STATISTICS 1

Keijo Ruohonen

(Translation by Jukka-Pekka Humaloja and Robert Piché)

2011

Table of Contents

Preface

This document is the lecture notes for the course "MAT-33317 Statistics 1", and is a translation of the notes for the corresponding Finnish-language course. The laborious bulk translation was taken care of by Jukka-Pekka Humaloja and the material was then checked by professor Robert Piché. I want to thank the translation team for their effort.

The lecture notes are based on chapters 8, 9, 10, 12 and 16 of the book WALPOLE, R.E. & MYERS, R.H. & MYERS, S.L. & YE, K.: *Probability & Statistics for Engineers & Scientists*, Pearson Prentice Hall (2007). The book (denoted WMMY in the following) is one of the most popular elementary statistics textbooks in the world. The corresponding sections in WMMY are indicated in the right margin. These notes are however much more compact than WMMY and should not be considered as a substitute for the book, for example for self-study. There are many topics where the presentation is quite different from WMMY; in particular, formulas that are nowadays considered too inaccurate have been replaced with better ones. Additionally, a chapter on stochastic simulation, which is not covered in WMMY, is included in these notes.

The examples are mostly from the book WMMY. The numbers of these examples in WMMY are given in the right margin. The examples have all been recomputed using MATLAB, the statistical program JMP, or web-based calculators. The examples aren't discussed as thoroughly as in WMMY and in many cases the treatment is different.

An essential prerequisite for the course “MAT-33317 Statistics” is the course “MAT-20501 Probability Calculus” or a corresponding course that covers the material of chapters 1–8 of WMMY. MAT-33317 only covers the basics of statistics. The TUT mathematics department offers many advanced courses that go beyond the basics, including “MAT-34006 Statistics 2”, which covers statistical quality control, design of experiments, and reliability theory, “MAT-51706 Bayesian methods”, which introduces the Bayesian approach to solving statistical problems, “MAT-51801 Mathematical Statistics”, which covers the theoretical foundations of statistics, and “MAT-41281 Multivariate Statistical Methods”, which covers a wide range of methods including regression.

Keijo Ruohonen

Chapter 1

FUNDAMENTAL SAMPLING DISTRIBUTIONS AND DATA DESCRIPTIONS

This chapter is mostly a review of basic Probability Calculus. Additionally, some methods for visualisation of statistical data are presented.

1.1 Random Sampling [8.1]

A *population* is a collection of all the values that may be included in a sample. A numerical value or a classification value may exist in the sample multiple times. A *sample* is a collection of certain values chosen from the population. The *sample size*, usually denoted by n, is the number of these values. If these values are chosen at random, the sample is called a *random sample.*

A sample can be considered a sequence of random variables: $X_1, X_2, \ldots, X_n$ ("the first sample variable", "the second sample variable", ...) that are independent and identically distributed. A concrete realized sample as a result of sampling is a sequence of values (numerical or classification values): $x_1, x_2, \ldots, x_n$. Note: random variables are denoted with upper case letters, realized values with lower case letters.

The sampling considered here is actually *sampling with replacement.* In other words, if a population is finite (or countably infinite), an element taken from the sample is replaced before taking another element.

Sampling without replacement is not considered in this course.

1.2 Some Important Statistics [8.2]

A *statistic* is some individual value calculated from a sample: $f(X_1, \ldots, X_n)$ (random variables) or $f(x_1, \ldots, x_n)$ (realized values). A familiar statistic is the *sample mean*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{or} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The former is a random variable while the latter is a numerical value called the realized sample mean.

Another familiar statistic is the *sample variance*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \text{or} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Again, the former is a random variable and the latter is a realized numerical value. The sample variance can be written also in the form

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\overline{X}^2$$

Expand the square $(X_i - \overline{X})^2$.

(and s^2 similarly). The *sample standard deviation*, denoted by S (random variable) or s (realized value), is the positive square root of the sample variance. Other important statistics are the *sample maximum* and the *sample minimum*

$$X_{\max} = \max(X_1, \ldots, X_n) \quad \text{or} \quad x_{\max} = \max(x_1, \ldots, x_n),$$
$$X_{\min} = \min(X_1, \ldots, X_n) \quad \text{or} \quad x_{\min} = \min(x_1, \ldots, x_n)$$

and their difference, the *sample range.*

$$R = X_{\max} - X_{\min} \quad \text{or} \quad r = x_{\max} - x_{\min}.$$

1.3 Data Displays and Graphical Methods [8.3]

In addition to the familiar bar chart or histogram, there are other very common ways to visualize data.

Example. *In this example nicotine content was measured in a random sample of* $n = 40$ *cigaretters:* [8.3]

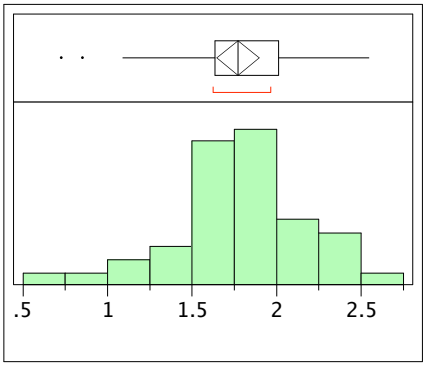
1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97	0.85	1.24
1.58	2.03	1.70	2.17	2.55	2.11	1.86	1.90	1.68	1.51
1.64	0.72	1.69	1.85	1.82	1.79	2.46	1.88	2.08	1.67
1.37	1.93	1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

The statistical software package JMP prints the following (a little tidied up) graphical display:

Nicotinedata: Distribution

Distributions

Content



Quantiles		
100.0%	maximum	2.5500
99.5%		2.5500
97.5%		2.5478
90.0%		2.3070
75.0%	quartile	2.0150
50.0%	median	1.7700
25.0%	quartile	1.6325
10.0%		1.2530
2.5%		0.7232
0.5%		0.7200
0.0%	minimum	0.7200

Moments	
Mean	1.77425
Std Dev	0.3904559
Std Err Mean	0.0617365
upper 95% Mean	1.8991239
lower 95% Mean	1.6493761
N	40

The box-and-whiskers *plot in the upper left depicts the distribution of data. The box denotes the part of the data that lies between the lower* $q(0.25)$ *and upper* $q(0.75)$ *quartiles (quartiles are explained below). Inside the box there is also a vertical line denoting the sample median (see next page). The whiskers show the sample maximum and the sample minimum. Other quantiles can also be marked in the whiskers (see next page). (Inside the box there is the mean value square that denotes the confidence interval that will be considered in section 3.8.)*

In most cases, one or more outliers are removed from the sample. An outlier *is a sample value that differs from the others so remarkably, that it can be considered an error in the sample. There are various criteria to classify outliers. In the picture, outliers are marked with dots (there are two of them).*

Instead of the bar chart, some people prefer a *stem-and-leaf diagram* to visualize data. If a d-decimal presentation is used, the $d-1$ first decimals are chosen as the stem and the rest of the decimals are the leaves. Data is typically displayed in the form

$$1.2 \,|\, 0227779$$

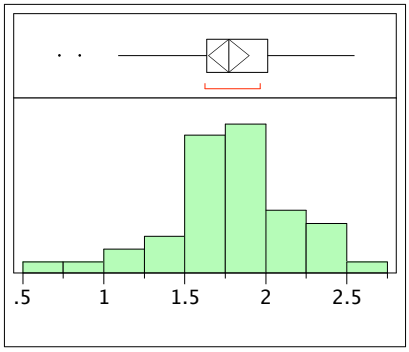
which in this case means, that the stem is 1.2, and the following values are included in the sample: 1.20 once, 1.22 twice, 1.27 thrice and 1.29 once (1.21 for example isn't included). The leaves may be written in multiple rows due to space issues.

Example. *(Continued) JMP prints the following stem and leaf diagram (again, a little tidied up compared to the default output)* **[8.3]**

Nicotinedata: Distribution

Distributions

Content



Quantiles

100.0%	maximum	2.5500
99.5%		2.5500
97.5%		2.5478
90.0%		2.3070
75.0%	quartile	2.0150
50.0%	median	1.7700
25.0%	quartile	1.6325
10.0%		1.2530
2.5%		0.7232
0.5%		0.7200
0.0%	minimum	0.7200

Stem and Leaf

Stem	Leaf	Count
2	6	1
2	45	2
2	233	3
2	00111	5
1	88888999999	11
1	6666777777	10
1	4455	4
1	2	1
1	1	1
0	9	1
0	7	1

0|7 represents 0.7

In this case, the values have first been rounded off to two decimals.

The *sample quantile* $q(f)$ is a numerical value, such that $100f$ % of the sample values are $\leq q(f)$. In particular, it is defined that $q(0) = x_{\min}$ and $q(1) = x_{\max}$. In addition to the minimum and the maximum, other common sample quantiles are the *sample median* $q(0.5)$, the *lower quartile*

$q(0.25)$ and the *upper quartile* $q(0.75)$. Yet other commonly used sample quantiles are the *quintiles*

$$q(0.2)\ ,\ q(0.4)\ ,\ q(0.6)\ ,\ q(0.8),$$

the *deciles*

$$q(0.1)\ ,\ q(0.2)\ ,\ q(0.3)\ ,\ q(0.4)\ ,\ q(0.5)\ ,\ q(0.6)\ ,\ q(0.7)\ ,\ q(0.8)\ ,\ q(0.9)$$

and the *centiles*

$$q(0.01)\ ,\ q(0.02)\ ,\ q(0.03)\ ,\ldots,\ q(0.99).$$

The difference $q(0.75) - q(0.25)$ is the *interquartile range.*

The following may be a better definition to the sample quantile: $q(f)$ is such a numerical value that at most $100f$ % of the sample values are $< q(f)$ and at most $(1-f)100$ % of the sample values are $> (q(f))$. The sample quantiles are however not unambiguously defined this way. There are many ways to define the sample quantiles so that they will be unambiguous (see exercises). Statistical programs usually print a collection of sample quantiles according to one of such definitions (see the previous example).

The sample quantiles mentioned above are realized values. It is of course possible to define the corresponding random variables $Q(f)$, for example the sample median $Q(0.5)$. The probability distributions of these variables are however very complicated.

A *quantile plot* is obtained by first sorting the sample values $x_1, x_2, \ldots, x_n$ in increasing order:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

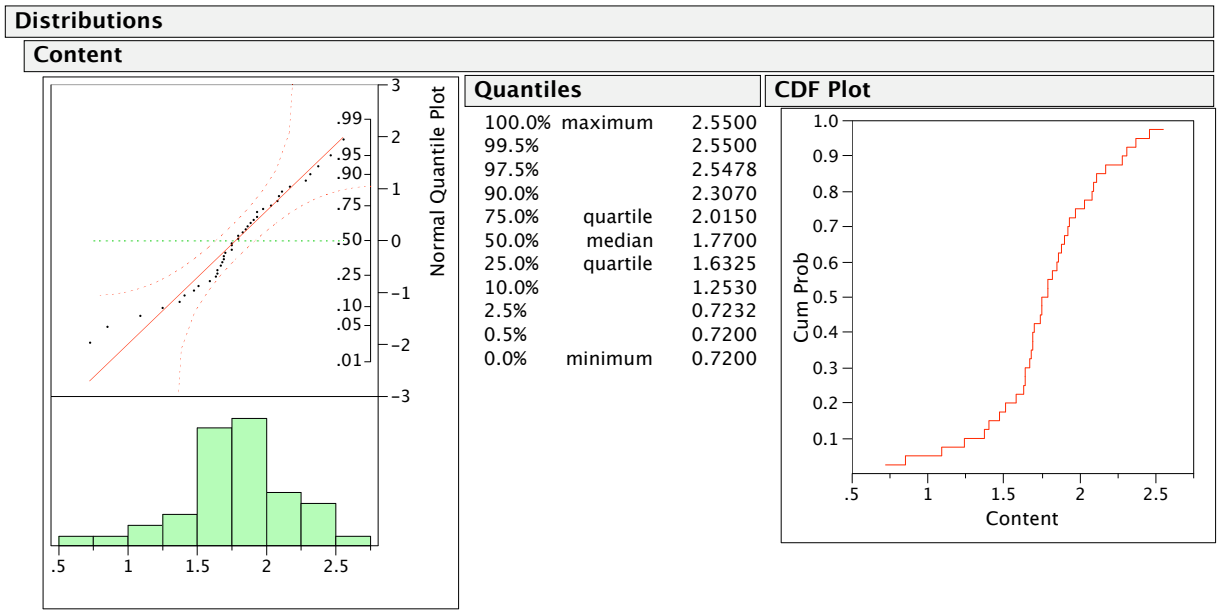
(where $x_{(i)}$ is the i:th smallest sample value). Then a suitable number f is computed for every sample value $x_{(i)}$. Such number is often chosen to be

$$f_i = \frac{i - 3/8}{n + 1/4}.$$

Finally, the dots $(f_i, x_{(i)})$ $(i = 1, \ldots, n)$ can be plotted as a point plot or a step line. The result is a quantile plot. If the data is displayed using a step plot, the result is an *empirical cumulative distribution function.*

Example. *(Continued) JMP plots exactly the cumulative distribution* [**8.3**]

function (the figure on the right):



Population values have a distribution that can be very difficult to define accurately. There are often though good reasons to assume that the distribution is somewhat normal. In other words, the cumulative distribution function is often fairly well approximated by some cumulative distribution function of the normal distribution $\mathrm{N}(\mu, \sigma^2)$. If in doubt, the first thing to do is to examine a graphical display. This can be done by comparing the sample quantiles to the relatives of the normal distribution.

Often also the last!

If the cumulative distribution function is F, its *quantile* $q(f)$ is a number such that $F\big(q(f)\big) = f$. If the quantiles of the normal distribution $\mathrm{N}(\mu, \sigma^2)$ are denoted by $q_{\mu,\sigma}(f)$, then

Note that in spite their similar notation, the distribution's quantile and the sample quantile are different concepts.

$$q_{\mu,\sigma}(f) = \mu + \sigma\Phi^{-1}(f),$$

where Φ is the cumulative distribution function of the standard normal distribution $\mathrm{N}(0, 1)$.

Quite a good approximation is $\Phi^{-1}(f) \cong 4.91 f^{0.14} - 4.91(1-f)^{0.14}$.

By plotting the points $\big(x_{(i)}, q_{0,1}(f_i)\big)$ $(i = 1, \ldots, n)$ as a scatter plot or a step line, the result is a *normal quantile plot*. If the population distribution actually is $\mathrm{N}(\mu, \sigma^2)$, then the plot should be somewhat a straight line, because then ideally

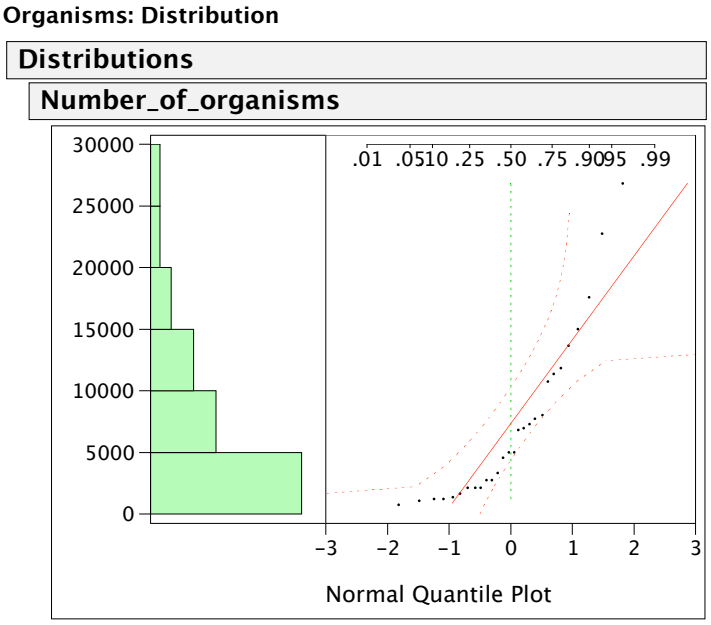
$$q_{0,1}(f_i) = \Phi^{-1}(f_i) = \frac{q_{\mu,\sigma}(f_i) - \mu}{\sigma} \cong \frac{x_{(i)} - \mu}{\sigma}.$$

Near the ends of the plot there may be some scattering, but at least in the middle the plot should be a quite straight line. If that is not the case, it can be tentatively concluded that the population distribution is not normal. In the previous example the plot on the left is a normal quantile plot. The population distribution can be, according to this figure, considered normal although some scattering can be observed.

Example. *In this example, the number of organisms (per square me-* **[8.5]**

ter) has been measured $n = 28$ times. JMP prints the following normal quantile plot, from which it can be seen that the population distribution cannot be considered normal. This can naturally be clearly seen from the bar chart as well.

The axes are reversed!



There are other graphical methods to examine normality, for example the *normal probability plot.*

1.4 Sampling distributions [8.4]

The distribution of a random variable is the *sampling distribution.* The distributions of some random variables are often complicated, although the population distribution itself may be ”nice” (for example normal). Such variables are especially sample quantiles when considered random variables.

1.4.1 Sampling distributions of means [8.5]

If the expectation of the sampling distribution is μ and its variance is σ^2, then the expectation of the sample mean is

$$\mathrm{E}(\overline{X}) = \mu$$

and its variance is

$$\mathrm{var}(\overline{X}) = \frac{\sigma^2}{n}$$

(n is sample size). The standard deviation of the sample mean or its *standard error* is $\sigma/\sqrt{n}$ and it decreases as the sample size increases.

If the population distribution is a normal distribution $\mathrm{N}(\mu, \sigma^2)$, then the distribution of the sample mean is also a normal distribution, namely $\mathrm{N}(\mu, \sigma^2/n)$. The distribution of $(\overline{X})$ is however almost always normal in

Not all distributions have an expectation. Some distributions on the other hand have only an expectation but not a finite variance.

other cases, if just n is great enough (and the population distribution has an expected value and a finite variance). This is ensured by a classical approximation result:

The central limit theorem. *If the expectation of the population distribution is μ and its (finite) variance is σ, then the cumulative distribution function of the standardized random variable*

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

approaches the cumulative distribution function Φ of the standard normal distribution in the limit as n increases.

