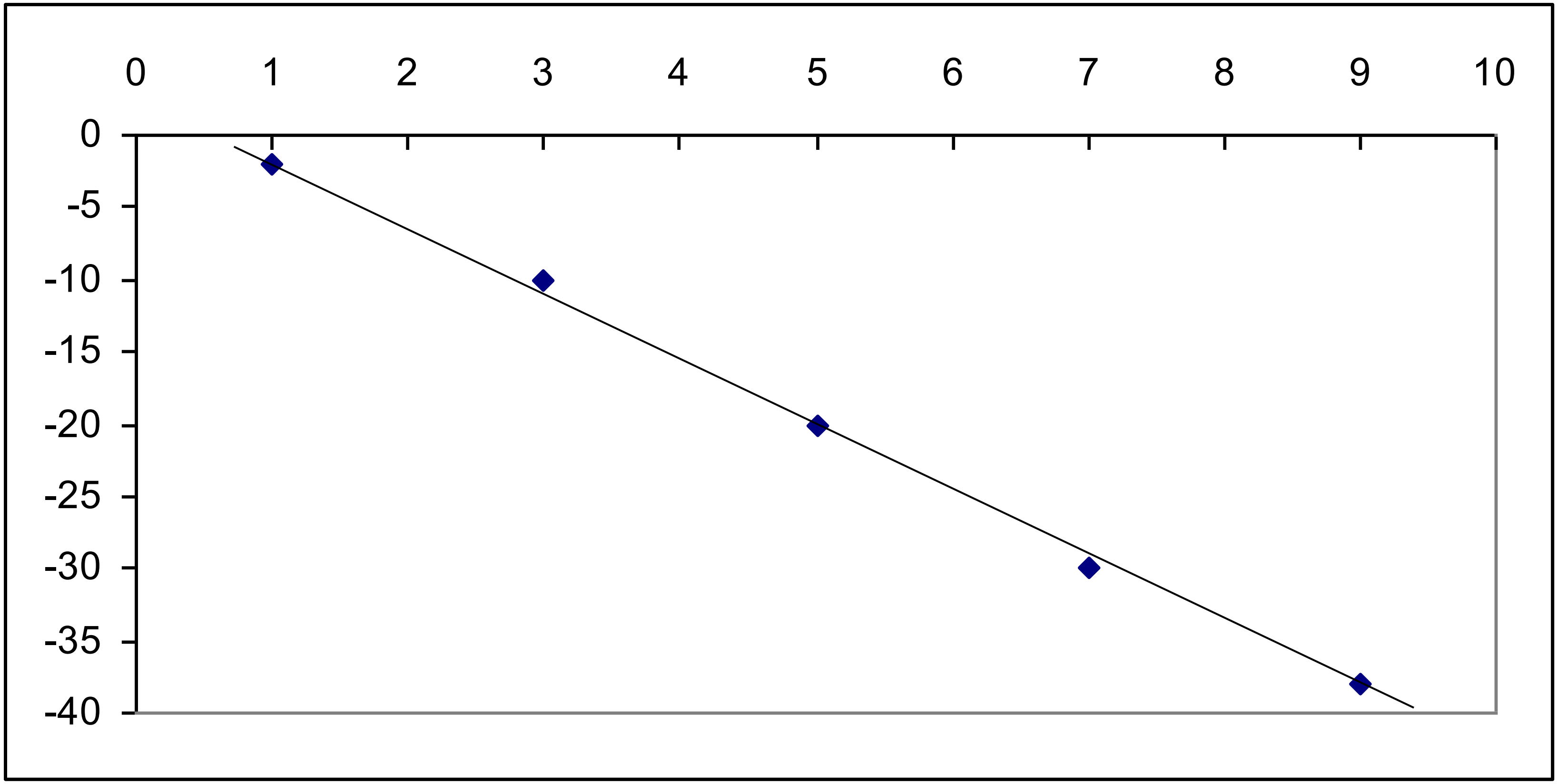
$$s_a^2 = s^2 \frac{n}{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2} \qquad s_b^2 = s^2 \frac{\left(\sum x_i^2\right)}{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

The sample correlation coefficient r

$$r = \frac{cov\left(x_i, y_i\right)}{\sqrt{var\left(x_i\right) var\left(y_i\right)}} = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}} \qquad S_{\alpha\beta} = \sum_{i=1}^{n} (\alpha_i - \bar{\alpha})\left(\beta_i - \bar{\beta}\right)$$

r lies between -1 and +1. r>0 indicates a positive relationship and r<0 a negative relationship between x and y. r=0 indicates no linear relationship between x and y.

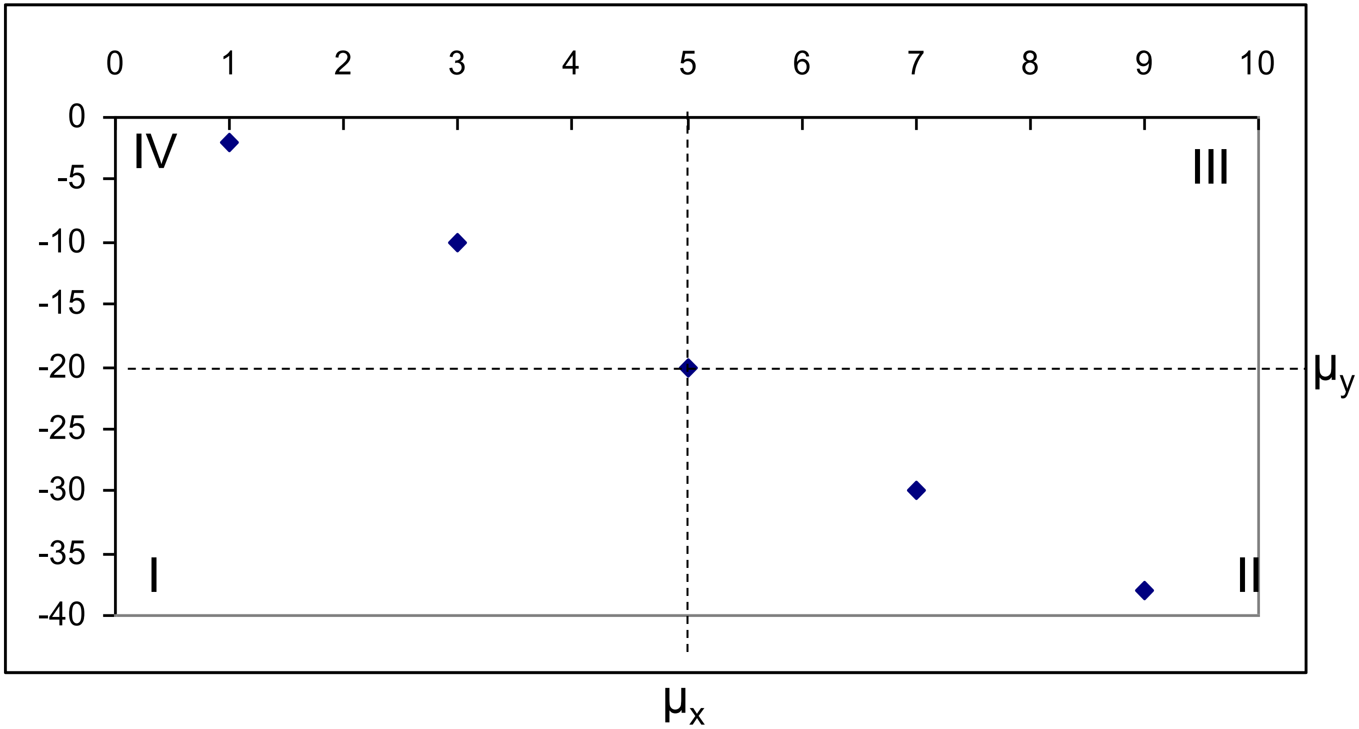
Linear regression - example



	x [m]	y [kg]	x*x	x*y	y-a*x-b	(y-a*x-b)^2	x-xsr	y-ysr
	1	-2	1	-2	-0.4	0.16	-4	18
	3	-10	9	-30	0.8	0.64	-2	10
	5	-20	25	-100	0	0	0	0
	7	-30	49	-210	-0.8	0.64	2	-10
	9	-38	81	-342	0.4	0.16	4	-18
Sum:	**25**	**-100**	**165**	**-684**	**0.00**	**1.6**	**0**	**0**

a=	-4.6 kg/m		
b=	3 kg		
s^2=	0.5333	s=	0.7303 kg
sa^2=	0.0133	sa=	0.1155
sb^2=	0.44	sb=	0.6633
xsr=	5	cov(x,y)=	-36.8000
ysr=	-20	var(x)=	8.0000
		var(y)=	169.6000
		r(x,y)=	-0.9991

More about correlation - quadrants



Quadrants:

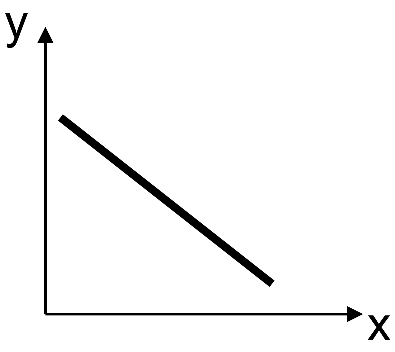
I	$x-\mu_x<0$	$y-\mu_y<0$	$(x-\mu_x)(y-\mu_y)>0$
II	$x-\mu_x>0$	$y-\mu_y<0$	$(x-\mu_x)(y-\mu_y)<0$
III	$x-\mu_x>0$	$y-\mu_y>0$	$(x-\mu_x)(y-\mu_y)>0$
IV	$x-\mu_x<0$	$y-\mu_y>0$	$(x-\mu_x)(y-\mu_y)<0$