There are also versions of the theorem where the distributions are not assumed to be identical, only independent. Then, if the expectations of the sample values $X_1, \ldots, X_n$ are $\mu_1, \ldots, \mu_n$ and their variances are $\sigma_1, \ldots, \sigma_n$, let's choose

$\mu = \frac{1}{n}(\mu_1 + \cdots + \mu_n)$,

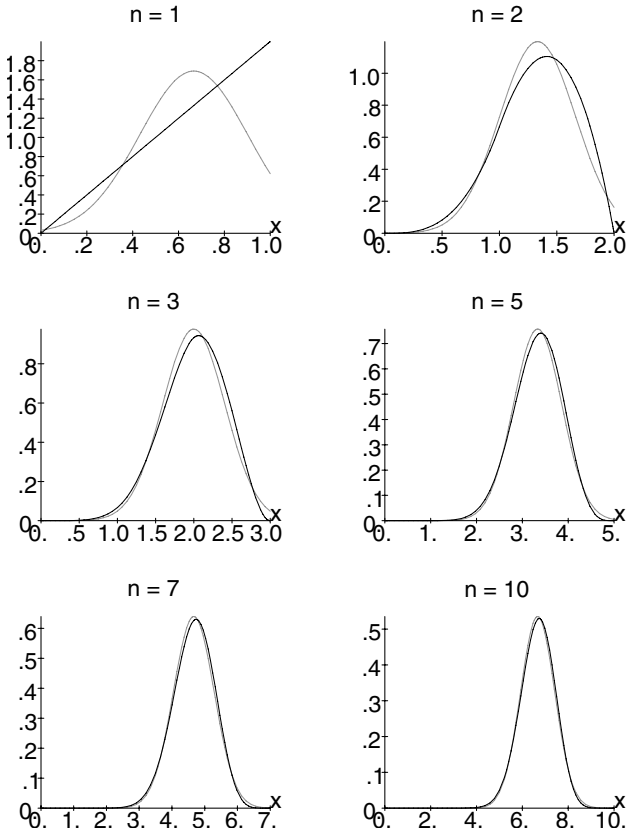
$\sigma^2 = \frac{1}{n}(\sigma_1^2 + \cdots + \sigma_n^2)$.

Now the theorem holds as long as yet some other (weak) assumption is made. A famous such assumption is *Lindeberg's condition*. Jarl Lindeberg (1876–1932), by the way, was a Finnish mathematician!

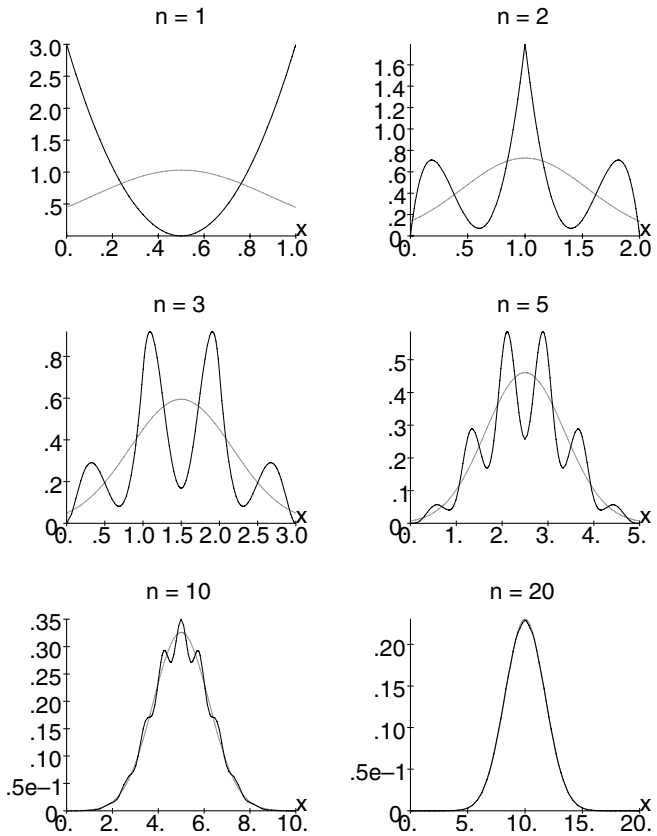
Usually a sample size of $n = 30$ is enough to normalize the distribution of $\overline{X}$ accurately enough. If the population distribution is "well-shaped" (unimodal, almost symmetric) to begin with, a smaller sample size is enough (for example $n = 5$).

Example. *Starting from a strongly asymmetric distribution, density functions of the sum $X_1 + \cdots + X_n$ for different sample sizes are formed according to the first plot series on the next page (calculated with Maple). If, on the other hand, in the beginning there is a symmetric, but strongly bimodal, distribution, the density functions of the sum $X_1 + \cdots + X_n$ resemble ones in the second plot series on the next page. The sample size of $n = 7$ is indeed enough to normalize the distribution $\overline{X}$ quite accurately in the first case, but in the second the sample size of $n = 20$ is required.*

1. plot series:



2. plot series:



Example. *The diameter of a machine part should be $\mu = 5.0$ mm (the expectation). It is known that the population standard deviation is $\sigma = 0.1$ mm. By measuring the diameter of $n = 100$ machine parts a sample mean of $\overline{x} = 5.027$ mm was calculated. Let's calculate the probability that a random sample from a population having the distribution $N(5, 0.1^2)$ would have a sample mean that differs from 5 at least as much as this sample does:* [8.7]

$$\mathrm{P}(|\overline{X} - \mu| \geq 0.027 \text{ mm}) = 2\mathrm{P}\Big(\frac{\overline{X} - 5.0}{0.1/\sqrt{100}} \geq 2.7\Big) = 0.0069$$

(from the standard normal distribution according to the Central limit theorem). This probability is quite small, which raises suspicion: It is quite probable that the actual μ is greater. The calculations in MATLAB are:

```
>> mu=5.0;
   sigma=0.1;
   n=100;
   x_viiva=5.027;

>> 2*(1-normcdf(x_viiva,mu,sigma/sqrt(n)))

ans =
    0.0069
```

An expectation and a variance can be calculated for the difference of two independent samples $\overline{X}_1$ and $\overline{X}_2$

If the random variables X and Y are independent, then $\mathrm{var}(X \pm Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.

$$\mathrm{E}(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2 \quad \text{and} \quad \mathrm{var}(\overline{X}_1 - \overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

where μ_1, μ_2 and σ_1^2, σ_2^2 are the corresponding expectations and variances of the population standard deviations and n_1, n_2 are the sample sizes. If the sample sizes are great enough, a standardized random variable

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

has, according to the Central limit theorem, a distribution that is close (when considering cumulative distributions) to a normal distribution $\mathrm{N}(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$. (The distribution is exactly normal if the population distributions are normal.)

The sum and the difference of two normally distributed random variables are also normally distributed

Example. *The drying times of two paints A and B were compared by measuring $n = 18$ samples. The population variances of the paints are known to be $\sigma_\mathrm{A} = \sigma_\mathrm{B} = 1.0$ h. The difference of the sample means was $\overline{x}_\mathrm{A} - \overline{x}_\mathrm{B} = 1.0$ h. Could this result be possible, even though the population expectations are the same (meaning $\mu_\mathrm{A} = \mu_\mathrm{B}$)? Let's calculate* [8.8]

$$\mathrm{P}(\overline{X}_\mathrm{A} - \overline{X}_\mathrm{B} \geq 1.0 \text{ h}) = \mathrm{P}\Big(\frac{\overline{X}_\mathrm{A} - \overline{X}_\mathrm{B} - 0}{\sqrt{1.0^2/18 + 1.0^2/18}} \geq 3.0\Big) = 0.0013.$$

The probability is so small that the result most likely isn't a coincidence, so indeed $\mu_\mathrm{A} > \mu_\mathrm{B}$. If there had been $\overline{x}_\mathrm{A} - \overline{x}_\mathrm{B} = 15$ min, the result would be

$$\mathrm{P}(\overline{X}_\mathrm{A} - \overline{X}_\mathrm{B} \geq 0.25 \text{ h}) = 0.2266,$$

This result might well be a coincidence. These calculations in MATLAB *are:*

```
>> mu=0;         % The paints have the same expectations
   sigma_A=1.0;
   sigma_B=1.0;
   n_A=18;
   n_B=18;
   difference=1.0;   % The sample mean of paint A - the sample mean of paint B

> 1-normcdf(difference,mu,sqrt(sigma_A/n_A+sigma_B/n_B))

ans =
    0.0013

>> difference=0.25;

>> 1-normcdf(difference,mu,sqrt(sigma_A/n_A+sigma_B/n_B))

ans =
    0.2266
```

1.4.2 The sampling distribution of the sample variance

[8.6]

The proofs are quite complicated and are omitted

The sampling distribution of the sample variance is a difficult concept, unless it can be assumed that the population distribution is normal. Let's make this assumption, so the sampling distribution of the sample variance can be formed using the χ^2-distribution.

If random variables $U_1, \ldots, U_v$ have the standard normal distribution and they are independent, a random variable

$$V = U_1^2 + \cdots + U_v^2$$

has the χ^2-*distribution.* Here v is a distribution's parameter, the *number of degrees of freedom.* The density function of the distribution is

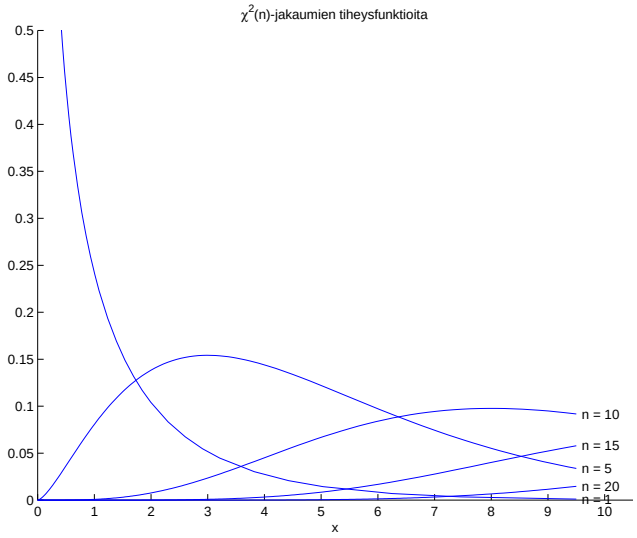
"chi-square-distribution"

$$g(x) = \begin{cases} \dfrac{1}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})} x^{\frac{v-2}{2}} e^{-\frac{x}{2}}, \text{ when } x > 0 \\ 0, \text{ when } x \leq 0, \end{cases}$$

where Γ is the gamma-function $\Gamma(y) = \int_0^\infty t^{y-1}e^{-t}\,\mathrm{d}t$. Despite its difficult form, the probabilities of the χ^2-distribution are numerically quite easily computed. Here are presented some density functions of the χ^2-distribution (the number of degrees of freedom is denoted by n, the func-

The gamma-function is a continuous generalization of the factorial $n!$. It can be easily seen that $\Gamma(1) = 1$ and by partial integration that $\Gamma(y+1) = y\Gamma(y)$. Thus $\Gamma(n) = (n-1)!$ when n is a positive integer. It is more difficult to see that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

tions are calculated with MATLAB):



It is easily seen that $\mathrm{E}(V) = v$ and it can be shown that $\mathrm{var}(V) = 2v$. As a consequence of the Central limit theorem for large values of v (about $v \geq 30$) the χ^2-distribution is very close to normal distribution $\mathrm{N}(v, 2v)$.

That is the reason why the χ^2-distribution is tabulated to at most 30–40 degrees of freedom.

If $X_1, \ldots, X_n$ is a sample of $\mathrm{N}(\mu, \sigma^2)$-distributed population, then the random variables $(X_i - \mu)/\sigma$ have the standard normal distribution and they are independent. Additionally, the sum

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

is χ^2-distributed with n degrees of freedom. But the sum is not the sample variance! On the other hand a similar random variable

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}$$

calculated from the sample variance is also χ^2-distributed , but with $n-1$ degrees of freedom. It is important to notice that in this case there is no approximation such as the Central limit theorem that can be used. The distribution has to be normal.

This is difficult to prove!

Example. *The lifetimes of $n = 5$ batteries have been measured. The standard deviation is supposed to be $\sigma = 1.0$ y. The measured lifetimes were 1.9 y, 2.4 y, 3.0 y, 3.5 y and 4.2 y. Sample variance can be calculated to be $s^2 = 0.815$ y^2. Furhermore* [8.10]

$$\mathrm{P}(S^2 \geq 0.815\ y^2) = \mathrm{P}\Big(\frac{(n-1)S^2}{\sigma^2} \geq 3.260\Big) = 0.5153$$

(by using χ^2-distribution with $n - 1 = 4$ degrees of freedom.) The value s^2 is thus quite "common" (close to median). There's no reason to doubt the supposed standard deviation 1.0 y. The calculations with MATLAB:

```
>> mu=3;
```

```
   sigma=1;
   n=5;
   otos=[1.9 2.4 3.0 3.5 4.2];

>> s=std(otos)

s =
    0.9028

>> 1-chi2cdf((n-1)*s^2/sigma^2,n-1)

ans =
    0.5153
```

1.4.3 t-Distribution

[8.7]

Earlier when considering the sample mean, it was required to know the standard deviation σ. If the standard deviation is not known, it is possible to proceed, but instead of a normal distribution, a t-distribtution (or Student's distribution) is used. Additionally, the Central limit theorem isn't used, but the population distribution has to be normal.

Again, the proofs are complicated and will be omitted

If random variables U and V are independent, U has the standard normal distribution and V is χ^2-distributed with v degrees of freedom, a random variable

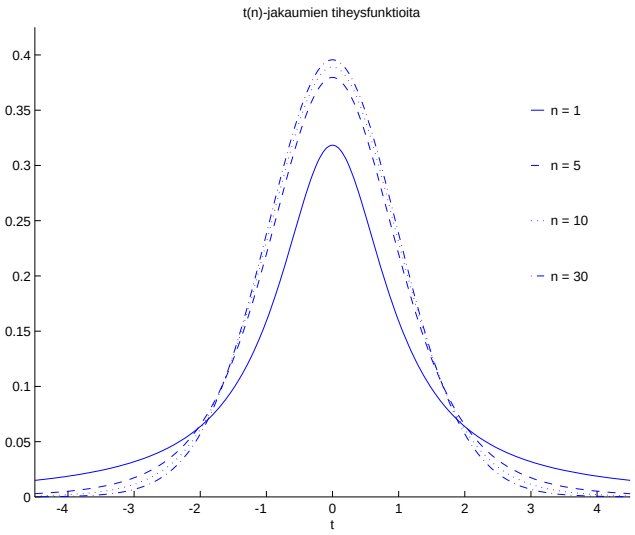
$$T = \frac{U}{\sqrt{V/v}}$$

has a *t-distribution with v degrees of freedom.* The density function of the distribution is

The distribution was originally used by chemist William Gosset (1876–1937) a.k.a. "Student".

$$g(x) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi v}\,\Gamma(\frac{v}{2})}\left(1+\frac{1}{v}x^2\right)^{-\frac{v+1}{2}}.$$

Here are a few examples of density functions of the t-distribution (with n degrees of freedom, calculated with MATLAB):



The t-distribution is unimodal and symmetric about the origin and somewhat resembles the standard normal distribution. It approaches the standard normal distribution in the limit as $v \to \infty$, but that is not because of the Central limit theorem.

But what?

If the population distribution is normal, then the sample mean $\overline{X}$ and the sample variance s^2 are independent random variables. Because of this, the random variables

This independence is quite difficult to prove and somewhat surprising!

$$U = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad V = \frac{(n-1)S^2}{\sigma^2}$$

calculated from those are also independent. The preceding has the standard normal distribution and the latter has χ^2-distribution with $n-1$ degrees of freedom. Thus a random variable

$$T = \frac{U}{\sqrt{V/(n-1)}} = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has the t-distribution with $n-1$ degrees of freedom.

Example. *The outcome of a chemical process is measured. The outcome should be $\mu = 500$ g/ml (supposed population expectation). The outcome was measured in $n = 25$ batches, when the sample mean was $\overline{x} = 518$ g/ml and the standard deviation $s = 40$ g/ml. Let's calculate*

$$\mathrm{P}\Big(\frac{\overline{X} - \mu}{S/\sqrt{n}} \geq \frac{518 - 500}{40/\sqrt{25}}\Big) = \mathrm{P}(T \geq 2.25) = 0.0169$$

(by using a t-distribution with $n - 1 = 24$ degrees of freedom.) This probability is quite small, so the result wasn't a coincidence and thus the outcome is actually better than it was thought to be. The calculations with MATLAB:

```
>> mu=500;
   n=25;
   x_viiva=518;
   s=40;

>> 1-tcdf((x_viiva-mu)/(s/sqrt(n)),n-1)

ans =
    0.0169
```

Although the t-distribution is derived with the assumption that the population distribution is normal, it is still quite *robust*, for the preceding random variable T is almost t-distributed as long as the population distribution is normal-like (unimodal, almost symmetric). That is because the standard deviation S of population distributions with relatively large sample sizes n is so accurately $= \sigma$, that the Central limit theorem is used in some ways. Thus the t-distribution is very useful in many situations.

1.4.4 F-distribution [8.8]

Ronald Fisher (1880–1962), a pioneer in statistics

George Snedecor (1881–1974)

Comparing the standard deviations of two samples can be done with their sample variances and by using the F-distribution (also known as Fisher's distribution or Snedecor's distribution).

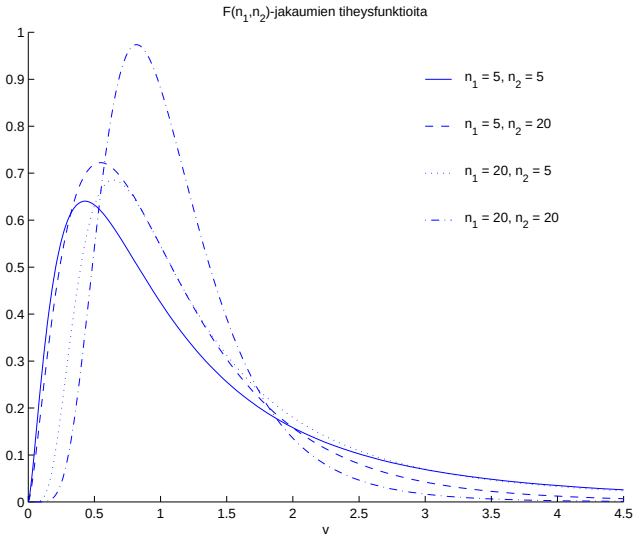
If random variables V_1 and V_2 are independent and they are χ^2-distributed with v_1 and v_2 degrees of freedom correspondingly, a random variable

$$F = \frac{V_1/v_1}{V_2/v_2}$$

has the *F-distribution with v_1 and v_2 degrees of freedom.* In that case, random variable $1/F$ has also F-distribution, namely with v_2 and v_1 degrees of freedom. The formula for the density function of the F-distribution is quite complicated:

$$g(x) = \begin{cases} \left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} \frac{\Gamma(\frac{v_1+v_2}{2})}{\Gamma(\frac{v_1}{2})\Gamma(\frac{v_2}{2})} x^{\frac{v_1-2}{2}} \left(1 + \frac{v_1}{v_2}x\right)^{-\frac{v_1+v_2}{2}}, & \text{when } x > 0 \\ 0, \text{ when } x \leq 0. \end{cases}$$

A few examples of these density functions (with n_1 and n_2 degrees of freedom, calculated with MATLAB):



If S_1^2 and S_2^2 are the sample variances of two independent samples, the corresponding populations are normally distributed with standard deviations σ_1 and σ_2 and the sample sizes are n_1 and n_2, then random variables

$$V_1 = \frac{(n_1-1)S_1^2}{\sigma_1^2} \quad \text{and} \quad V_2 = \frac{(n_2-1)S_2^2}{\sigma_2^2}$$

are independent, χ^2-distributed with n_1-1 and n_2-1 degrees of freedom. Thus a random variable

$$F = \frac{V_1/(n_1-1)}{V_2/(n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

is F-distributed with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

The F-distribution can be used to compare sample variances using samples, see sections 2.9 and 3.7. It is however a fairly limited tool for that purpose, and statistical software usually use other methods.