$$cov(x,y)=\frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{n}$$

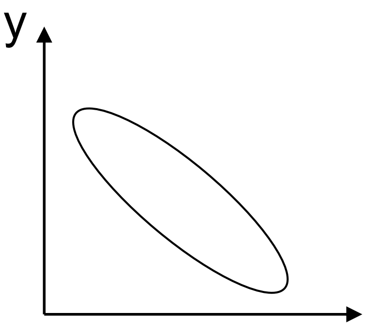
$$-\infty<cov(x,y)<\infty$$

Linear regression coefficient

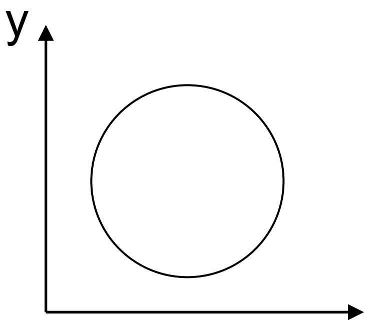
$$r(x,y) = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$$



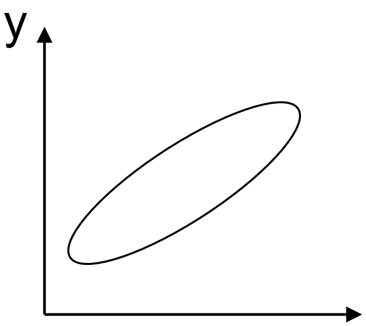
r=-1



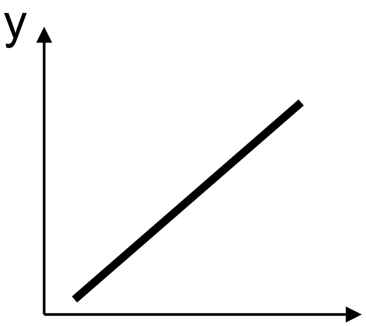
-1<r<0



r=0



0<r<1



r=1

Matrices

A set of linear equations

$$a_{11}\,x_1 + a_{12}\,x_2 + a_{13}\,x_3 = b_1$$
$$a_{21}\,x_1 + a_{22}\,x_2 + a_{23}\,x_3 = b_2$$
$$a_{31}\,x_1 + a_{32}\,x_2 + a_{33}\,x_3 = b_3$$

Let's define matrices

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

The matrix notation $\boldsymbol{Ax} = \boldsymbol{b}$

Matrix algebra

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}$$

Equality of matrices

A = ***B*** , when a_{11}=b_{11} , a_{12}=b_{12} , a_{13}=b_{13} , a_{21}=b_{21} , a_{22}=b_{22} , a_{23}=b_{23}

Multiplication of a matrix with a scalar

$$c\boldsymbol{A} = c\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} = \begin{bmatrix} ca_{11} & ca_{12} & ca_{13} \\ ca_{21} & ca_{22} & ca_{23} \end{bmatrix}$$

Sum of matrices

$$\boldsymbol{A} + \boldsymbol{B} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & a_{13} + b_{13} \\ a_{21} + b_{21} & a_{22} + b_{22} & a_{23} + b_{23} \end{bmatrix}$$

Notice: the matrices must have the same dimensions

Multiplication of matrices

$$\boldsymbol{a} = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$\boldsymbol{ab} = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = a_1 b_1 + a_2 b_2 + a_3 b_3$$

Notice: the number of columns in the first matrix must be the same as the number of rows in the second matrix

$$\boldsymbol{C} = \boldsymbol{AB} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} =$$

$$= \begin{bmatrix} a_{11} b_{11} + a_{12} b_{21} & a_{11} b_{12} + a_{12} b_{22} & a_{11} b_{13} + a_{12} b_{23} \\ a_{21} b_{11} + a_{22} b_{21} & a_{21} b_{12} + a_{22} b_{22} & a_{21} b_{13} + a_{22} b_{23} \end{bmatrix}$$

If a matrix ***C*** (m×p) is a product of matrices ***A*** (m×n) and ***B*** (n×p), then

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

i=1 do m , *j*=1 do p

Multiplication of matrices

$$\boldsymbol{A} = \begin{bmatrix} 3 & 2 \\ 4 & 5 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 2 \\ 8 & 4 & 3 \end{bmatrix}$$

$$\boldsymbol{C} = \boldsymbol{AB} = \begin{bmatrix} 3 & 2 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 \\ 8 & 4 & 3 \end{bmatrix} =$$

$$\begin{bmatrix} 3*1+2*8 & 3*0+2*4 & 3*2+2*3 \\ 4*1+5*8 & 4*0+5*4 & 4*2+5*3 \end{bmatrix} = \begin{bmatrix} 19 & 8 & 12 \\ 44 & 20 & 23 \end{bmatrix}$$

How to multiply matrices?