E.g. Bartlett's test or Levene's test.

Example. *Let's consider a case where realized sample variances are* $s_1^2 = 0.20$ *and* $s_2^2 = 0.14$ *and the sample sizes are* $n_1 = 25$ *and* $n_2 = 30$. *Additionally, the corresponding standard deviations are the same meaning* $\sigma_1 = \sigma_2$. *Let's calculate*

$$\mathrm{P}\Big(\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \geq \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}\Big) = \mathrm{P}(F \geq 1.429) = 0.1787$$

(by using a F-distribution with $n_1 - 1 = 24$ *and* $n_2 - 1 = 29$ *degrees of freedom). The tail probability is therefore quite large, the value is in the "common" area of the distribution and there's no actual reason to doubt that the sample deviations wouldn't be the same. The calculations with MATLAB:*

```
>> n_1=25;
   n_2=30;
   s_1_toiseen=0.20;
   s_2_toiseen=0.14;

>> 1-fcdf(s_1_toiseen/s_2_toiseen,n_1-1,n_2-1)

ans =
    0.1787
```

Primarily the F-distribution is used in analysis of variance that will be considered later.

Chapter 2

ONE- AND TWO-SAMPLE ESTIMATION

Estimation of a numerical value related to the population distribution is, together with hypothesis testing, a basic method in the field of *classical statistical inference.*

Another basic field in statisticals methods is *Bayesian statistics* that is not considered in this course.

2.1 Point Estimation and Interval Estimation

[9.3]

The purpose of *point estimation* is to estimate some population-related numerical value, a *parameter* θ, by using the sample. Such a parameter is for example the population's expectation μ, which can be estimated by the sample mean $\overline{X}$. The realized value calculated from the sample is a numerical value that estimates θ. This value is called the *estimate*, and it is denoted by $\hat{\theta}$. The estimate is calculated from the sample values by using some formula or some numerical algorithm.

On the other hand, the estimate calculated by applying the estimation formula or algorithm to a sequence of random variables $X_1, \ldots, X_n$ is a random variable as well, and it is denoted by $\hat{\Theta}$. This random variable is called the *estimator.*

Remember: random variables are denoted with upper case letters, realized values with lower case letters.

There may be different estimators for the same parameter, and different parameters can be estimated by the same function of the sample. For example the population expectation could be estimated by the sample median. The quality of the estimates depends on the symmetry of the population distribution about its expectation. Moreover, the sample mean is an estimator of the population median—a better estimator of the population median is of course the sample median.

When estimating the population mean μ, variance σ^2 and median m the above mentioned concepts are:

Parameter θ	Estimate $\hat{\theta}$	Estimator $\hat{\Theta}$
μ	$\hat{\mu} = \overline{x}$	$\overline{X}$
σ^2	$\widehat{\sigma^2} = s^2$	S^2
m	$\hat{m} = q(0.5)$	$Q(0.5)$

A random variable that is used as an estimator of a population parameter is called a *point estimator*. If there is no systematic error in its value, in other words its expectation $\mathrm{E}(\hat{\Theta})$ equals the actual parameter value, it is said that the estimator is *unbiased*. If, on the other hand, $\mathrm{E}(\hat{\Theta}) \neq \theta$, then it's said that the estimator $\mathrm{E}(\hat{\Theta})$ is *biased*. (Here it is assumed, of course, that $\mathrm{E}(\hat{\Theta})$ exists!)

If μ is the population expectation, then the estimator $\overline{X}$ (the sample mean as a random variable) is unbiased, because $\mathrm{E}(\overline{X}) = \mu$. It will now be shown that the sample variance S^2 is an unbiased estimator for the population variance σ^2. Firstly, S^2 can be written in the form

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

Write $X_i - \overline{X} = (X_i - \mu) - (\overline{X} - \mu)$ and expand the square.

Thus,

$$\begin{aligned}\mathrm{E}(S^2) &= \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}\big((X_i - \mu)^2\big) - \frac{n}{n-1}\mathrm{E}\big((\overline{X} - \mu)^2\big) \\ &= \frac{n}{n-1}\sigma^2 - \frac{n}{n-1}\frac{\sigma^2}{n} = \sigma^2.\end{aligned}$$

The smaller the variance

$$\mathrm{var}(\hat{\Theta}) = \mathrm{E}\big((\hat{\Theta} - \theta)^2\big)$$

of the unbiased point estimator $\hat{\Theta}$ is, the more probable it is that it is close to its expectation. It's said that an estimator is more *efficient*, the smaller its variance is. A biased estimator can be good as well, in the sense that its *mean square error* $\mathrm{E}\big((\hat{\Theta} - \theta)^2\big)$ is small.

The purpose of *interval estimation* is, by calculating from a sample, to create an interval in which the correct parameter value θ belongs, at least at some known, high enough probability. The interval may be one- or two-sided. In a two-sided interval, both the endpoints θ_{L} (left or lower) and θ_{U} (right or upper) are estimated. In one-sided interval, only the other endpoint is estimated (the other is trivial, for example $\pm\infty$ or 0.) Let's consider first the two-sided intervals.

Here also, the estimates $\hat{\theta}_{\mathrm{L}}$ and $\hat{\theta}_{\mathrm{U}}$ are realized values calculated from the sample. The estimators $\hat{\Theta}_{\mathrm{L}}$ and $\hat{\Theta}_{\mathrm{U}}$, for their part, are random variables. The basic idea is to find estimators, in one way or another, so that

So the endpoints $\hat{\Theta}_{\mathrm{L}}$ and $\hat{\Theta}_{\mathrm{U}}$ are the random variables, not the parameter θ!

$$\mathrm{P}(\hat{\Theta}_{\mathrm{L}} < \theta < \hat{\Theta}_{\mathrm{U}}) = 1 - \alpha,$$

where α is a *given* value (often 0.10, 0.05 or 0.01). The realized interval $(\hat{\theta}_{\mathrm{L}}, \hat{\theta}_{\mathrm{U}})$ is then called a $100(1-\alpha)$ % *confidence interval*. The value $1-\alpha$ is the interval's *degree of confidence*, and its endpoints are the *lower* and the *upper confidence limit*.

The greater the degree of confidence is required, the wider the confidence interval will be, and a degree of confidence close to 100 % usually leads to intervals that are too wide to be interesting. Additionally, the

condition $\mathrm{P}(\hat{\Theta}_\mathrm{L} < \theta < \hat{\Theta}_\mathrm{U}) = 1 - \alpha$ doesn't tell, how the interval is chosen. It is often required that the interval is *symmetric,* in other words

$$\mathrm{P}(\theta \leq \hat{\Theta}_\mathrm{L}) = \mathrm{P}(\theta \geq \hat{\Theta}_\mathrm{U}) = \frac{\alpha}{2}.$$

(Another alternative would be to seek an interval that is the shortest possible but that often leads to complicated calculations.)

2.2 Single Sample: Estimating the Mean [9.4]

When point estimating the population expectation μ, a natural unbiased estimator is the sample mean $\overline{X}$, whose variance is σ^2/n. Here σ^2 is the population variance, which is for now supposed to be known. With large sample sizes n such estimation is quite accurate indeed.

The interval estimation of the expectation is based on the fact that the distribution of the random variable

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

approaches, according to the Central limit theorem, the standard normal distribution $\mathrm{N}(0,1)$ in the limit as n increases. Let's now choose a quantile $z_{\alpha/2}$ of the distribution so that $\mathrm{P}(Z \geq z_{\alpha/2}) = 1 - \Phi(z_{\alpha/2}) = \alpha/2$, so that (by symmetry) also $\mathrm{P}(Z \leq -z_{\alpha/2}) = \Phi(-z_{\alpha/2}) = \alpha/2$. Then

Φ is the cumulative distribution function of the standard normal distribution.

$$\mathrm{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

On the other hand, the double inequality

$$-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

is equivalent to the double inequality

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

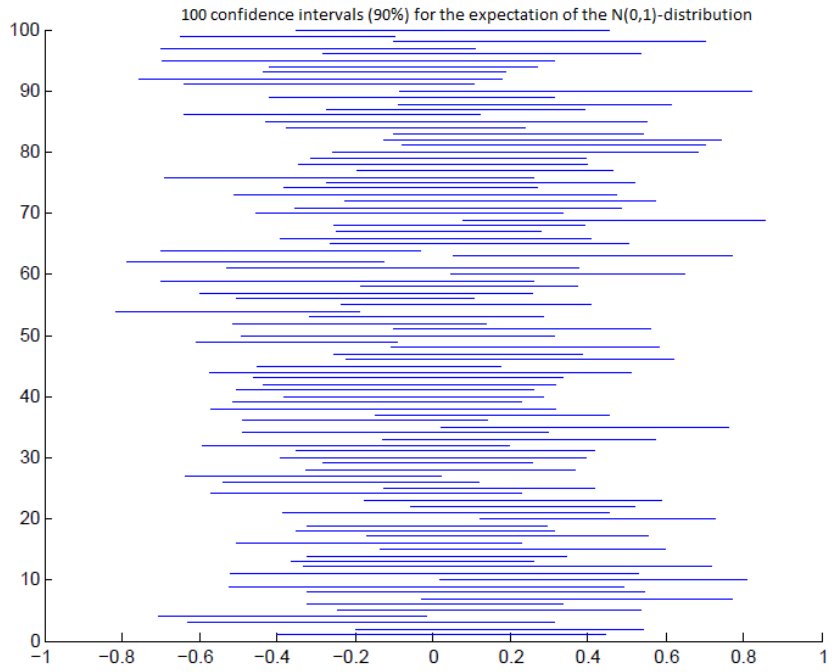
Thus, if the realized sample mean is $\overline{x}$, the $100(1-\alpha)$ % confidence limits are chosen to be

$$\hat{\mu}_\mathrm{L} = \overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \hat{\mu}_\mathrm{U} = \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

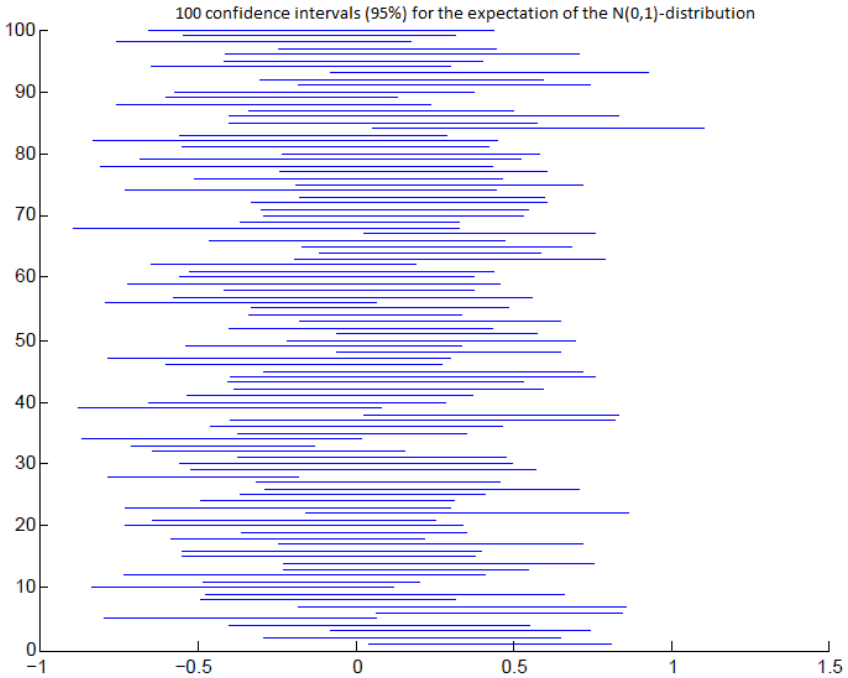
Here are presented 100 cases of 90 %, 95 % and 99 % confidence intervals for the standard normal distribution simulated with MATLAB.

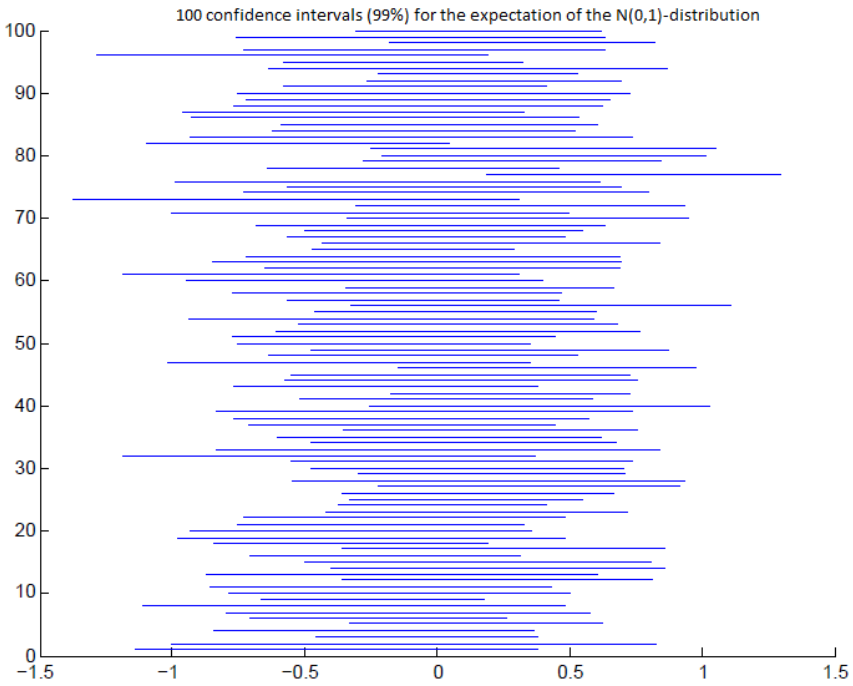
For each 95 % confidence interval case we generate twenty standard normal numbers, compute $\overline{X}$, and plot the line segment with endpoints $\overline{X} \pm 1.96/\sqrt{20}$.

Let's begin with the 90 % confidence intervals.



Note how about ten intervals don't include the correct expectation $\mu = 0$. Many of the intervals are even disjoint. When moving to a higher degree of confidence, the intervals become longer but are more likely to include the correct expectation:





Example. *This is about zinc concentration in* $n = 36$ *different locations. The sample mean of the measurements is* $\overline{x} = 2.6$ *g/ml. The population standard deviation is known to be* $\sigma = 0.3$ *g/ml. If* $\alpha = 0.05$*, when* $z_{0.025} = 1.960$*, by calculating we get* $\hat{\mu}_{\mathrm{L}} = 2.50$ *g/ml and* $\hat{\mu}_{\mathrm{U}} = 2.70$ *g/ml. If again* $\alpha = 0.01$*, when* $z_{0.005} = 2.575$*, we get* $\hat{\mu}_{\mathrm{L}} = 2.47$ *g/ml and* $\hat{\mu}_{\mathrm{U}} = 2.73$ *g/ml so the interval is longer.* [**9.2**]

If a confidence interval is determined by a symmetric distribution, which is the case for the expectation, the limits are of the form $\hat{\theta} \pm b$, where $\hat{\theta}$ is the point estimate. The value b is in that case called the *estimation error*. For the expectation the estimation error is $b = z_{\alpha/2}\sigma/\sqrt{n}$. So if the estimation error is wanted to be at most b_0, the sample size n must be chosen so that

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq b_0 \quad \text{so} \quad n \geq \left(\frac{z_{\alpha/2}\sigma}{b_0}\right)^2.$$

Thus, if in the previous example the estimation error is wanted to be at most $b_0 = 0.05$ g/ml, the sample size should be at least $n = 139$.

In the above, the confidence intervals have always been two-sided. If only the *lower confidence limit* is wanted for the sample mean μ, let's choose a quantile z_α of the standard normal distribution, for which $\mathrm{P}(Z \geq z_\alpha) = 1 - \Phi(z_\alpha) = \alpha$; then also $\mathrm{P}(Z \leq -z_\alpha) = \Phi(-z_\alpha) = \alpha$. Now the inequality

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha$$

is equivalent with the inequality

$$\mu > \overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}$$

and we obtain the wanted $100(1-\alpha)$ % *lower confidence limit.*

$$\hat{\mu}_{\mathrm{L}} = \overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Correspondingly, the $100(1-\alpha)$ % *upper confidence limit* is $\hat{\mu}_{\mathrm{U}} = \overline{x} + z_\alpha \sigma/\sqrt{n}$.

Example. *A certain reaction time was measured on $n = 25$ subjects. From previous tests it is known that the standard deviation of the reaction times is $\sigma = 2.0$ s. The measured sample mean of the samples is $\overline{x} = 6.2$ s. Now $z_{0.05} = 1.645$ and the 95 % upper confidence limit for the expectations of the reaction times is $\hat{\mu}_{\mathrm{U}} = 6.86$ s.* **[9.4]**

Above it was required that population variance σ^2 is known. If the population variance is not known, it is possible to proceed but a normal distribution will be replaced with a t-distribution (The central limit theorem isn't used here: the population distribution has to be normal.) Let's now begin with a random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

that has a t-distribution with $n-1$ degrees of freedom. Let's find a quantile $t_{\alpha/2}$ for which holds $\mathrm{P}(T \geq t_{\alpha/2}) = \alpha/2$. Then, because of the symmetry of the t-distribution, $\mathrm{P}(T \leq -t_{\alpha/2}) = \alpha/2$ and $\mathrm{P}(-t_{\alpha/2} < T < t_{\alpha/2}) = 1-\alpha$, just like for the normal distribution. By proceeding as above, we obtain the $100(1-\alpha)$ % confidence limits for the population expectation μ

$$\hat{\mu}_{\mathrm{L}} = \overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \quad \text{and} \quad \hat{\mu}_{\mathrm{U}} = \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}.$$

The estimation error of the estimate $\overline{x}$ is obviously in this case $b = t_{\alpha/2}s/\sqrt{n}$. But it is not known beforehand.