				B	
			1	0	2
			8	4	3
A	3	2	3*1+2*8	3*0+2*4	3*2+2*3
	4	5	4*1+5*8	4*0+5*4	4*2+5*3
					C

Transposed matrix

$$\boldsymbol{C}^T = (\boldsymbol{AB})^T = \boldsymbol{B}^T \boldsymbol{A}^T = \begin{bmatrix} 1 & 8 \\ 0 & 4 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 3 & 4 \\ 2 & 5 \end{bmatrix} = \begin{bmatrix} 19 & 44 \\ 8 & 20 \\ 12 & 23 \end{bmatrix}$$

Inverse matrix 2×2

A is a square matrix n×n
A^-1^ is an inverse matrix
I is an identity matrix

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$$

A matrix 2×2 and its inverse

Determinant of ***A***

$$\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad \boldsymbol{A}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

$$\boldsymbol{det}\ \boldsymbol{A} = ad - bc$$

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$= \frac{1}{ad-bc}\begin{pmatrix} da-bc & db-bd \\ -ca+ac & -cb+ad \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \boldsymbol{I}$$

Example: $\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \qquad \boldsymbol{A}^{-1} = -\frac{1}{2}\begin{pmatrix} 4 & -2 \\ -3 & 1 \end{pmatrix}$

$$\boldsymbol{A}^{-1}\boldsymbol{A} = -\frac{1}{2}\begin{pmatrix} 4 & -2 \\ -3 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = -\frac{1}{2}\begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \boldsymbol{I}$$

$$\boldsymbol{A}\,\boldsymbol{A}^{-1} = \left(-\frac{1}{2}\right)\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} 4 & -2 \\ -3 & 1 \end{pmatrix} = -\frac{1}{2}\begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \boldsymbol{I}$$

Excess system of linear equations

$$\boldsymbol{y_1 - (ax_1 + b) = \varepsilon_1}$$

$$\boldsymbol{y_2 - (ax_2 + b) = \varepsilon_2}$$

$$\boldsymbol{y_3 - (ax_3 + b) = \varepsilon_3}$$

$$\boldsymbol{y_4 - (ax_4 + b) = \varepsilon_4}$$

$$\boldsymbol{y_5 - (ax_5 + b) = \varepsilon_5}$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} \qquad \boldsymbol{J} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ x_4 & 1 \\ x_5 & 1 \end{bmatrix} \qquad \boldsymbol{a} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

Matrix notation

$$\boldsymbol{y - Ja = \varepsilon}$$

Excess system of linear equations

$$[\mathbf{y}-\mathbf{Ja}]=\boldsymbol{\varepsilon}$$

$$\sum_{i=1}^{n}\varepsilon_i^2=[\varepsilon_1 \quad \varepsilon_2 \quad \dots \quad \varepsilon_n]\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \dots\\ \varepsilon_n\end{bmatrix}=\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}=[\mathbf{y}-\mathbf{Ja}]^{\mathrm{T}}[\mathbf{y}-\mathbf{Ja}]$$

We search for a solution **a**, where the value of $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}$ is minimal.

$$[\mathbf{y}-\mathbf{Ja}]^{\mathrm{T}}[\mathbf{y}-\mathbf{Ja}]=[\mathbf{y}^{\mathrm{T}}-\mathbf{a}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}][\mathbf{y}-\mathbf{Ja}]=$$

$$=\mathbf{y}^{\mathrm{T}}\mathbf{y}+\mathbf{a}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{Ja}-\mathbf{a}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{y}-\mathbf{y}^{\mathrm{T}}\mathbf{Ja}=\mathbf{y}^{\mathrm{T}}\mathbf{y}+\mathbf{a}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{Ja}-2\mathbf{a}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{y}$$

$$\frac{\partial\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\partial\mathbf{a}}=2\mathbf{J}^{\mathrm{T}}\mathbf{Ja}-2\mathbf{J}^{\mathrm{T}}\mathbf{y}=\mathbf{0}$$

$$\mathbf{J}^{\mathrm{T}}\mathbf{Ja}=\mathbf{J}^{\mathrm{T}}\mathbf{y}$$

$$\boxed{\mathbf{a}=\left(\mathbf{J}^{\mathrm{T}}\mathbf{J}\right)^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{y}}$$

The optimal values of parameters **a** which minimize the sum of squares

Example of the matrix representation

$$\boldsymbol{y} = \begin{bmatrix} -2 \\ -10 \\ -20 \\ -30 \\ -38 \end{bmatrix} \qquad \boldsymbol{J} = \begin{bmatrix} 1 & 1 \\ 3 & 1 \\ 5 & 1 \\ 7 & 1 \\ 9 & 1 \end{bmatrix} \qquad \boldsymbol{a} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

$$\boldsymbol{J^T J} = \begin{bmatrix} 1 & 3 & 5 & 7 & 9 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 3 & 1 \\ 5 & 1 \\ 7 & 1 \\ 9 & 1 \end{bmatrix} = \begin{bmatrix} 165 & 25 \\ 25 & 5 \end{bmatrix} \qquad det\boldsymbol{J^T J} = 200$$

$$(\boldsymbol{J^T J})^{-\boldsymbol{1}} = \begin{bmatrix} 0.025 & -0.125 \\ -0.125 & 0.825 \end{bmatrix} \qquad \boldsymbol{J^T y} = \begin{bmatrix} 1 & 3 & 5 & 7 & 9 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -2 \\ -10 \\ -20 \\ -30 \\ -38 \end{bmatrix} = \begin{bmatrix} -684 \\ -100 \end{bmatrix}$$

$$\boldsymbol{a} = (\boldsymbol{J^T J})^{-\boldsymbol{1}}\boldsymbol{J^T y} = \begin{bmatrix} 0.025 & -0.125 \\ -0.125 & 0.825 \end{bmatrix} \begin{bmatrix} -684 \\ -100 \end{bmatrix} = \begin{bmatrix} -4.6 \\ 3 \end{bmatrix}$$

Variance/covariance

Variance of the variable y $s_y^2 = \dfrac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{n-2}$

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{J}\boldsymbol{a} = \begin{bmatrix} -2 \\ -10 \\ -20 \\ -30 \\ -38 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 3 & 1 \\ 5 & 1 \\ 7 & 1 \\ 9 & 1 \end{bmatrix} \begin{bmatrix} -4.6 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ -10 \\ -20 \\ -30 \\ -38 \end{bmatrix} - \begin{bmatrix} -1.6 \\ -10.8 \\ -20 \\ -29.2 \\ -38.4 \end{bmatrix} = \begin{bmatrix} -0.4 \\ 0.8 \\ 0 \\ -0.8 \\ 0.4 \end{bmatrix}$$

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 5$$

$$\bar{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = -20$$

$$s_y^2 = \frac{1}{5-2}\begin{bmatrix} -0.4 & 0.8 & 0 & -0.8 & 0.4 \end{bmatrix} \begin{bmatrix} -0.4 \\ 0.8 \\ 0 \\ -0.8 \\ 0.4 \end{bmatrix} = \frac{1.6}{3} = 0.5333$$

$$\boldsymbol{x} - \bar{\boldsymbol{x}} = \begin{bmatrix} -4 \\ -2 \\ 0 \\ 2 \\ 4 \end{bmatrix} \qquad \boldsymbol{y} - \bar{\boldsymbol{y}} = \begin{bmatrix} 18 \\ 10 \\ 0 \\ -10 \\ -18 \end{bmatrix}$$

$$cov(x,y) = \frac{1}{5-1}(\boldsymbol{x}-\bar{\boldsymbol{x}})^T(\boldsymbol{y}-\bar{\boldsymbol{y}}) = \frac{1}{4}\begin{bmatrix} -4 & -2 & 0 & 2 & 4 \end{bmatrix} \begin{bmatrix} 18 \\ 10 \\ 0 \\ -10 \\ -18 \end{bmatrix} = -46$$

$$var(x) = \frac{1}{5-1}(\boldsymbol{x}-\bar{\boldsymbol{x}})^T(\boldsymbol{x}-\bar{\boldsymbol{x}}) = \frac{1}{4}\begin{bmatrix} -4 & -2 & 0 & 2 & 4 \end{bmatrix} \begin{bmatrix} -4 \\ -2 \\ 0 \\ 2 \\ 4 \end{bmatrix} = 10$$

$$var(y) = \frac{1}{5-1}(\boldsymbol{y}-\bar{\boldsymbol{y}})^T(\boldsymbol{y}-\bar{\boldsymbol{y}}) = \frac{1}{4}\begin{bmatrix} 18 & 10 & 0 & -10 & -18 \end{bmatrix} \begin{bmatrix} 18 \\ 10 \\ 0 \\ -10 \\ -18 \end{bmatrix} = 212$$

Linear regression coefficient

$$r(x,y) = \frac{cov(x,y)}{\sqrt{var(x)var(y)}} = -0.99906$$

Correlation coefficient

$$s_y^2 = 0.5333$$

$$\begin{bmatrix} s_a^2 & cov(a,b) \\ cov(a,b) & s_b^2 \end{bmatrix} = s_y^2(\boldsymbol{J^T J})^{-\mathbf{1}} = 0.5333 \begin{bmatrix} 0.025 & -0.125 \\ -0.125 & 0.825 \end{bmatrix} = \begin{bmatrix} 0.0133 & -0.0667 \\ -0.0667 & 0.44 \end{bmatrix}$$

Linear correlation coefficient *r(a,b)*

$$r(a,b) = \frac{cov(a,b)}{\sqrt{var(a)var(b)}} = -0.87$$

Closer to zero is the value of r(a,b) the better is a determination of parameters a and b. They can be determined independently.