The corresponding one-sided confidence limits are

$$\hat{\mu}_{\mathrm{L}} = \overline{x} - t_\alpha \frac{s}{\sqrt{n}} \quad \text{and} \quad \hat{\mu}_{\mathrm{U}} = \overline{x} + t_\alpha \frac{s}{\sqrt{n}},$$

where quantile t_α is chosen so that $\mathrm{P}(T \geq t_\alpha) = \alpha$.

Example. *The contents of seven similar containers of sulfuric acid were measured. The mean value of these measurements is $\overline{x} = 10.0$ l, and their standard deviation is $s = 0.283$ l. Now $t_{0.025} = 2.447$ and the 95 % confidence interval is $(9.74\ l, 10.26\ l)$.* **[9.5]**

2.3 Prediction Intervals [9.6]

Often after interval estimation a corresponding *prediction interval* is wanted for the next measurement x_0. Naturally the corresponding random variable X_0 is considered independent of the sample's random variables $X_1, \ldots, X_n$ and identically distributed to them.

Assuming that the population distribution is a normal distribution $\mathrm{N}(\mu, \sigma^2)$, it is known that the difference $X_0 - \overline{X}$ is also normally distributed and

The sum and the difference of two independent normally distributed random variables are also normally distributed.

$$\mathrm{E}(X_0 - \overline{X}) = \mathrm{E}(X_0) - \mathrm{E}(\overline{X}) = \mu - \mu = 0$$

and

If random variables X and Y are independent, then $\mathrm{var}(X \pm Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.

$$\mathrm{var}(X_0 - \overline{X}) = \mathrm{var}(X_0) + \mathrm{var}(\overline{X}) = \sigma^2 + \frac{\sigma^2}{n} = \left(1 + \frac{1}{n}\right)\sigma^2.$$

Thus, the random variable

$$Z = \frac{X_0 - \overline{X}}{\sigma\sqrt{1 + 1/n}}$$

has the standard normal distribution. Here it is again assumed that the population variance σ^2 is known.

By proceeding just like before, but replacing $\sigma/\sqrt{n}$ with $\sigma\sqrt{1 + 1/n}$, we obtain the $100(1 - \alpha)$ % confidence interval for x_0

$$\overline{x} - z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}} < x_0 < \overline{x} + z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}},$$

in which it belongs to with probability $1 - \alpha$. The probability has to be interpreted so that it is the probability of an event

$$\overline{X} - z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}} < X_0 < \overline{X} + z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}}.$$

Thus the prediction interval takes into account the uncertainty of both the expectation and the random variable X_0.

Again, if the population standard deviation σ is not known, the sample standard deviation s must be used instead and instead of a normal distribution, a t-distribution must be used with $n-1$ degrees of freedom. A random variable $X_0 - \overline{X}$ is namely independent of the sample variance S^2, so

Again a difficult fact to prove.

$$T = \frac{Z}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{X_0 - \overline{X}}{S\sqrt{1 + 1/n}}$$

is t-distributed with $n-1$ degrees of freedom. The $100(1-\alpha)$ % prediction interval obtained for the value x_0 is then

$$\overline{x} - t_{\alpha/2}s\sqrt{1 + \frac{1}{n}} < x_0 < \overline{x} + t_{\alpha/2}s\sqrt{1 + \frac{1}{n}}.$$

Example. *The percentage of meat was measured in $n = 30$ packages of a low-fat meat product. The distribution was supposed to be normal. The sample mean is $\overline{x} = 96.2$ %, and the population standard deviation is $s = 0.8$ %. By using a t-quantile $t_{0.005} = 2.756$ (with 29 degrees of freedom) the 99 % confidence interval for the percentage of meat measured in yet another sample is obtained* (93.96 %, 98.44 %). [9.7]

Don't confuse the meat percentages with the confidence interval percentages!

One use of prediction intervals is to *find outliers*. An observation is considered to be an outlier if it doesn't belong to the prediction interval that is obtained after the observation in question is removed from the sample.

See the example in section 1.3

One-sided prediction intervals could be also formulated by using similar methods.

2.4 Tolerance Limits [9.7]

One form of interval estimation is the *tolerance interval* that is used in, among other things, defining the statistical behavior of processes.

If a population distribution is a known normal distribution $\mathrm{N}(\mu, \sigma^2)$, its $100(1-\alpha)$ % tolerance interval is an interval $(\mu - k\sigma, \mu + k\sigma)$ such that $100(1-\alpha)$ % of the distribution belongs to it. The interval is given by giving the corresponding value of k and is often presented in the form $\mu \pm k\sigma$. Thus, for example a 95 % tolerance interval is $\mu \pm 1.96\sigma$. This requires that μ and σ are known.

The μ ja σ of a population are usually however unknown. The tolerance interval is then obtained by using the corresponding statistics $\overline{x}$ and s, as follows

Sometimes $\overline{x} \pm k\frac{s}{\sqrt{n}}$.

$$\overline{x} \pm ks.$$

These are however realized values of the random variables $\overline{X} \pm kS$ and thus, the tolerance interval is correct only with the probability of $1-\gamma$, which depends on the chosen value of k (and the sample size n). That's why k is chosen so that the interval $\overline{X} \pm kS$ contains at least $100(1-\alpha)$ % of the distribution at the probability of $1-\gamma$ (significance).

The distribution of the endpoints of a tolerance interval is somewhat complicated.[1] Quantiles related to these distributions (the choosing of k)

[1] For those who might be interested! With a little thinking one can note that when constructing the *upper* confidence interval such a value for k must be found that

$$\mathrm{P}\Big(\frac{\overline{X} + kS - \mu}{\sigma} \geq z_\alpha\Big) = 1 - \gamma.$$

If denoting like before,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad V = \frac{(n-1)S^2}{\sigma^2},$$

then Z is standard-normally distributed and V is χ^2-distributed with $n-1$ degrees of freedom and Z and V are independent. The problem can thus be written so that no population parameters are needed: When α, γ and n are given, a number k must be found such that

$$\mathrm{P}\Big(\frac{Z}{\sqrt{n}} + \frac{k\sqrt{V}}{\sqrt{n-1}} \geq z_\alpha\Big) = 1 - \gamma.$$

are tabulated in statistics books (and in particular, in WMMY). There are also web-based calculators for these intervals. Accurate values for k are tabulated in the Appendix.

Values given on the web may however be based on crude approximate formulas and not very accurate.

Example. *A sample of $n = 9$ machine-produced metal pieces are measured and the statistics $\overline{x} = 1.0056$ cm and $s = 0.0246$ cm are obtained. Then at least 95 % of the population values are included in the tolerance interval $1.0056 \pm k0.0246$ cm (where $k = 4.5810$, see the Appendix) at the probability of 0.99. The corresponding 0.99 % confidence interval is shorter:* $(0.9781\ cm, 1.0331\ cm)$.

One-sided tolerance intervals are also possible.

2.5 Two Samples: Estimating the Difference between Two Means [9.8]

The expectations the variances of two populations are μ_1, μ_2 and σ_1^2, σ_2^2 respectively. A sample is taken from both populations, and sample sizes are n_1 and n_2. According to the Central limit theorem, the sample means obtained are $\overline{X}_1$ and $\overline{X}_2$ (as random variables) and they are nearly normally distributed. Thus also their difference $\overline{X}_1 - \overline{X}_2$ is (nearly) normally distributed, and the expectation and the variance of that population are $\mu_1 - \mu_2$ and $\sigma_1^2/n_1 + \sigma_2^2/n_2$. Furthermore, the distribution of the random variable

Naturally, the samples are independent also in this case.

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

is then nearly the standard normal distribution.

By using a quantile $z_{\alpha/2}$ of the standard normal distribution like before and by noticing that inequalities

$$-z_{\alpha/2} < \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{\alpha/2}$$

and

$$(\overline{X}_1 - \overline{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

are equivalent, the $100(1-\alpha)$ % confidence limits for the difference $\mu_1 - \mu_2$ are obtained:

$$(\overline{x}_1 - \overline{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

where $\overline{x}_1$ and $\overline{x}_2$ are the realized sample means. Here it was again assumed that the population variances σ_1^2 and σ_2^2 are known.

Because of the independence, the density function of the joint distribution of Z and V is $\phi(z)g(v)$, where g is the density function of the χ^2-distribution (with $n-1$ degrees of freedom) and ϕ is the density function of the standard normal distribution. By using that, the left side probability is obtained as an integral formula, and an equation is obtained for k. It shouldn't be a surprise that this is difficult and requires a numerical solution! In case of two-sided tolerance interval the situation is even more complicated.

Example. *The gas mileage of two different types of engines A and B was measured by driving cars having these engines,* $n_{\mathrm{A}} = 50$ *times for engine A and* $n_{\mathrm{B}} = 75$ *times for engine B. The sample means obtained are* $\overline{x}_{\mathrm{A}} = 36$ *mpg (miles per gallon) and* $\overline{x}_{\mathrm{B}} = 42$ *mpg. By using the quantile* $z_{0.02} = 2.054$ *of the standard normal distribution, for the difference* $\mu_{\mathrm{B}} - \mu_{\mathrm{A}}$ *the calculated* 96 *% confidence limits are* 3.43 *mpg and* 8.57 *mpg.* [**9.9**]

If the population variances σ_1^2 and σ_2^2 are not known, the situation becomes more complicated. Then naturally we try to use the sample variances s_1^2 and s_2^2 obtained from the sample.

A nice feature of the χ^2-distribution is that if V_1 and V_2 are independent χ^2-distributed random variables with v_1 and v_2 degrees of freedom, then their sum $V_1 + V_2$ is also χ^2-distributed with $v_1 + v_2$ degrees of freedom. By considering the sample variances to be random variables S_1^2 and S_2^2, it is known that random variables

This is quite difficult to prove. It is however somewhat apparent, if you remember that V_1 and V_2 can be presented as a sum of squares of independent standard normal distributions.

$$V_1 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \quad \text{and} \quad V_2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

have the χ^2-distributions with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, and they are also independent. Thus the random variable

$$V = V_1 + V_2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} + \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

has the χ^2-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Let's first consider a case where σ_1^2 and σ_2^2 are known to be equal $(= \sigma^2)$, although it is not known what σ^2 is. Then

$$V = \frac{1}{\sigma^2}\big((n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\big)$$

which is χ^2-distributed with $n_1 + n_2 - 2$ degrees of freedom. For more concise notation, let's denote

$$S_{\mathrm{p}}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

the *pooled sample variance.* Correspondingly, we obtain s_{p}^2 from the realized sample variances s_1^2 and s_2^2.'

Because the random variables Z (defined earlier) and V are independent, the random variable

This is also difficult to prove.

Note how the population standard deviations σ_1 and σ_2 can't be eliminated from the formula of T if they are unequal or the ratio σ_1/σ_2 is unknown.

$$T = \frac{Z}{\sqrt{V/(n_1 + n_2 - 2)}} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_{\mathrm{p}}\sqrt{1/n_1 + 1/n_2}}$$

has the t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

By using the quantile $t_{\alpha/2}$ of the t-distribution (with $n_1 + n_2 - 2$ degrees of freedom) and by noticing that the double inequalities

$$-t_{\alpha/2} < \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_{\mathrm{p}}\sqrt{1/n_1 + 1/n_2}} < t_{\alpha/2}$$

and

$$(\overline{X}_1-\overline{X}_2)-t_{\alpha/2}S_{\mathrm{p}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}<\mu_1-\mu_2<(\overline{X}_1-\overline{X}_2)+t_{\alpha/2}S_{\mathrm{p}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

are equivalent, for the difference $\mu_1-\mu_2$ we now obtain the $100(1-\alpha)$ % confidence limits

$$(\overline{x}_1-\overline{x}_2)\pm t_{\alpha/2}s_{\mathrm{p}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},$$

where $\overline{x}_1$ and $\overline{x}_2$ are the realized sample means.

Example. *A diversity index was measured in two locations monthly. The measurements lasted one year $n_1=12$ in location 1 and for ten months ($n_2=10$) in location 2. The obtained statistics were* [**9.10**]

$$\overline{x}_1=3.11\quad,\quad s_1=0.771\quad,\quad \overline{x}_2=2.04\quad \textit{and}\quad s_2=0.448.$$

The calculated pooled sample variance is $s_{\mathrm{p}}^2=0.417$, so $s_{\mathrm{p}}=0.646$. The required t-quantile (with 20 degrees of freedom) is $t_{0.05}=1.725$, by using which we obtain for the difference $\mu_1-\mu_2$ the calculated confidence interval $(0.593, 1.547)$.

If the population variances are not known nor they are known to be equal, the situation becomes difficult. It can often be however noted that if the population variances are approximately equal, the method mentioned above can be used. (The equality of variances can be tested for example by using the F-distribution, see section 3.7.) The method is often used, even when the population variances are known to differ, if the sample sizes are (approximately) equal.

This is known as the Behrens–Fisher-problem.

This however has little theoretical basis.

A widely used method when the population variances cannot be supposed to be even approximately equal, is the following *Welch–Satterthwaite-approximation:* A random variable

Bernard Welch (1911–1989), Franklin Satterthwaite

$$W=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{S_1^2/n_1+S_2^2/n_2}}$$

is nearly t-distributed with

$$v=\frac{(a_1+a_2)^2}{a_1^2/(n_1-1)+a_2^2/(n_2-1)},$$

degrees of freedom, where $a_1=s_1^2/n_1$ and $a_2=s_2^2/n_2$. This v isn't usually an integer, but that is no problem because the t-distribution is defined also in cases when its degree of freedom is not an integer. By using this information we obtain for the difference $\mu_1-\mu_2$ the *approximative* $100(1-\alpha)$ % confidence limits

When using tabulated values v must be rounded off to closest integer or interpolated.

$$(\overline{x}_1-\overline{x}_2)\pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},$$

where again $\overline{x}_1$ and $\overline{x}_2$ are the realized sample means.

The accuracy of this approximation is a controversial subject. Some people recommend that it always be used when there is even a little doubt of the equality of the population variances. Others warn about the inaccuracy of the approximation when the population variances differ greatly.

Example. *The amount of orthophosphate is measured at two different stations.* $n_1 = 15$ *measurements were made at station 1 and* $n_2 = 12$ *at station 2. Population variances are unknown. The obtained statistics were (mg/l)*

$$\overline{x}_1 = 3.84 \quad , \quad s_1 = 3.07 \quad , \quad \overline{x}_2 = 1.49 \quad \textit{and} \quad s_2 = 0.80.$$

By using the (approximative) t-quantile $t_{0.025} = 2.117$ *with* $v = 16.3$ *degrees of freedom we obtain for the difference* $\mu_1 - \mu_2$ *the (approximative)* 95 *% confidence interval* (0.60 *mg/l*, 4.10 *mg/l*).

The same interval is obtained at given precision by rounding off the degree of freedom to 16.

2.6 Paired observations [9.9]

Often two populations examined are connected element by element. For example a test subject on two different occasions, a product before and after some treatment or a product now and a year later and so on. Let's denote the expectation of the first population by μ_1 and the second by μ_2. Let's take a random sample of matched pairs from the two populations:

$$X_{1,1}, \ldots, X_{1,n} \quad \text{and} \quad X_{2,1}, \ldots, X_{2,n}.$$

Let's denote by D_i the value in population 1 minus the corresponding value in population 2:

$$D_1 = X_{1,1} - X_{2,1} \ , \ldots , \ D_n = X_{1,n} - X_{2,n}$$

and correspondingly the realized differences

$$d_1 = x_{1,1} - x_{2,1} \ , \ldots , \ d_n = x_{1,n} - x_{2,n}.$$

Now the differences are considered the actual population (either random or realized values). Thus, the sample means $\overline{D}$ and $\overline{d}$ and the sample variances S^2 ja s^2 are obtained.

Clearly, $\mathrm{E}(\overline{D}) = \mu_1 - \mu_2$. On the other hand, the counterparts $X_{1,i}$ and $X_{2,i}$ aren't generally independent or uncorrelated, so there actually isn't too much information about the variance of $\overline{D}$. In order to make statistical analysis, let's suppose that the distribution of the differences of the population values is (approximately) normal.

This isn't saying anything about the actual population distributions, they don't need to be even close to normal.

Just like before in section 2.2, we note that the random variable

$$T = \frac{\overline{D} - (\mu_1 - \mu_2)}{S/\sqrt{n}}$$

has the t-distribtution with $n - 1$ degrees of freedom. Thus, we obtain from the realized samples the $100(1 - \alpha)$ % confidence limits for the difference of the population expectations $\mu_1 - \mu_2$

$$\overline{d} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}.$$

Example. *TCDD levels in plasma (population 1) and fat tissue (population 2) were measured on $n = 20$ veterans who were exposed to Agent Orange -toxin during the Vietnam war. The mean of the differences of the sample values was $\overline{d} = -0.87$ and the standard deviation was $s = 2.98$. The t-quantile with 19 degrees of freedom is $t_{0.025} = 2.093$ and thus, we obtain for the difference $\mu_1 - \mu_2$ the 95 % confidence interval $(-2.265, 0.525)$.* [9.12]

2.7 Estimating a Proportion [9.10]

When estimating a proportion, the information we obtain is if the sample values are of a certain type ('success') or not ('failure'). The number of successes is denoted by X (a random variable) or by x (a realized numerical value). If the sample size is n and the probability of a successful case in the population is p (ratio), the distribution of X is a binomial distribution $\mathrm{Bin}(n, p)$ and

$$\mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

For this distribution it is known that

$$\mathrm{E}(X) = np \quad \text{and} \quad \mathrm{var}(X) = np(1-p).$$

Because $p(1-p) \leq 1/4$, it follows that $\mathrm{var}(X) \leq n/4$. The natural point estimator and estimate of the ratio p are

The maximum of the function $x(1-x)$ is $1/4$.

$$\hat{P} = \frac{X}{n} \quad \text{and} \quad \hat{p} = \frac{x}{n}.$$

$\hat{P}$ is unbiased, in other words $\mathrm{E}(\hat{P}) = p$, and

$$\mathrm{var}(\hat{P}) = \frac{1}{n^2}\mathrm{var}(X) = \frac{p(1-p)}{n} \leq \frac{1}{4n}.$$

Again the variance of the estimator decreases as n increases. We also note that if the standard deviation of $\hat{P}$ is wanted to be at most b, it is enough to choose n such that $n \geq \dfrac{1}{4b^2}$.