Jacobian

The model function in the linear regression y= a*x + b.
Jacobian is a matrix of derivatives over parameters a, b in all points of data
i = 1,2,...,n

$$\mathbf{J} = \begin{bmatrix} \left(\frac{\partial y}{\partial a}\right)_1 & \left(\frac{\partial y}{\partial b}\right)_1 \\ \left(\frac{\partial y}{\partial a}\right)_2 & \left(\frac{\partial y}{\partial b}\right)_2 \\ \dots & \dots \\ \left(\frac{\partial y}{\partial a}\right)_n & \left(\frac{\partial y}{\partial b}\right)_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \dots & \dots \\ x_n & 1 \end{bmatrix}$$

When fitting the data to the polynomial of the 2nd order
$y= a_0 + a_1*x + a_2*x^2$, then the Jacobian takes a form of:

$$\mathbf{J} = \begin{bmatrix} \left(\frac{\partial y}{\partial a_0}\right)_1 & \left(\frac{\partial y}{\partial a_1}\right)_1 & \left(\frac{\partial y}{\partial a_2}\right)_1 \\ \left(\frac{\partial y}{\partial a_0}\right)_2 & \left(\frac{\partial y}{\partial a_1}\right)_2 & \left(\frac{\partial y}{\partial a_2}\right)_2 \\ \dots & \dots & \dots \\ \left(\frac{\partial y}{\partial a_0}\right)_n & \left(\frac{\partial y}{\partial a_1}\right)_n & \left(\frac{\partial y}{\partial a_2}\right)_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

Transformation to a linear form

$$y = a * x + b$$

$$a, b, c, d, f - constants, \quad x, r - independent, \quad y, s - dependent\ variables$$

$$s = c * \frac{1}{r} + d \qquad y = s \quad a = c \quad x = \frac{1}{r} \quad b = d$$

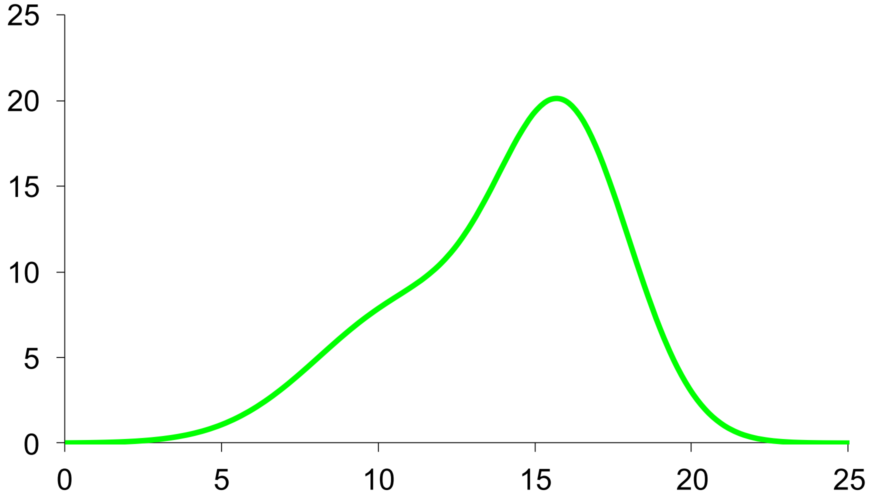
$$s = c * r^2 + d \qquad y = s \quad a = c \quad x = r^2 \quad b = d$$

$$s = e^{c*r+d} \quad ln(s) = c * r + d \qquad y = ln(s) \quad a = c \quad x = r \quad b = d$$

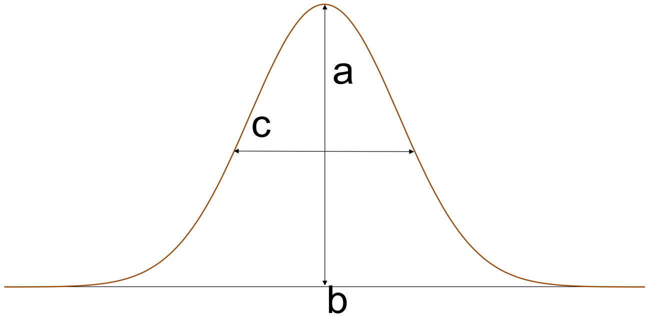
$$s = \frac{d}{c * r + f} \quad \frac{1}{s} = \frac{c}{d} * r + \frac{f}{d} \qquad y = \frac{1}{s} \quad a = \frac{c}{d} \quad x = r \quad b = \frac{f}{d}$$

Deconvolution of a complex band

Experimental band



The band should be expressed as a sum of Gaussian curves



$$P_k(x) = a_k e^{-\frac{(x-b_k)^2}{2c_k^2}}$$

a - height
b - position
c - width

The least squares method

$\{a_k\}$, $k=1:M$,   M fitted parameters

The error function (sum over n points):
$\Phi\{a_k\} = \Sigma_j\ [y_j(\text{exp}) - y_j(\{a_k\}]^2$

Problem
To minimize Φ through modification of $\{a_k\}$
using the starting values of parameters $\{a_k\}_0$