If the realized number of successful elements is x, then in interval estimation we obtain the lower limit of the $100(1-\alpha)$ % confidence interval for p by requiring that

By considering how the probability on the left changes as p decreases, you see that it indeed is the lower limit

$$\mathrm{P}(X \geq x) = \frac{\alpha}{2}.$$

Thus, we obtain an equation for $\hat{p}_{\mathrm{L}}$

$$\sum_{i=x}^{n} \binom{n}{i} \hat{p}_{\mathrm{L}}^i (1-\hat{p}_{\mathrm{L}})^{n-i} = \frac{\alpha}{2}.$$

Correspondingly, the upper confidence limit $\hat{p}_{\mathrm{U}}$ for two-sided interval is obtained by requiring that

$$\mathrm{P}(X \leq x) = \frac{\alpha}{2},$$

and it's obtained by solving the equation

This accurate interval estimate is called the *Clopper–Pearson estimate.*

$$\sum_{i=0}^{x} \binom{n}{i} \hat{p}_{\mathrm{U}}^{i} (1-\hat{p}_{\mathrm{U}})^{n-i} = \frac{\alpha}{2}.$$

These two equations are difficult to solve numerically, especially if n is large. The solution is implemented in MATLAB, and there are also web-based calculators.

A special function, the beta function, is often used in the solution.

One-sided confidence intervals are obtained similarly, just replace $\alpha/2$ on the right hand side by α.

Instead of the above exact interval estimate, one of the many approximate methods can be used to compute the interval estimate. According to the Central limit theorem, the random variable X has nearly a normal distribution $\mathrm{N}\big(np, np(1-p)\big)$. Thus, the random variable

$$Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

has nearly the standard normal distribution. When the realized estimate $\hat{p} = x/n$ is obtained for p, the approximative $100(1-\alpha)$ % confidence limits are then obtained by solving the second order equation:

This estimate is called the *Wilson estimate.*

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \pm z_{\alpha/2} \quad \text{or} \quad (\hat{p} - p)^2 = \frac{z_{\alpha/2}^2}{n} p(1-p).$$

The estimate $\hat{p}$ can be used also in the denominator, because the random variable

$$Z' = \frac{\hat{P} - p}{\sqrt{\hat{P}(1-\hat{P})/n}}$$

is also nearly normally distributed. With this, the approximative confidence intervals can then be calculated very similarly as before when considering a normally distributed population. The result isn't however always too accurate, and nowadays exact methods are preferable.

The *Wald estimate.*

There are many other approximative interval estimates for binomial distribution, that differ in their behavior. The above mentioned exact estimate is the most conservative but also the most reliable.

Example. $n = 500$ *households were chosen at random and asked if they subscribe to a certain cable TV channel.* $x = 340$ *had ordered the TV channel in question. Then* $\hat{p} = 340/500 = 0.680$ *and the* 95 *% confidence interval for the ratio* p *is* $(0.637, 0.721)$. [**9.13**]

Here n is large and the correct p is in the "middle", so the normal distribution approximation works fine.

2.8 Single Sample: Estimating the Variance [**9.12**]

A natural point estimator for the population variance σ^2 is the sample variance S^2; the corresponding point estimate would be the realized sample variance s^2. As noted, S^2 is unbiased, that is $\mathrm{E}(S^2) = \sigma^2$, no matter what the population distribution is (as long as it has a variance!)

For the interval estimation it has to be assumed that the population distribution is normal (accurately enough). The χ^2-distribution to be used is namely quite vulnerable to abnormality. The random variable

$$V = \frac{(n-1)S^2}{\sigma^2}$$

has then the χ^2-distribution with $n-1$ degrees of freedom. Let's now choose quantiles $h_{1,\alpha/2}$ and $h_{2,\alpha/2}$ of the χ^2-distribution in question so that

Because the χ^2-distribution is not symmetric, these quantiles aren't connected.

$$\mathrm{P}(V \le h_{1,\alpha/2}) = \mathrm{P}(V \ge h_{2,\alpha/2}) = \frac{\alpha}{2}.$$

Then

$$\mathrm{P}(h_{1,\alpha/2} < V < h_{2,\alpha/2}) = 1 - \alpha.$$

The double inequalities

$$h_{1,\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < h_{2,\alpha/2}$$

and

$$\frac{(n-1)S^2}{h_{2,\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{h_{1,\alpha/2}}$$

are equivalent. Thus, from the realized sample variance s^2, confidence limits are obtained for σ^2

$$\frac{(n-1)s^2}{h_{2,\alpha/2}} \quad \text{and} \quad \frac{(n-1)s^2}{h_{1,\alpha/2}}.$$

One-sided confidence limits are obtained similarly just by using another χ^2-quantile, $h_{1,\alpha}$ for the upper and $h_{2,\alpha}$ lower confidence limit.

Example. *$n = 10$ packages of grass seed were weighted. The weights are supposed to be normally distributed. The obtained sample variance is $s^2 = 28.62\ g^2$. By using the χ^2-quantiles $h_{1,0.025} = 2.700$ and $h_{2,0.025} = 19.023$ (with 9 degrees of freedom), the calculated 95 % confidence interval for the population variance σ^2 is $(13.54\ g^2, 95.40\ g^2)$.*

The square roots of the confidence limits for variance σ^2 are the confidence limits for the population standard deviation σ.

These limits are exact, contrary to what is claimed in WMMY

2.9 Two Samples: Estimating the Ratio of Two Variances [9.13]

If two samples (sample sizes n_1 and n_2, sample variances S_1^2 and S_2^2) are taken from two populations whose variances are σ_1^2 ja σ_2^2, then the obvious point estimator for the ratio σ_1^2/σ_2^2 is the ratio S_1^2/S_2^2. Corresponding point estimate is s_1^2/s_2^2, the ratio of the realized sample variances s_1^2 and s_2^2.

Independent samples, of course!

This isn't usually unbiased. For example, when considering normally distributed populations, the corresponding unbiased estimator is

$$\frac{n_2 - 3}{n_2 - 1}\frac{S_1^2}{S_2^2}$$

(supposing that $n_2 > 3$).

For interval estimation, it has to be supposed that the populations are normally distributed. The F-distribution isn't robust in this respect, and

using it with non-normal populations leads easily to inaccurate results. The random variable

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2}$$

is F-distributed with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Let's choose, for the interval estimation, quantiles $f_{1,\alpha/2}$ and $f_{2,\alpha/2}$ of the F-distribution in question such that

$$\mathrm{P}(F \le f_{1,\alpha/2}) = \mathrm{P}(F \ge f_{2,\alpha/2}) = \frac{\alpha}{2}.$$

Then

$$\mathrm{P}(f_{1,\alpha/2} < F < f_{2,\alpha/2}) = 1 - \alpha.$$

Like the χ^2-distribution, the F-distribution is asymmetric, so the quantiles $f_{1,\alpha/2}$ and $f_{2,\alpha/2}$ are not directly connected. They aren't however completely unrelated either. We remember that the random variable $F' = 1/F$ is F-distributed with $n_2 - 1$ ja $n_1 - 1$ degrees of freedom. If quantiles $f'_{1,\alpha/2}$ and $f'_{2,\alpha/2}$ are obtained for the F-distribution in question, then $f'_{1,\alpha/2} = 1/f_{2,\alpha/2}$ and $f'_{2,\alpha/2} = 1/f_{1,\alpha/2}$. In particular, if the sample sizes are equal, in other words $n_1 = n_2$, then the distributions of F and F' are the same and $f_{1,\alpha/2} = 1/f_{2,\alpha/2}$.

This is exploited in tables: The values are tabulated often only for the end tail quantiles $f_{2,\alpha/2}$ or the first degree of freedom is smaller.

Because the inequalities

$$f_{1,\alpha/2} < \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} < f_{2,\alpha/2}$$

and

$$\frac{S_1^2}{S_2^2}\frac{1}{f_{2,\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{f_{1,\alpha/2}}$$

are equivalent, from the realized sample variances s_1^2 and s_2^2 we can calculate the $100(1-\alpha)$ % confidence limits for the ratio σ_1^2/σ_2^2

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{2,\alpha/2}} \quad \text{and} \quad \frac{s_1^2}{s_2^2}\frac{1}{f_{1,\alpha/2}}.$$

The one-sided confidence limits are obtained similarly, but by using only one F-quantile, quantile $f_{1,\alpha}$ for upper and $f_{2,\alpha}$ for lower confidence limit. Furthermore, the square roots of the confidence limits of the ratio σ_1^2/σ_2^2 (population variances) are the confidence limits for the ratio σ_1/σ_2 (population standard deviations).

These limits are exact, contrary to what is claimed in WMMY.

Example. *Let's return to the orthophosphate measurements of an example in section 2.5. The sample sizes were* $n_1 = 15$ *and* $n_2 = 12$*, the obtained population standard deviations were* $s_1 = 3.07$ *mg/l and* $s_2 = 0.80$ *mg/l. By using the F-quantiles* $f_{1,0.01} = 0.2588$ *and* $f_{2,0.01} = 4.2932$ *(with* 14 *and* 11 *degrees of freedom), the calculated* 98 *% confidence interval for the ratio* σ_1^2/σ_2^2 *is* $(3.430, 56.903)$*. Because the number* 1 *is not included in this interval, it seems to be correct to assume–as was done in the example–that the population variances aren't equal. The* 98 *% confidence limits for the ratio* σ_1/σ_2 *are the (positive) square roots of the previous limits* $(1.852, 7.543)$*.* [**9.18**]

Chapter 3

TESTS OF HYPOTHESES

3.1 Statistical Hypotheses [10.1]

A *statistical hypotheses* means some attribute of the population distribution(s) that it (they) either has (have) or does (do) not have. Such an attribute often involves the parameters of the population distributions, distribution-related probabilities or something like that. By *hypothesis testing* we try to find out, by using the sample(s), whether the population distribution(s) has (have) the attribute in question or not. The testing is based on random samples, so the result ("yes" or "no") is not definite, but it can be considered a random variable. The probability of an incorrect result should of course be small and quantizable.

Traditionally a *null hypothesis* (denoted by H_0) and an *alternative hypothesis* (denoted by H_1) are presented. A test is made with an assumption that the null hypothesis is true. The result of the test may then prove that the assumption is probably wrong, in other words the realized result is very improbable if H_0 is true. The result of hypothesis testing is either of the following:

- Strong enough evidence has been found to reject the null hypothesis H_0. We'll continue by assuming that the alternate hypothesis H_1 is true. This may require further testing.
- The sample and the test method used haven't given strong enough evidence to reject H_0. This may result because H_0 is true or because the test method wasn't strong enough. We'll continue by assuming that H_0 is true.

Because of random sampling, both of the results may be wrong, ideally though only with a small probability.

3.2 Hypothesis Testing [10.2]

A hypothesis is tested by calculating some suitable statistic from the sample. If this produces a value that is highly improbable when assuming that the null hypothesis H_0 is true, evidence has been found to reject H_0. The result of hypothesis testing may be wrong in two different ways:

Type I error: H_0 is rejected, although it's true ("false alarm").

Type II error: H_0 isn't rejected, although it's false.

The actual attributes of the population distribution(s) and the error types divide the results to four cases:

	H_0 is true	H_0 is false
H_0 isn't rejected	The right decision	Type II error
H_0 is rejected	Type I error	The right decision

The probability of type I error is called the *risk* or the *level of significance* of the test and it is often denoted by α. The greatest allowed level of significance α is often a starting point of hypothesis testing.

The probability of type II error can't often be calculated, for H_0 may be false in many ways. Often some sort of an (over) estimate is calculated by assuming a typical relatively insignificant way for H_0 to break down. This probability is usually denoted by β. The value $1 - \beta$ is called the *power* of the test. The more powerful a test is, the smaller deviation it notices from H_0.

Example. *Let's consider a normally-distributed population, whose expectation is supposed to be μ_0 (hypothesis H_0). The population variance σ^2 is considered to be known. If the realized sample mean $\overline{x}$ is a value that is in the tail area of the $N(\mu_0, \sigma^2/n)$-distribution and outside a wide interval $(\mu_0 - z, \mu_0 + z)$, there is a reason to reject H_0. Then α is obtained by calculating the total tail probability for the $N(\mu_0, \sigma^2/n)$-distribution. By increasing the sample size n the probability α can be made as small as wanted.*

The distribution of $\overline{X}$ narrows and the tail probabilities decrease.

The value for probability β cannot be calculated, for if the population expectation isn't μ_0, it can be almost anything. The larger the deviation between the population expectation and μ_0, the smaller the actual β is. If we however consider a deviation of size d to be good enough reason to reject H_0, with of course $|d| > z$, we could estimate β by calculating the probability of the $N(\mu_0 + d, \sigma^2/n)$ distribution between the values $\mu_0 \pm z$. This probability also decreases as the sample size n increases, for the distribution of $\overline{X}$ concentrates around an expected value that doesn't belong to the interval $(\mu_0 - z, \mu_0 + z)$, and the probability of the interval in question decreases.

By increasing the sample size we can usually make both α and (estimated) β decrease as small as wanted. The sensitivity of the test shouldn't though be always increased this way. If for example the population values are given to just a few decimals, then the sensitivity (the sample size) shouldn't be increased so much that it observes differences smaller than the data accuracy. Then the test would reject the null hypothesis very often and become useless!

3.3 One- and Two-Tailed Tests [10.3]

Often a hypothesis concerns some population parameter θ. Because the parameter is numerical, there are three different types of basic hypotheses

concerning it: two one-tailed tests and one two-tailed test. The same can be said for comparing corresponding parameters of two populations. The testing of hypotheses like this at a risk level α comes back to constructing the $100(1-\alpha)$ % confidence intervals for θ. The basic idea is to try to find a confidence interval that lies in an area where H_0 should be rejected. If this is not possible, there is no reason to reject H_0 at the risk level used, in other words the risk to make the wrong decision is too large.

The *one-tailed* hypothesis pairs are

$$\mathrm{H}_0 : \theta = \theta_0 \quad \text{vs.} \quad \mathrm{H}_1 : \theta > \theta_0$$

and

$$\mathrm{H}_0 : \theta = \theta_0 \quad \text{vs.} \quad \mathrm{H}_1 : \theta < \theta_0,$$

where the reference value θ_0 is given.

The pair $\mathrm{H}_0 : \theta = \theta_0$ vs. $\mathrm{H}_1 : \theta > \theta_0$ is tested at the level of significance α by calculating from the realized sample the *lower* $100(1-\alpha)$ % confidence limit $\hat{\theta}_\mathrm{L}$ for parameter θ in the manner presented earlier. The null hypothesis H_0 is rejected if the reference value θ_0 isn't included in the obtained confidence interval, in other words if $\theta_0 \leq \hat{\theta}_\mathrm{L}$.

Correspondingly, the pair $\mathrm{H}_0 : \theta = \theta_0$ vs. $\mathrm{H}_1 : \theta > \theta_0$ is tested at the level of significance α by calculating from the realized sample the *upper* $100(1-\alpha)$ % confidence limit $\hat{\theta}_\mathrm{U}$ for parameter θ in ways presented earlier. The null hypothesis H_0 is rejected if the reference value θ_0 isn't included in the obtained confidence interval, in other words $\theta_0 \geq \hat{\theta}_\mathrm{U}$.

All the parameter values aren't included in one-tailed tests. In above for example while testing the hypothesis pair $\mathrm{H}_0 : \theta = \theta_0$ vs. $\mathrm{H}_1 : \theta > \theta_0$ it was assumed that the correct value of the parameter θ cannot be less than θ_0. What if it however is? Then in a way a type II error cannot occur: H_0 is certainly false, but H_1 isn't true either. On the other hand, the lower confidence limit $\hat{\theta}_\mathrm{L}$ decreases and the probability of type I error α decreases. The case is similar if while testing the hypothesis pair $\mathrm{H}_0 : \theta = \theta_0$ vs. $\mathrm{H}_1 : \theta < \theta_0$ the correct value of the parameter θ is greater than θ_0.

In terms of testing the situation just gets better!

Example. *The average life span of $n = 100$ deceased persons was $\overline{x} = 71.8$ y. According to earlier studies, the population standard deviation is assumed to be $\sigma = 8.9$ y. According to this information, could it be concluded that the average life span μ of the population is greater than 70 y? The life span is supposed to be normally distributed. The hypothesis pair to be tested is* [**10.3**]

$$\mathrm{H}_0 : \mu = 70 \ y \quad \textit{vs.} \quad \mathrm{H}_1 : \mu > 70 \ y.$$

The risk of the test is supposed to be $\alpha = 0.05$, when $z_\alpha = 1.645$. Let's calculate the lower 95 % confidence limit for μ

$$\hat{\mu}_\mathrm{L} = \overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}} = 70.34 \ y.$$

The actual life span is thus, with a probability of at least 95 %, greater than 70.34 y and H_0 has to be rejected.

The hypothesis pair of a *two-tailed* test is

$$\mathrm{H}_0 : \theta = \theta_0 \quad \text{vs.} \quad \mathrm{H}_1 : \theta \neq \theta_0.$$

In order to test this at the level of significance α let's first calculate the *two-tailed* $100(1-\alpha)$ confidence interval $(\hat{\theta}_\mathrm{L}, \hat{\theta}_\mathrm{U})$ for the parameter θ. Now H_0 is rejected if the reference value θ_0 isn't included in the interval.

Example. *A manufacturer of fishing equipment has developed a new synthetic fishing line that he claims has a breaking strength of* 8.0 *kg while the standard deviation is* $\sigma = 0.5$ *kg. The deviation is supposed to be accurate. In order to test the claim, a sample of* 50 *fishing lines was taken and the mean breaking strength was found to be* $\overline{x} = 7.8$ *kg. The risk of the test was supposed* $\alpha = 0.01$*. Here the test is concerned with the two-tailed hypothesis pair* $\mathrm{H}_0 : \mu = 8.0$ *vs.* $\mathrm{H}_1 : \mu \neq 8.0$*. Now the* $100(1-\alpha) = 99$ *% confidence interval for the population expectation* μ *is* $(7.62$ *kg*$, 7.98$ *kg*$)$*, and the value* 8.0 *kg isn't included in this interval. Thus* H_0 *is rejected with the risk* 0.01*.* [10.4]

3.4 Test statistics [10.4]

If a hypothesis concerns a population distribution parameter θ, the hypothesis testing can be done using the confidence interval for θ. On the other hand, the testing doesn't require the confidence interval itself. The task is only to verify if the value $\theta = \theta_0$ given by the null hypothesis is included in the confidence interval or not, and this can be usually done without constructing the empirical confidence interval, by using a *test statistic*. This is the only way to test hypotheses that don't concern parameters.