The error function and Jacobian

$$P_k(x) = a_k e^{-\frac{(x-b_k)^2}{2c_k^2}}$$

$$P(x) = \sum_{k=1}^{N} P_k(x)$$

Decomposition over N bands

Elements of the Jacobian

$$\frac{\partial P_k}{\partial a_k} = e^{-\frac{(x-b_k)^2}{2c_k^2}}$$

$$\frac{\partial P_k}{\partial b_k} = a_k \frac{(x-b_k)}{c_k^2} e^{-\frac{(x-b_k)^2}{2c_k^2}}$$

$$\frac{\partial P_k}{\partial c_k} = a_k \frac{(x-b_k)^2}{c_k^3} e^{-\frac{(x-b_k)^2}{2c_k^2}}$$

Algorythm

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \qquad J = \begin{bmatrix} \left(\frac{\partial P}{\partial a_1}\right)_1 & \left(\frac{\partial P}{\partial b_1}\right)_1 & \left(\frac{\partial P}{\partial c_1}\right)_1 & \left(\frac{\partial P}{\partial a_2}\right)_1 & \dots \\ \left(\frac{\partial P}{\partial a_1}\right)_2 & \left(\frac{\partial P}{\partial b_1}\right)_2 & \dots & \dots & \\ \dots & \dots & \dots & \dots & \\ \left(\frac{\partial P}{\partial a_1}\right)_n & \left(\frac{\partial P}{\partial b_1}\right)_n & \dots & \dots & \end{bmatrix} \qquad \Delta a = \begin{bmatrix} \Delta a_1 \\ \Delta b_1 \\ \Delta c_1 \\ \Delta a_2 \\ \Delta b_2 \\ \Delta c_2 \end{bmatrix}$$

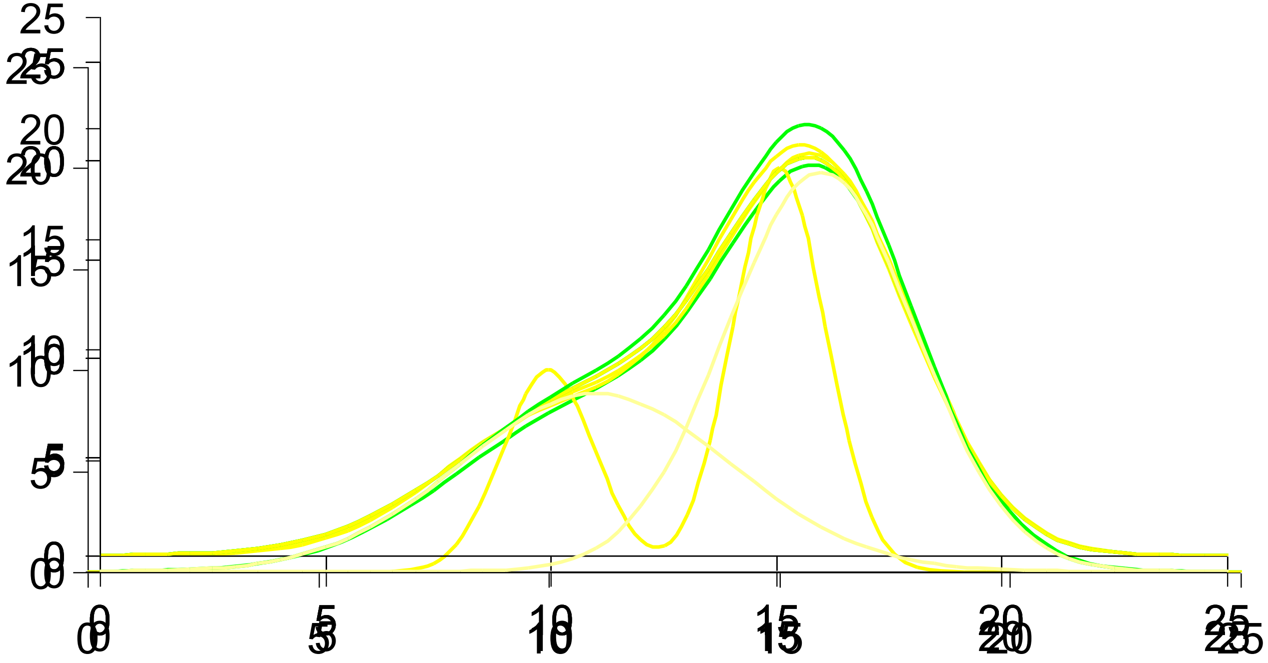
Corrections to the values of parameters $\{a_k\}$

$$\Delta a = \left(J^T J\right)^{-1} J^T Y$$

The least squares method

Pasmo rozłożone na 2 składowe

Krok 4

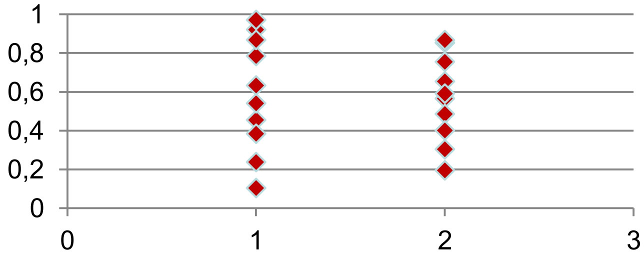


Relations between parameters 1+1>2

Objects	Par 1	Par 2
A	0.455	0.566
B	0.921	0.853
C	0.785	0.195
D	0.104	0.304
E	0.541	0.654
F	0.384	0.486
G	0.633	0.591
H	0.868	0.755
I	0.238	0.401
J	0.971	0.867

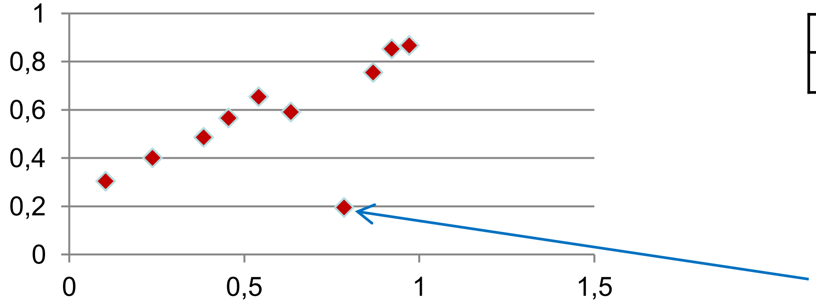
Do all objects originate from the same source?

The separate analysis of parameters:



Each parameter spans from 0.1 to 1.0.

The correlation analysis of parameters:



The correlation graph

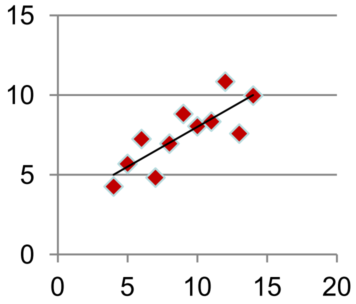
1	
0.644603	1

Correlation coefficient

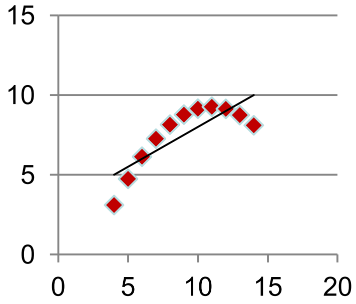
The object C is different.

The Anscombe‘s quartet

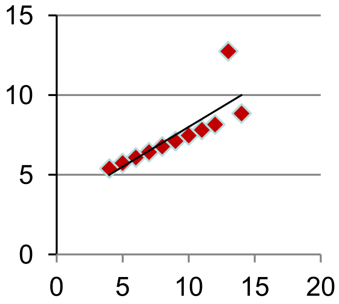
Attribute	Value
Aritmethic mean of x	9
Variance of x	11
Aritmethic mean of y	7.50 (two decimal digits identical)
Variance of y	4.122 do 4.127 (two decimal digits identical)
Correlation coefficient between x and y	0.816 (three decimal digits identical)
The equation of linear regression	y=3.00+0.500x (two and three decimal digits identical, respectively)



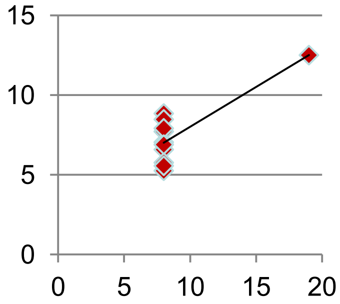
A



B



C



D

The Anscombe‘s quartet

The source data:

	A		B		C		D	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
mean	9	7.500909	9	7.500909	9	7.5	9	7.500909
variance	11	4.127269	11	4.127629	11	4.12262	11	4.123249
Corr.coef.	0.816421		0.816237		0.816287		0.816521	
Linear regr.	0.500091	3.000091	0.5	3.000909	0.499727	3.002455	0.499909	3.001727

Models

- Generally, in chemometrics no exact functional relation $f(x_1, x_2,..., x_n)$ exists
- Empirical models are used based on a form of function and a list of variables
- Often the linear model is used as a starting point:

$$y(x)=b_0+b_1x_1+b_2x_2+...+b_mx_m$$

Empirical multivariant model

Linear function $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im}$.

Example

Significance of a variable: x_k is significant for the model, if the confidence interval for the coefficient b_k does not include the zero value

Orthogonal variables: if the variables are orthogonal, then the removal or inclusion of a variable does not change the coefficients b_k for remaining variables

Significance of a model:

Variance of rests

$$s^2 = \frac{\sum_{i=1}^{n} \left(y_i^{\exp} - y_i^{cal}\right)^2}{n-m-1}$$

Variability of the model

$$s_y^2 = \frac{\sum_{i=1}^{n} \left(y_i^{\exp} - \overline{y}\right)^2}{n-1}$$

Statistics F:

$$F = \frac{s_y^2}{s^2}$$

if $F > F_{kr}$, then the model accordingly describes the variability of objects