In above, the confidence intervals were constructed by using a random variable, whose (approximative) distribution doesn't depend on the parameter studied: Z (standard normal distribution), T (t-distribution), V (χ^2-distribution), X (binomial distribution) and F (F-distribution). The confidence interval was obtained by presenting the suitable quantile(s) of the distribution and by changing the (double) inequality concerning it (them) to concern the parameter. Thus, if a confidence interval is used to test a hypothesis, it can be also done straightforward by using the inequality concerning the "original" random variable. The *test statistic* is then that particular formula that connects the random variable to sample random variables presented for realized values. The area where the value of the test statistic leads to rejecting the null hypothesis is the *critical region.*

Example. *Let's return to the previous example concerning average life spans. The confidence interval was constructed by using the standard normally distributed random variable* [10.3]

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

The value that agrees with the null hypothesis $\mu = \mu_0$ is included in the confidence interval used when

$$\mu_0 > \overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}},$$

or when the realized value of Z in accordance with H_0

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

is smaller that the quantile z_α. Thus, H_0 *is rejected if $z \geq z_\alpha$. Here z is the test statistic and the critical region is the interval $[z_\alpha, \infty)$. In the example, the realized value of Z is $z = 2.022$ and it is greater than $z_{0.05} = 1.645$.*

Example. *In the example concerning synthetic fishing lines above the realized value of Z is $z = -2.83$ and it is less than $-z_{0.005} = -2.575$. The critical region consists of the intervals $(-\infty, -2.575]$ and $[2.575, \infty)$.* **[10.4]**

All the hypotheses testing based on confidence intervals in previous chapter can in this way be returned to using a suitable test statistic. The critical area consists of one or two tail areas restricted by suitable quantiles.

In certain cases the use of test statistics is somewhat easier that the use of confidence intervals. This is the case for example when testing hypotheses concerning ratios by using binomial distribution. If for example we'd like to test the hypothesis pair $\mathrm{H}_0 : p = p_0$ vs. $\mathrm{H}_1 : p > p_0$ at the risk α, this could be done by finding the lower confidence limit for p by solving $\hat{p}_\mathrm{L}$ from the equation

$$\sum_{i=x}^{n} \binom{n}{i} \hat{p}_\mathrm{L}^i (1 - \hat{p}_\mathrm{L})^{n-i} = \alpha.$$

Like it was noted earlier, this can be numerically challenging. Here the test variable can be chosen to be x itself and then it can be checked whether the tail probability is

$$\mathrm{P}(X \geq x) = \sum_{i=x}^{n} \binom{n}{i} p_0^i (1 - p_0)^{n-i} \leq \alpha$$

(in which case H_0 is rejected) or not. Testing can be somewhat difficult, but it is nevertheless easier than calculating the lower confidence limit $\hat{p}_\mathrm{L}$. The critical region consists of the values $x_1, \ldots, n$, where

If n is large, the binomial coefficients can be very large and the powers of p_0 very small.

$$\sum_{i=x_1}^{n} \binom{n}{i} p_0^i (1 - p_0)^{n-i} \leq \alpha \quad \text{and} \quad \sum_{i=x_1-1}^{n} \binom{n}{i} p_0^i (1 - p_0)^{n-i} > \alpha.$$

Example. *A certain vaccine is known to be efficient only in 25 % of the cases after two years. A more expensive vaccine is claimed to be more effective. In order to test the claim, $n = 100$ subjects were vaccinated with the more expensive vaccine and followed for two years. The hypothesis*

In reality, way larger sample sizes are required in medical exams.

pair tested is $\mathrm{H}_0 : p = p_0 = 0.25$ *vs.* $\mathrm{H}_1 : p > 0.25$. *The risk is wanted to be at most* $\alpha = 0.01$. *By trial-and-error (web-calculators) or by calculating with MATLAB we find that now* $x_1 = 36$. *If the more expensive vaccine provides immunity after two years in at least* 36 *cases, it can be decided that* H_0 *is rejected and find the more expensive vaccine better than the cheaper vaccine. The calculations on MATLAB are:*

```
>> p_0=0.25;
   n=100;
   alfa=0.01;

>> binoinv(1-alfa,n,p_0)+1

ans =
    36
```

In a similar way we can test the hypothesis pair $\mathrm{H}_0 : p = p_0$ vs. $\mathrm{H}_1 : p < p_0$. The critical region consists of the values $0, \ldots, x_1$, where

$$\sum_{i=0}^{x_1} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \alpha \quad \text{and} \quad \sum_{i=0}^{x_1+1} \binom{n}{i} p_0^i (1-p_0)^{n-i} > \alpha.$$

In a two-tailed test the hypothesis pair is $\mathrm{H}_0 : p = p_0$ vs. $\mathrm{H}_1 : p \neq p_0$ and the critical area consists of the values $0, \ldots, x_1$ and $x_2, \ldots, n$, where

$$\sum_{i=0}^{x_1} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \frac{\alpha}{2} \quad \text{and} \quad \sum_{i=0}^{x_1+1} \binom{n}{i} p_0^i (1-p_0)^{n-i} > \frac{\alpha}{2}$$

and

$$\sum_{i=x_2}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \frac{\alpha}{2} \quad \text{and} \quad \sum_{i=x_2-1}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} > \frac{\alpha}{2}.$$

3.5 P-probabilities [10.4]

Many statistical analysts prefer to announce the result of a test with a *P-probability.* The P-probability of a hypothesis test is the smallest risk at which H_0 can be rejected based on the sample. In practice, the P-probability of a one-tailed test is obtained by calculating the tail probability corresponding the realized statistic (assuming that H_0 is true).

Example. *If in the vaccine example mentioned above the realized number of uninfected is* $x = 38$, *the P-probability is the tail probability*

$$P = \sum_{i=38}^{100} \binom{100}{i} 0.25^i (1-0.25)^{100-i} = 0.0027.$$

Calculating with MATLAB this is obtained as follows:

```
>> p_0=0.25;
  n=100;
  x=38;

>> 1-binocdf(x-1,n,p_0)

ans =
    0.0027
```

In two-tailed testing the P-value is obtained by choosing the smaller of the two tail probabilities corresponding the realized test statistic, and by multiplying the result by two. For example in a two-sided test concerning ratios the P-probability is the smaller of the values

Usually it is completely clear which number is smaller.

$$\sum_{i=0}^{x} \binom{n}{i} p_0^i (1-p_0)^{n-i} \quad \text{and} \quad \sum_{i=x}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}$$

multiplied by two.

Example. *In the example concerning synthetic fishing lines above the realized value of the test statistic was* $z = -2.83$*. The corresponding (clearly) smaller tail probability is* 0.0023 *(left tail). Thus, the P-probability is* $P = 0.0046$*.* **[10.4]**

The P-probability is a random variable (if we consider a sample to be random) and varies when the test is repeated using different samples. Ideally, when using the P-probability, a wanted smallest risk α is chosen beforehand and H_0 is rejected if the (realized) P-probability is $\leq \alpha$. In many cases however, no risk α is set beforehand, but the realized value of the P-probability is calculated and the conclusions are made according to it. Because at least sometimes the realized P-probability is quite small, the obtained insight of the risk of the test may be completely wrong in these cases. For this reason (and others) not every statistician favors the use of the P-probability.

3.6 Tests Concerning Expectations [10.5–8]

Earlier the testing of the population expectation μ has been presented when its variance σ^2 is known. According to the Central limit theorem a test statistic can be formulated based on the (approximative) standard normal distribution, namely the statistic

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}.$$

The different test situations are the following, when the null hypothesis is $H_0 : \mu = \mu_0$ and the wanted risk is α:

H_1	Critical region	P-probability
$\mu > \mu_0$	$z \geq z_\alpha$	$1 - \Phi(z)$
$\mu < \mu_0$	$z \leq -z_\alpha$	$\Phi(z)$
$\mu \neq \mu_0$	$\lvert z\rvert \geq z_{\alpha/2}$	$2\min\big(\Phi(z), 1-\Phi(z)\big)$

Here Φ is the cumulative distribution function of the standard normal distribution.

Let's then consider a case where the population distribution is normal (at least approximatively) and the population variance σ^2 is unknown. The testing of the expectation μ can be done by using the t-distribution with $n-1$ degrees of freedom, and we obtain the test statistic

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

from the realized statistics. Like before, the different test situations are the following for the null hypothesis $\mathrm{H}_0 : \mu = \mu_0$ and the risk α:

H_1	Critical region	P-probability
$\mu > \mu_0$	$t \geq t_\alpha$	$1 - F(t)$
$\mu < \mu_0$	$t \leq -t_\alpha$	$F(t)$
$\mu \neq \mu_0$	$\lvert t\rvert \geq t_{\alpha/2}$	$2\min\big(F(t), 1 - F(t)\big)$

Here F is the cumulative distribution function of the t-distribution with $n - 1$ degrees of freedom.

These tests are used often even when there is no accurate information about the normality of the population distribution, as long as it is unimodal and nearly symmetric. The result of course isn't always very accurate.

The t-distribution is namely quite robust in that respect.

Example. *In $n = 12$ households, the annual energy consumption of a vacuum cleaner was measured. The average value was $\overline{x} = 42.0$ kWh and the sample standard deviation $s = 11.9$ kWh. The distribution is assumed to be closely enough normal. Could it, according to this information, be assumed that the expected annual consumption is less than $\mu_0 = 46$ kWh? The hypothesis pair to be tested is $\mathrm{H}_0 : \mu = \mu_0 = 46$ kWh vs. $\mathrm{H}_1 : \mu < 46$ kWh, and the risk of the test may be at most $\alpha = 0.05$. The realized value of the test statistic is now $t = -1.16$, and on the other hand, $-t_{0.05} = -1.796$ (with 11 degrees of freedom). Thus, H_0 won't be rejected, and the annual consumption cannot be considered to be less than 46 kWh. Even the P-probability is $P = 0.135$.* [**10.5**]

When comparing the expectations μ_1 ja μ_2 of two different populations, when their variances σ_1^2 ja σ_2^2 are known, we end up, according to the Central limit theorem, with the (approximative) standard normal distribution and the test statistic

$$z = \frac{\overline{x}_1 - \overline{x}_2 - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}},$$

where $\overline{x}_1$ and $\overline{x}_2$ are the realized sample means, n_1 and n_2 are the sample sizes and d_0 is the difference of the population expectations given by the null hypothesis.

For the null hypothesis $\mathrm{H}_0 : \mu_1 - \mu_2 = d_0$ and the risk α, the tests are the following:

H_1	Critical region	P-probability
$\mu_1 - \mu_2 > d_0$	$z \geq z_\alpha$	$1 - \Phi(z)$
$\mu_1 - \mu_2 < d_0$	$z \leq -z_\alpha$	$\Phi(z)$
$\mu_1 - \mu_2 \neq d_0$	$\lvert z\rvert \geq z_{\alpha/2}$	$2\min\big(\Phi(z), 1 - \Phi(z)\big)$

If, while comparing population expectations μ_1 ja μ_2, the population variances are unknown, but they are known to be equal, we may continue by assuming that the populations are normally distributed (at least quite

accurately) and the test statistic is obtained by using the t-distribution (with $n_1 + n_2 - 2$ degrees of freedom)

$$t = \frac{\overline{x}_1 - \overline{x}_2 - d_0}{s_{\mathrm{p}}\sqrt{1/n_1 + 1/n_2}},$$

where

$$s_{\mathrm{p}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

(pooled sample variance) and s_1^2, s_2^2 are the realized sample variances. Then, for the null hypothesis $\mathrm{H}_0 : \mu_1 - \mu_2 = d_0$ and the risk α, the tests are the following:

H_1	Critical region	P-probability
$\mu_1 - \mu_2 > d_0$	$t \geq t_\alpha$	$1 - F(t)$
$\mu_1 - \mu_2 < d_0$	$t \leq -t_\alpha$	$F(t)$
$\mu_1 - \mu_2 \neq d_0$	$\lvert t\rvert \geq t_{\alpha/2}$	$2\min\big(F(t), 1 - F(t)\big)$

Here again, F is the cumulative distribution function of the t-distribution, now with $n_1 + n_2 - 2$ degrees of freedom.

Example. *The abrasive wears of two different laminated materials were compared. The average wear of material 1 was obtained in* $n_1 = 12$ *tests to be* $\overline{x}_1 = 85$ *(on some suitable scale) while the sample standard deviation was* $s_1 = 4$*. The average wear of material 2 was obtained in* $n_2 = 10$ *tests to be* $\overline{x}_2 = 81$ *and the sample standard deviation was* $s_2 = 5$*. The distributions are assumed to be close to normal with equal variances. Could we, at the risk* $\alpha = 0.05$*, conclude that the wear of material 1 exceeds that of material 2 by more than* $d_0 = 2$ *units?* [**10.6**]

The hypothesis pair to be tested is $\mathrm{H}_0 : \mu_1 - \mu_2 = d_0 = 2$ *vs.* $\mathrm{H}_1 : \mu_1 - \mu_2 > 2$*. By calculating from the realized statistics we obtain the pooled standard deviation* $s_{\mathrm{p}} = 4.48$ *and the test statistic* $t = 1.04$*. The P-probability calculated from those is* $P = 0.155$ *(t-distribution with* 20 *degrees of freedom). This is clearly greater than the greatest allowed risk* $\alpha = 0.05$*, so, according to these samples,* H_0 *cannot be rejected, and we cannot claim that the average wear of material 1 exceeds that of material 2 by more than* 2 *units.*

If the population variances cannot be considered to be equal, the testing proceeds similarly but by using the Welch–Satterthwaite-approximation. The test statistic is then

$$t = \frac{\overline{x}_1 - \overline{x}_2 - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

and the (approximative) t-distribution is used with

$$v = \frac{(a_1 + a_2)^2}{a_1^2/(n_1 - 1) + a_2^2/(n_2 - 1)}$$

degrees of freedom, where $a_1 = s_1^2/n_1$ and $a_2 = s_2^2/n_2$. Like the corre- The Behrens–Fisher-problem again!

sponding confidence interval, the usability and utility value of this test are a controversial subject.

When considering paired observations the test statistic is

See section 2.6.

$$t = \frac{\overline{d} - d_0}{s/\sqrt{n}}.$$

The tests are exactly the same as before when considering one sample by using the t-distribution (with $n-1$ degrees of freedom).

3.7 Tests Concerning Variances [10.13]

If a population is normally distributed, its variance σ^2 can be tested. The null hypothesis is then $\mathrm{H}_0 : \sigma^2 = \sigma_0^2$, and the test statistic is

$$v = \frac{(n-1)s^2}{\sigma_0^2},$$

and by using the χ^2-distribution (with $n-1$ degrees of freedom), at the risk α we obtain the tests

H_1	Critical region	P-probability
$\sigma^2 > \sigma_0^2$	$v \geq h_{2,\alpha}$	$1 - F(v)$
$\sigma^2 < \sigma_0^2$	$v \leq h_{1,\alpha}$	$F(v)$
$\sigma^2 \neq \sigma_0^2$	$v \leq h_{1,\alpha/2}$ tai $v \geq h_{2,\alpha/2}$	$2\min\big(F(v), 1 - F(v)\big)$

where F is the cumulative distribution function of the χ^2-distribution with $n-1$ degrees of freedom. This test is quite a sensitive to exceptions from the normality of the population distribution. If the population distribution isn't close enough to normal, H_0 will often be rejected in vain.

Unlike the t-distribution, χ^2-distribution isn't robust to deviation from normality

Example. *A manufacturer of batteries claims that the life of his batteries is approximatively normally distributed with a standard deviation of* $\sigma_0 = 0.9$ *y. A sample of* $n = 10$ *of these batteries has a standard deviation of* 1.2 *y. Could we conclude that the standard deviation is greater than the claimed* 0.9 *y? The risk is assumed to be* $\alpha = 0.05$*. The hypothesis pair to be tested is* $\mathrm{H}_0 : \sigma^2 = \sigma_0^2 = 0.9^2 = 0.81$ *vs.* $\mathrm{H}_1 : \sigma^2 > 0.81$*. The realized value for the test statistic is* $v = 16.0$*. The corresponding P-probability is obtained from the right side tail probability of the* χ^2*-distribution (with* 9 *degrees of freedom), and it is* $P = 0.067$*. Thus,* H_0 *isn't rejected.* [10.13]

The P-probability is however quite close to α, so some doubts may still remain about the matter.

Let there be two normally distributed populations with variances σ_1^2 and σ_2^2. The ratio σ_1^2/σ_2^2 can be similarly tested by using the F-distribution. The null hypothesis is of the form $\mathrm{H}_0 : \sigma_1^2 = k\sigma_2^2$, where k is a given value (ratio). The test statistic is

Often $k = 1$, when the equality of the population variances is being tested.

$$f = \frac{1}{k}\,\frac{s_1^2}{s_2^2}.$$

By using the F-distribution with $n_1 - 1$ and $n_2 - 1$ we obtain, at the risk α, the tests

H_1	Critical region	P-probability
$\sigma_1^2 > k\sigma_2^2$	$f \geq f_{2,\alpha}$	$1 - G(f)$
$\sigma_1^2 < k\sigma_2^2$	$f \leq f_{1,\alpha}$	$G(f)$
$\sigma_1^2 \neq k\sigma_2^2$	$f \leq f_{1,\alpha/2}$ tai $f \geq f_{2,\alpha/2}$	$2\min\big(G(f), 1 - G(f)\big)$

where G is the cumulative distribution function of the F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Like the χ^2-distribution, F-distribution is not robust to exceptions from normality, so the normality of the population distributions has to be clear. There are also more robust tests to compare variances, and these are available in statistical software.

Example. *Let's return to the example above concerning the abrasive wear of the two materials. The sample standard deviations that were obtained are* $s_1 = 4$ *and* $s_2 = 5$*. The sample sizes were* $n_1 = n_2 = 10$*. Could we assume that the variances are equal, as we did? The hypothesis pair to be tested is thus* $H_0 : \sigma_1^2 = \sigma_2^2$ *vs.* $H_1 : \sigma_1^2 \neq \sigma_2^2$ *(and so* $k = 1$*). The risk is supposed to be only* $\alpha = 0.10$*. Now* $f_{1,0.05} = 0.3146$ *and* $f_{2,0.05} = 3.1789$ *(with* 9 *and* 9 *degrees of freedom) and the critical region consists of the values that aren't included in that interval. The realized test statistic is* $f = 0.64$*, and it's not in the critical region. No proof about the inequality of the variances was obtained, so* H_0 *isn't rejected. (The P-probability is* $P = 0.517$*.)* **[10.6, 10.14]**

3.8 Graphical Methods for Comparing Means [10.10]

A glimpse to a graphical display obtained from the population often gives quite a good image about the matter, at least when considering the expectations. In a graphical display, a usual element is a *means diamond* $\Diamond$. In the middle of it there is the sample mean and the vertices give the 95 % confidence interval (by assuming that the population distribution is at least nearly normal).

As a some sort of rule of thumb it is often mentioned that if the quantile box of either of the samples doesn't include the median of the other sample, then the population expectations aren't equal.

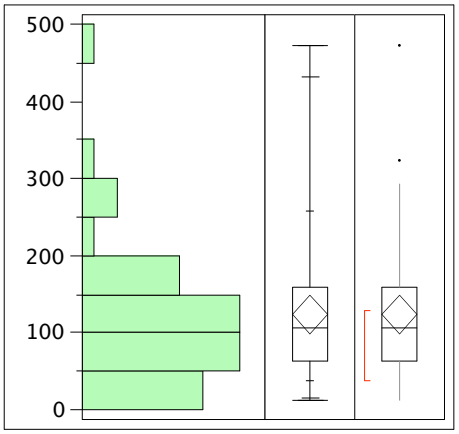
See section 1.3.

Example. *Let's consider the committed robberies and assaults in* 50 *states of USA during a certain time span, the unit is crimes per* 100000 *inhabitants. The JMP-program prints the following graphical display:*

This is not an actual sample, except with respect to the time span.

Distributions

robbery



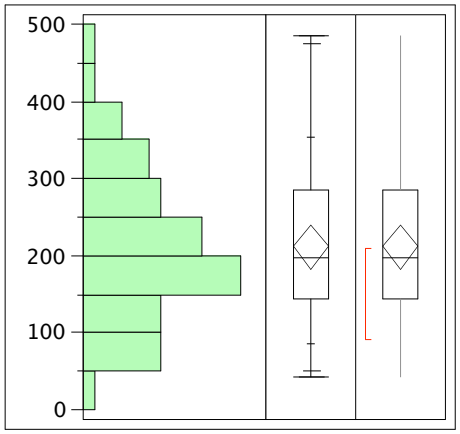
Quantiles

100.0%	maximum	472.60
99.5%		472.60
97.5%		431.49
90.0%		256.84
75.0%	quartile	160.03
50.0%	median	106.05
25.0%	quartile	63.85
10.0%		38.75
2.5%		14.57
0.5%		13.30
0.0%	minimum	13.30

Moments

Mean	124.092
Std Dev	88.348567
Std Err Mean	12.494374
upper 95% Mean	149.20038
lower 95% Mean	98.983615
N	50

assault



Quantiles

100.0%	maximum	485.30
99.5%		485.30
97.5%		475.35
90.0%		353.84
75.0%	quartile	284.73
50.0%	median	197.60
25.0%	quartile	143.43
10.0%		86.20
2.5%		49.27
0.5%		43.80
0.0%	minimum	43.80

Moments

Mean	211.3
Std Dev	100.25305
Std Err Mean	14.177922
upper 95% Mean	239.7916
lower 95% Mean	182.8084
N	50

The two outliers are New York and Nevada (Las Vegas, Reno). The hook-like (red) intervals are the *shortest halves* or the *densest halves* of the sample.

When measuring by using the above mentioned criterion, these two types of crime don't occur similarly according to expectations. Additionally, the distribution of robberies doesn't seem to be normal.

Chapter 4

χ^2-TESTS

Karl (Carl) Pearson (1857–1936), the "father" of statistics

By "χ^2-tests" it's not usually meant the preceding test concerning variance but a group of tests based on the Pearson-approximation and contingency tables.

4.1 Goodness-of-Fit Test [10.14]

The population distribution is often assumed to be known, for example a normal distribution, and its parameters are known. But are the assumptions correct? This is a hypothesis and it can be tested statistically.

Let's begin with a finite discrete distribution. There are a finite number of possible population cases, say the cases $T_1, \ldots, T_k$. The (point) probabilities of these

$$\mathrm{P}(T_1) = p_1 \ , \ldots , \ \mathrm{P}(T_k) = p_k$$

are supposed to be known, which is the null hypothesis H_0 of the test. The alternative hypothesis H_1 is that at least for one i $\mathrm{P}(T_i) \neq p_i$.

Actually at least for two, for $p_1 + \cdots + p_k = 1$.

For the test, let's take a sample with n elements, from which we determine the realized (absolute) frequencies $f_1, \ldots, f_k$ of the cases $T_1, \ldots, T_k$. These can be also considered to be random variables $F_1, \ldots, F_k$ and $\mathrm{E}(F_i) = np_i$. The test is based on the fact that the random variable

Cf. the expectation of the binomial distribution, just merge (pool) cases other than T_i.

$$H = \sum_{i=1}^{k} \frac{(F_i - np_i)^2}{np_i}$$

has nearly the χ^2-distribution with $k-1$ degrees of freedom. This is the *Pearson approximation.* As an additional restriction it is however often mentioned that none of the values $np_1, \ldots, np_k$ should be less than 5.

A result difficult to prove!

Some people however claim that 1.5 is already enough.

The test statistic is thus

$$h = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}$$

and when testing with it, only the tail of the χ^2-distribution is used. The deviations in the realized frequencies $f_1, \ldots, f_k$ namely result in increasing of h. There are web-based calculators to calculate this test statistic.

Example. *Let's consider a case where a dice is rolled* $n = 120$ *times. The expected probability of each face to occur is of course* $1/6$*, but is it actually so? The null hypothesis is* $\mathrm{H}_0 : p_1 = \cdots = p_6 = 1/6$ *and* $np_1 = \cdots = np_6 = 20$*. The obtained frequencies of each face are the following:*

Face i	1	2	3	4	5	6
Frequency f_i	20	22	17	18	19	24

By calculating from these, we obtain $h = 1.70$*. On the other hand, for example* $h_{0.05} = 11.070$ *(with* 5 *degrees of freedom) is much greater and there is no evidence to reject* H_0*.*

The testing of a continuous population distribution is done in a similar way. Then the range is divided into a finite number of intervals (cases $T_1, \ldots, T_k$). The probabilities $p_1, \ldots, p_k$ of these, according to the expected population distribution, are known (if H_0 is true) and the testing is done by using the Pearson-approximation as before.

Another widely used test for continuous distributions is the *Kolmogorov–Smirnov test*, which is not considered in this course.

Example. *Let's consider a case, where the population distribution is supposed to be a normal distribution: the expectation* $\mu = 3.5$ *and the standard deviation* $\sigma = 0.7$*. The range was divided into four intervals, the probabilities of which are obtained from the* $\mathrm{N}(3.5, 0.7^2)$ *distribution. The sample size is* $n = 40$*. The following results were obtained:*

i	1	2	3	4
T_i	$(-\infty, 2.95]$	$(2.95, 3.45]$	$(3.45, 3.95]$	$(3.95, \infty)$
p_i	0.2160	0.2555	0.2683	0.2602
np_i	8.6	10.2	10.7	10.4
f_i	7	15	10	8

By calculating from these, the value $h = 3.156$ *is obtained for the test statistic. Because* $h_{0.05} = 7.815$ *(with* 3 *degrees of freedom), the null hypothesis isn't rejected at the risk* $\alpha = 0.05$*.*

In above, the supposed population distribution has to be known in order to calculate probabilities related to it. There are also tests that test if the distribution is normal or not, without knowing its expectation or variance. Such a test is the *Lilliefors test* (and the *Geary's test* mentioned in WMMY). Also a χ^2-test similar to that in the preceding example can be performed using an expectation $\overline{x}$ estimated from the sample and a standard deviation s. The number of degrees of freedom is then however $k - 3$ and the precision suffers as well.

Also known as the Kolmogorov–Smirnov–Lilliefors test or the KSL test.

Hubert Lilliefors

4.2 Test for Independence. Contingency Tables [10.15]

The Pearson-approximation is suitable in many other situations. One such a situation is the testing of statistical independence of two different populations. In order the result to be interesting, the populations

must of course have some connection. The sampling is made in both the populations simultaneously .

Let's also here first consider a population, whose distributions are finite and discrete. The cases of the population 1 are $T_1, \ldots, T_k$ and their (point) probabilities are

$$\mathrm{P}(T_1) = p_1, \ldots, \mathrm{P}(T_k) = p_k.$$

These are often presented as vectors:

$$\mathbf{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix} \text{ and } \mathbf{q} = \begin{pmatrix} q_1 \\ \vdots \\ q_l \end{pmatrix}.$$

The cases of the population 2 are $S_1, \ldots, S_l$ and their (point) probabilities are

$$\mathrm{P}(S_1) = q_1, \ldots, \mathrm{P}(S_l) = q_l.$$

Additionally, we need the (point) probabilities

$$\mathrm{P}(T_i \cap S_j) = p_{i,j} \quad (i = 1, \ldots, k \text{ ja } j = 1, \ldots, l).$$

This is often presented as a matrix:

$$\mathbf{P} = \begin{pmatrix} p_{1,1} & \cdots & p_{1,l} \\ \vdots & & \vdots \\ p_{k,1} & \cdots & p_{k,l} \end{pmatrix}.$$

None of these probabilities is however supposed to be known; the testing is purely based on the values obtained from the samples. Let's introduce the following notations.. The frequencies of the cases $T_1, \ldots, T_k$ as random variables are $F_1, \ldots, F_k$ and as realized values in the sample $f_1, \ldots, f_k$. The frequencies of the cases $S_1, \ldots, S_l$ as random variables are $G_1, \ldots, G_l$ and as realized values from the sample $g_1, \ldots, g_l$. The frequency of the pooled case $T_i \cap S_j$ as a random variable is $F_{i,j}$ and as a realized value from the sample $f_{i,j}$.

These are presented in a *contingency table* in the following form, where n is the sample size:

	S_1	S_2	$\cdots$	S_l	Σ
T_1	$f_{1,1}$	$f_{1,2}$	$\cdots$	$f_{1,l}$	f_1
T_2	$f_{2,1}$	$f_{2,2}$	$\cdots$	$f_{2,l}$	f_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
T_k	$f_{k,1}$	$f_{k,2}$	$\cdots$	$f_{k,l}$	f_k
Σ	g_1	g_2	$\cdots$	g_l	n

A similar table could also be done for frequencies considered to be random variables.

Population distributions are independent when

This is the definition of independence, in a matrix form $\mathbf{P} = \mathbf{p}\mathbf{q}^{\mathsf{T}}$.

$$\mathrm{P}(T_i \cap S_j) = \mathrm{P}(T_i)\mathrm{P}(S_j) \quad \text{or} \quad p_{i,j} = p_i q_j \quad (i = 1, \ldots, k \text{ ja } j = 1, \ldots, l).$$

This independence is now the null hypothesis H_0. The alternative hypothesis claims that at least for one index pair i, j there holds $p_{i,j} \neq p_i q_j$. Thus, when H_0 is true, the frequencies should fulfill the corresponding equations (cf. the binomial distribution):

$$\mathrm{E}(F_{i,j}) = np_{i,j} = np_i q_j = \frac{1}{n}\mathrm{E}(F_i)\mathrm{E}(G_j).$$

Let's now form a test statistic like before in goodness-of-fit testing by considering the frequency $f_{i,j}$ to be realized and the value $f_i g_j / n$ given by the right hand side to be expected, that is according to H_0:

$$h = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(f_{i,j} - f_i g_j/n)^2}{f_i g_j/n}.$$

The formula could be presented in a matrix form as well.

There are also web-calculators to calculate this test statistic from the given contingency tables.

According to the Pearson-approximation, the corresponding random variable

$$H = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(F_{i,j} - F_i G_j/n)^2}{F_i G_j/n}.$$

has nearly the χ^2-distribution, but now with $(k-1)(l-1)$ degrees of freedom. The worse the equations $f_{i,j} \cong f_i g_j/n$ hold, the greater becomes the value of h. The critical region is again the right tail of the χ^2-distribution in question.

Example. *Let's as an example consider a case where a sample of* $n = 309$ *defective products. The product is made in three different production lines* L_1, L_2 *and* L_3 *and there are four different kinds of faults* V_1, V_2, V_3 *and* V_4. *The null hypothesis here is that the distributions of faults in terms of fault types and production lines are independent. The obtained contingency table is*

	V_1	V_2	V_3	V_4	Σ
L_1	15(22.51)	21(20.99)	45(38.94)	13(11.56)	94
L_2	26(22.90)	31(21.44)	34(39.77)	5(11.81)	96
L_3	33(28.50)	17(26.57)	49(49.29)	20(14.63)	119
Σ	74	69	128	38	309

The values in brackets are the numbers $f_i g_j/n$. *The realized calculated value of the test statistic is* $h = 19.18$. *This corresponds to the P-probability* $P = 0.0039$ *obtained from the* χ^2*-distribution (with* 6 *degrees of freedom). At the risk* $\alpha = 0.01$, H_0 *can thus be rejected and it can be concluded that the production line affects the type of the fault.*

Here also it's often recommended that all the values $f_i g_j/n$ should be at least 5. This certainly is the case in the previous example.

The independence of continuous distributions can also be tested in this manner. Then the ranges are divided into a finite number of intervals, just like in the goodness-of-fit test, and the testing is done as described above.

4.3 Test for Homogeneity [10.16]

In the test of independence, the sample is formed randomly in terms of both populations. A corresponding test is obtained when the number of elements taken into the sample is determined beforehand for one of the populations.

If in above, the values are determined for population 2, then the frequencies $g_1, \ldots, g_l$ are also determined beforehand when the sample size is $n = g_1 + \cdots + g_l$. The null hypothesis is however exactly similar to

the above mentioned. Only its meaning is different: Here H_0 claims that the distribution of population 1 is similar for different types of elements $S_1, \ldots, S_l$, in other words the population distribution is *homogeneous* in terms of element types $S_1, \ldots, S_l$. Note that here $S_1, \ldots, S_l$ aren't cases and they don't have probabilities. They are simply types, in which the elements of the population 1 can be divided to, and it is determined beforehand how much each of these types are being taken into the sample.

Now $f_{i,j}$ and $F_{i,j}$ denote the frequency of the population elements of the type S_j in the sample. If H_0 is true, then the probability that T_i occurs to elements of the type S_j is the same as the probability to the whole population, namely p_i. In terms of expectations then

Cf. the binomial distribution again.

$$\mathrm{E}(F_{i,j}) = g_j p_i = \frac{1}{n}\mathrm{E}(F_i) g_j \quad (i = 1, \ldots, k \text{ ja } j = 1, \ldots, l).$$

The test statistics H and h and the approximative χ^2-distribution related to them are thus exactly the same as before in the test of independence.

Example. *As an example we consider a case, where the popularity of a proposed law was studied in USA.* $n = 500$ *people were chosen as follows:* $g_1 = 200$ *Democrats,* $g_2 = 150$ *Republicans and* $g_3 = 150$ *independent. These people were asked if they were for or against the proposition or neither. The question of interest was, are the people with different opinions about the proposition identically distributed in different parties (this is* H_0*).*

The contingency table was obtained

	Democrat	*Republican*	*Independent*	Σ
Pro	82(85.6)	70(64.2)	62(64.2)	214
Con	93(88.8)	62(66.6)	67(66.6)	222
No opinion	25(25.6)	18(19.2)	21(19.2)	64
Σ	200	150	150	500

From this we can calculate the test statistic $h = 1.53$*. By using the* χ^2*-distribution (with* 4 *degrees of freedom) we obtain the P-probability* $P = 0.8213$*. There is practically no evidence to reject the null hypothesis* H_0 *according to this data.*

If $k = 2$ in the test of homogeneity, we have a special case, which is about the similarity test of the l binomial distributions' $\mathrm{Bin}(n_1, p_1), \ldots, \mathrm{Bin}(n_l, p_l)$ parameters $p_1, \ldots, p_l$. Then $g_1 = n_1, \ldots, g_l = n_l$ and the null hypothesis is

$$\mathrm{H}_0 : p_1 = \cdots = p_l \,(= p).$$

The common parameter value p is not assumed to be known.

The alternative hypothesis H_1 claims that at least two of the parameters aren't equal.

In order to examine the matter, we perform tests and obtain the numbers of realized favorable cases $x_1, \ldots, x_l$. The contingence table is in this case of the form

	$\mathrm{Bin}(n_1, p_1)$	$\mathrm{Bin}(n_2, p_2)$	$\cdots$	$\mathrm{Bin}(n_l, p_l)$	Σ
Favorable	x_1	x_2	$\cdots$	x_l	x
Unfavorable	$n_1 - x_1$	$n_2 - x_2$	$\cdots$	$n_l - x_l$	$n - x$
Σ	n_1	n_2	$\cdots$	n_l	n

where $x = x_1 + \cdots + x_l$ and $n = n_1 + \cdots + n_l$. The test proceeds similarly as before by using an approximative χ^2-distribution (now with $(2-1)(l-1) = l-1$ degrees of freedom). The test statistic can be written in various forms:

$$\begin{aligned} h &= \sum_{i=1}^{l} \frac{(x_i - xn_i/n)^2}{xn_i/n} + \sum_{i=1}^{l} \frac{\big(n_i - x_i - (n-x)n_i/n\big)^2}{(n-x)n_i/n} \\ &= \sum_{i=1}^{l} (x_i - xn_i/n)^2 \Big(\frac{1}{xn_i/n} + \frac{1}{(n-x)n_i/n} \Big) \\ &= \sum_{i=1}^{l} \frac{(x_i - xn_i/n)^2}{x(n-x)n_i/n^2} = \sum_{i=1}^{l} \frac{(x_i - n_i x/n)^2}{n_i(x/n)(1-x/n)}. \end{aligned}$$

The last form is perhaps most suitable for manual calculation, and from it the reason why we end up to χ^2-distribution can be seen. If the null hypothesis H_0 is true, the realized x/n is nearly p, and the random variable

Cf. the distribution of the sample variance of a normally distributed population.

$$\frac{X_i - n_i p}{\sqrt{n_i p(1-p)}}$$

is, by the normal approximation of the binomial distribution, nearly the standard normal distribution.

Example. *Let's consider, as an example, a situation before an election, where three different studies gave to a party the supporter numbers* $x_1 = 442$, $x_2 = 313$ *and* $x_3 = 341$ *while the corresponding sample sizes were* $n_1 = 2002$, $n_2 = 1532$ *and* $n_3 = 1616$. *Could these studies give every party the same percentage of support* (H_0)*? By calculating we obtain the realized test statistic* $h = 1.451$ *and the corresponding P-probability* $P = 0.4841$ *(χ^2-distribution with 2 degrees of freedom). According to this, there is practically no reason to doubt that the percentages of support given by the different studies wouldn't be equal.*

Chapter 5

MAXIMUM LIKELIHOOD ESTIMATION

5.1 Maximum Likelihood Estimation [9.14]

Many of the estimators above can be obtained by a common method. If the values to be estimated are the parameters $\theta_1, \ldots, \theta_m$ of the population distribution, and the density function of the distribution is $f(x; \theta_1, \ldots, \theta_m)$, then we try to obtain formulas for the estimators $\hat{\Theta}_1, \ldots, \hat{\Theta}_m$ by using the sample elements $X_1, \ldots, X_n$ considered to be random variables, or at least a procedure, by which the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_m$ can be calculated from the realized sample elements $x_1, \ldots, x_n$.

The parameters are included in the density function so that the dependence on them would be visible.

Because the sample elements $X_1, \ldots, X_n$ are taken independently in a random sampling, they all have the same density function and the density function of their pooled distribution is the product

$$g(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m) = f(x_1; \theta_1, \ldots, \theta_m) \cdots f(x_n; \theta_1, \ldots, \theta_m).$$

In *maximum likelihood estimation* or *MLE*, the estimators $\hat{\Theta}_1, \ldots, \hat{\Theta}_m$ are determined so that

$$g(X_1, \ldots, X_n; \theta_1, \ldots, \theta_m) = f(X_1; \theta_1, \ldots, \theta_m) \cdots f(X_n; \theta_1, \ldots, \theta_m)$$

obtains its greatest value when

$$\theta_1 = \hat{\Theta}_1 \ , \ldots , \ \theta_m = \hat{\Theta}_m.$$

Similarly, the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are obtained when we maximize

$$g(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m) = f(x_1; \theta_1, \ldots, \theta_m) \cdots f(x_n; \theta_1, \ldots, \theta_m).$$

The basic idea is to estimate the parameters so that the density probability of the observed values is the greatest.

In maximum likelihood estimation the notation

$$L(\theta_1, \ldots, \theta_m; X_1, \ldots, X_n) = f(X_1; \theta_1, \ldots, \theta_m) \cdots f(X_n; \theta_1, \ldots, \theta_m)$$

and similarly

$$L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n) = f(x_1; \theta_1, \ldots, \theta_m) \cdots f(x_n; \theta_1, \ldots, \theta_m)$$

and it's called the *likelihood function* or the *likelihood*. It's often easier to maximize the logarithm of the likelihood

$$\begin{aligned} l(\theta_1, \ldots, \theta_m; X_1, \ldots, X_n) &= \ln L(\theta_1, \ldots, \theta_m; X_1, \ldots, X_n) \\ &= \ln \big(f(X_1; \theta_1, \ldots, \theta_m) \cdots f(X_n; \theta_1, \ldots, \theta_m) \big) \\ &= \ln f(X_1; \theta_1, \ldots, \theta_m) + \cdots + \ln f(X_n; \theta_1, \ldots, \theta_m), \end{aligned}$$

the *loglikelihood (function)* and similarly

$$l(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n) = \ln f(x_1; \theta_1, \ldots, \theta_m) + \cdots + \ln f(x_n; \theta_1, \ldots, \theta_m).$$

With these notations, the result of estimation can be succinctly written in the form

$$(\hat{\theta}_1, \ldots, \hat{\theta}_m) = \operatorname*{argmax}_{\theta_1, \ldots, \theta_m} L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n)$$

or

$$(\hat{\theta}_1, \ldots, \hat{\theta}_m) = \operatorname*{argmax}_{\theta_1, \ldots, \theta_m} l(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n).$$

5.2 Examples [9.14]

Example. *The value to be estimated is the parameter λ of the Poisson distribution. The density function of the distribution is* [9.19]

$$f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

The likelihood (for the random variable sample) is thus

$$L(\lambda; X_1, \ldots, X_n) = \frac{\lambda^{X_1}}{X_1!} e^{-\lambda} \cdots \frac{\lambda^{X_n}}{X_n!} e^{-\lambda} = \frac{\lambda^{X_1 + \cdots + X_n}}{X_1! \cdots X_n!} e^{-n\lambda}$$

and the corresponding loglikelihood is

$$l(\lambda; X_1, \ldots, X_n) = -\ln(X_1! \cdots X_n!) + (X_1 + \cdots + X_n) \ln \lambda - n\lambda.$$

To find the maximum we set the derivative with respect to λ to zero

The case $X_1 = \cdots = X_n = 0$ must be considered separately. Then $\hat{\Lambda} = 0$.

$$\frac{\partial l}{\partial \lambda} = \frac{1}{\lambda}(X_1 + \cdots + X_n) - n = 0,$$

and solve it to obtain the maximum likelihood estimator:

$$\hat{\Lambda} = \frac{1}{n}(X_1 + \cdots + X_n) = \overline{X}.$$

By using the second derivative we can verify that we found the maximum. Similarly, we obtain as the maximum likelihood estimate the sample mean

This is of course natural since the expectation is λ.

$$\hat{\lambda} = \overline{x}.$$

Example. *The population distribution is a normal distribution* $\mathrm{N}(\mu, \sigma^2)$, *whose parameters are in this case* $\theta_1 = \mu$ *and* $\theta_2 = \sigma^2$. *The density function is then* [**9.20**]

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

and the likelihood (this time for the realized sample) is

$$\begin{aligned} L(\mu, \sigma^2; x_1, \ldots, x_n) &= \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2\sigma^2}(x_1-\mu)^2} \cdots \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2} \\ &= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}}\, e^{-\frac{1}{2\sigma^2}((x_1-\mu)^2+\cdots+(x_n-\mu)^2)} \end{aligned}$$

and the corresponding loglikelihood is

$$l(\mu, \sigma^2; x_1, \ldots, x_n) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\big((x_1-\mu)^2+\cdots+(x_n-\mu)^2\big).$$

To maximize, let's set the partial derivatives with respect to μ*:n and* σ^2 *to zero:*

Here the variable is σ^2, not σ.

$$\begin{cases} \dfrac{\partial l}{\partial \mu} = \dfrac{1}{\sigma^2}\big((x_1-\mu)+\cdots+(x_n-\mu)\big) = \dfrac{1}{\sigma^2}(x_1+\cdots+x_n-n\mu) = 0 \\ \dfrac{\partial l}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2(\sigma^2)^2}\big((x_1-\mu)^2+\cdots+(x_n-\mu)^2\big) = 0. \end{cases}$$

By solving the first equation, we obtain a familiar estimate for μ

$$\hat{\mu} = \frac{1}{n}(x_1+\cdots+x_n) = \overline{x}.$$

By inserting this in the second equation we obtain the maximum likelihood estimate for σ^2

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

By examining the second partial derivatives, we can verify that this is the maximum.

Surprisingly the result concerning σ^2 *isn't* the earlier used sample variance s^2. Because

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

is an unbiased estimate for σ^2, the maximum likelihood estimate of σ^2 for a normal distribution $\mathrm{N}(\mu, \sigma^2)$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

is thus a little biased.

This proves, that it's not favorable in all cases that the estimate is unbiased.

Example. *Let's consider, as an example, a case where the population distribution is a uniform distribution over the interval* $[a, b]$*, whose endpoints are unknown. If the realized sample values are* $x_1, \ldots, x_n$*, the most natural estimates would seem to be* $\min(x_1, \ldots, x_n)$ *for the endpoint* a *and* $\max(x_1, \ldots, x_n)$ *for the endpoint* b*. But are these the maximum likelihood estimates?*

The density function of the distribution is now

$$f(x; a, b) = \begin{cases} \dfrac{1}{b-a}, \text{ when } a \leq x \leq b \\ 0 \text{ otherwise.} \end{cases}$$

It is clear that in order to maximize the likelihood

$$L(a, b; x_1, \ldots, x_n) = f(x_1; a, b) \cdots f(x_n; a, b)$$

we have to choose endpoint estimates $\hat{a}$ *and* $\hat{b}$ *such that all the sample elements are included in the interval* $[\hat{a}, \hat{b}]$*, otherwise the likelihood would be* $= 0$ *and that's not the greatest possible. Under this condition, the likelihood is*

$$L(a, b; x_1, \ldots, x_n) = \frac{1}{(b-a)^n}$$

and it achieves its greatest value when $b - a$ *is the smallest possible. The estimates*

$$\begin{cases} \hat{a} = \min(x_1, \ldots, x_n) \\ \hat{b} = \max(x_1, \ldots, x_n) \end{cases}$$

are thus confirmed to be also the maximum likelihood estimates.

If under consideration was a uniform distribution over the open interval (a, b), the maximum likelihood estimates wouldn't exist at all.

Chapter 6

MULTIPLE LINEAR REGRESSION

6.1 Regression Models [12.1]

In linear (multiple) regression, a phenomenon is considered to be modeled mathematically in the form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon.$$

The different components in the model are the following:

1. $x_1, \ldots, x_k$ are the inputs of the model. These are given different names depending on the situation and the field of application: *independent variables*, *explanatory variables*, *regressors*, *factors* or *exogenous variables.*

   "Regressor" in the following.

2. y is the output of the model. It's also given different names, for example the *depending variable*, *response* or *endogenous variable.*

   "Response" in the following.

3. $\beta_0, \beta_1, \ldots, \beta_k$ are the *parameters of the model* or the *coefficients* of the model. They are fixed values, that are estimated from the obtained sample data when constructing the model. The parameter β_0 is the *intercept.*

4. ϵ is a random variable, whose expectation is $= 0$ and which has a variance σ^2, the *error term.* The response y is thus a random variable and its expectation is $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ and variance is σ^2.

The model functions so that its input are the regressors and its output is the value of the response, which is affected by the realized value of the error term.

The *linearity* of the model means that it's linear with respect to its parameters. Regressors may very well depend on one another. A usual model is for example a *polynomial model*

Correspondingly, we could consider and use nonlinear regression models.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon,$$

where the regressors are the powers of the single parameter x. Note that this as well is a linear model for it is linear with respect to its parameters.

6.2 Estimating the Coefficients. Using Matrices [12.2–3]

In order to fit the model, its parameters are estimated by using the sample data. The following n ordered k-tuples are chosen for the regressors

x_1	x_2	$\cdots$	x_k
$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,k}$
$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,k}$
$\vdots$	$\vdots$		$\vdots$
$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,k}$

The indices are chosen with the matrix presentation in mind.

Let's perform n experiments by using each k-tuples as an input and let's denote the obtained response values $y_1, y_2, \ldots, y_n$. The latter can be consider to be either realized values or random variables. The regressor k-tuples don't have to be distinct, the same tuple can be used more than once.

This may be even an advantage, for it improves the estimator of the variance σ^2.

From the table above we see that a matrix presentation could be very useful. Let's now denote

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Note especially the first column in the matrix $\mathbf{X}$.

and moreover for the parameters

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

With these markings we can write the results of the whole experiment series simply in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Data model.

Here $\epsilon_1, \ldots, \epsilon_n$ are either realized values of the random variable ϵ or independent random variables that all have the same distribution as ϵ. Note that if $\epsilon_1, \ldots, \epsilon_n$ are considered to be random variables, then $y_1, \ldots, y_n$ have to be considered similarly and then y_i depends only on ϵ_i.

To avoid confusion, these different interpretations *aren't* denoted differently unlike in the previous chapters. That is, lower case letters are used to denote also random variables. The case can be found out by its context.

Furthermore, note that if $y_1, \ldots, y_n$ are considered to be random variables or $\mathbf{y}$ is considered a random vector, then the expectation (vector) of $\mathbf{y}$ is $\mathbf{X}\boldsymbol{\beta}$. The matrix $\mathbf{X}$ is on the other hand a given matrix, which is usually called the *data matrix*. In many applications the matrix $\mathbf{X}$ is determined by circumstances outside the statistician's control, even though it can have a significant influence on the success of parameter estimation.

There is a whole field of statistics on how to make the best possible choice of $\mathbf{X}$, *experimental design*.

The idea behind the estimation of the parameters $\beta_0, \beta_1, \ldots, \beta_k$ (that is vector $\boldsymbol{\beta}$) is to fit the realized output vector $\mathbf{y}$ as well as possible to its expectation, that is $\mathbf{X}\boldsymbol{\beta}$. This can be done in many ways, the most

usual of which the *least sum of squares.* Then we choose the parameters $\beta_0, \beta_1, \ldots, \beta_k$, or the vector $\boldsymbol{\beta}$ so that

$$N(\beta_0, \beta_1, \ldots, \beta_k) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_k x_{i,k})^2$$

obtains its least value. Thus we obtain the parameter estimates

$$\hat{\beta}_0 = b_0 \ , \ \hat{\beta}_1 = b_1 \ , \ldots , \ \hat{\beta}_k = b_k,$$

in the form of vector $\hat{\boldsymbol{\beta}} = \mathbf{b}$, where

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}.$$

The estimates $b_0, b_1, \ldots, b_k$ are obtained by setting the partial derivatives of $N(\beta_0, \beta_1, \ldots, \beta_k)$ with respect to the parameters $\beta_0, \beta_1, \ldots, \beta_k$ equal to 0 and by solving for them from the obtained equations. These equations are the *normal equations.* The partial derivatives are

$$\begin{aligned}
\frac{\partial N}{\partial \beta_0} &= -2 \sum_{i=1}^{n} 1 \cdot (y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_k x_{i,k}), \\
\frac{\partial N}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_{i,1} (y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_k x_{i,k}), \\
&\vdots \\
\frac{\partial N}{\partial \beta_k} &= -2 \sum_{i=1}^{n} x_{i,k} (y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_k x_{i,k}).
\end{aligned}$$

When setting these equal to 0, we may cancel out the factor -2, and a matrix form equation is obtained for $\mathbf{b}$

$$\mathbf{X}^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0} \quad \text{or} \quad (\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{b} = \mathbf{X}^\mathsf{T}\mathbf{y}.$$

If $\mathbf{X}^\mathsf{T}\mathbf{X}$ is non-singular (invertible) matrix, as it's assumed in the following, we obtain the solution

If $\mathbf{X}^\mathsf{T}\mathbf{X}$ is singular or nearly singular (*multicollinearity*), statistical programs warn about it.

$$\mathbf{b} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

Estimation requires thus a lot of numerical calculations. There are web-calculators for the most common types of problems, but large problems have to be calculated with statistical programs.

Example. *Let's fit the regression model* [**12.4**]

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,1} x_1^2 + \beta_{2,2} x_2^2 + \beta_{1,2} x_1 x_2 + \epsilon.$$

Note that the regressors are independent and similarly indexed parameters!

Terms in a product form, like $x_1 x_2$ here, are called interaction terms. *Here x_1 is sterilization time (min) x_2 sterilization temperature (°C). The output y is the number of (organic) pollutants after sterilization. The test result are the following:*

	x_2		
x_1	$75\ ^\circ C$	$100\ ^\circ C$	$125\ ^\circ C$
$15\ min$	14.05	10.55	7.55
$15\ min$	14.93	9.48	6.59
$20\ min$	16.56	13.63	9.23
$20\ min$	15.85	11.75	8.78
$25\ min$	22.41	18.55	15.93
$25\ min$	21.66	17.98	16.44

By calculating from these we obtain the data matrix $\mathbf{X}$ *(remember that we should calculate all the columns corresponding to the five regressors). The result is a* 18×6*-matrix, from which here are a few rows and the corresponding responses:*

$$\mathbf{X} = \begin{pmatrix} 1 & 15 & 75 & 15^2 & 75^2 & 15 \cdot 75 \\ 1 & 15 & 100 & 15^2 & 100^2 & 15 \cdot 100 \\ 1 & 15 & 125 & 15^2 & 125^2 & 15 \cdot 125 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 20 & 75 & 20^2 & 75^2 & 20 \cdot 75 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \quad , \quad \mathbf{y} = \begin{pmatrix} 14.05 \\ 10.55 \\ 7.55 \\ \vdots \\ 16.56 \\ \vdots \end{pmatrix} .$$

In JMP-program, the data is inputed using a data editor or read from a file. The Added columns can easily be calculated in the editor (or formed when estimating):

Rows	Time	Temperature	Response	Time*Time	Temperature*Temperature	Time*Temperature
1	15	75	14.05	225	5625	1125
2	15	100	10.55	225	10000	1500
3	15	125	7.55	225	15625	1875
4	15	75	14.93	225	5625	1125
5	15	100	9.48	225	10000	1500
6	15	125	6.59	225	15625	1875
7	20	75	16.56	400	5625	1500
8	20	100	13.63	400	10000	2000
9	20	125	9.23	400	15625	2500
10	20	75	15.85	400	5625	1500
11	20	100	11.75	400	10000	2000
12	20	125	8.78	400	15625	2500
13	25	75	22.41	625	5625	1875
14	25	100	18.55	625	10000	2500
15	25	125	15.93	625	15625	3125
16	25	75	21.66	625	5625	1875
17	25	100	17.98	625	10000	2500
18	25	125	16.44	625	15625	3125

$\mathbf{X}^\mathsf{T}\mathbf{X}$ *is thus a* 6×6*-matrix. The numerical calculations are naturally also here done by computers and statistical programs. The obtained parameter estimates are*

$$b_0 = 56.4411\ , \quad b_1 = -2.7530\ , \ b_2 = -0.3619\ , \ b_{1,1} = 0.0817\ ,$$
$$b_{2,2} = 0.0008\ \ , \ b_{1,2} = 0.0031.$$

The (a little trimmed) print of the JMP-program is the following:

A lot of other information is included here, to which we'll return later.

Summary of Fit

RSquare	0.986408
RSquare Adj	0.980745
Root Mean Square Error	0.647809
Mean of Response	13.99556
Observations (or Sum Wgts)	18

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	365.47657	73.0953	174.1791
Error	12	5.03587	0.4197	Prob > F
C. Total	17	370.51244		<.0001

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	3	0.9211722	0.307057	0.6716
Pure Error	9	4.1147000	0.457189	Prob > F
Total Error	12	5.0358722		0.5906
				Max RSq
				0.9889

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	56.441111	7.994016	7.06	<.0001
Time	−2.753	0.550955	−5.00	0.0003
Temperature	−0.361933	0.110191	−3.28	0.0065
Time*Time	0.0817333	0.012956	6.31	<.0001
Temperature*Temperature	0.0008133	0.000518	1.57	0.1425
Time*Temperature	0.00314	0.001832	1.71	0.1123

From the result we could conclude that the regressor x_2^2 in the model isn't necessary and there's not much combined effect between regressors x_1 and x_2, but conclusions like this have to be statistically justified!

6.3 Properties of Parameter Estimators [12.4]

In the random variable interpretation, the obtained parameters b_i are considered to be random variables (estimators) that depend on the random variables ϵ_i according to the vectorial equation

$$\mathbf{b} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{\beta} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{\epsilon}.$$

Because $\mathrm{E}(\epsilon_1) = \cdots = \mathrm{E}(\epsilon_n) = 0$, from the equation above we can quite clearly see that $\mathrm{E}(b_i) = \beta_i$, in other words the *parameter estimators are unbiased.* Furthermore, by some short matrix calculation we can note that the $(k+1) \times (k+1)$"-matrix $\mathbf{C} = (c_{ij})$, where

$$\mathbf{C} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1},$$

and the indexes i and j go through values $0, 1, \ldots, k$, contains the information about the variances of the parameter estimators and about their mutual covariances in the form

$$\mathrm{var}(b_i) = c_{ii}\sigma^2 \quad \text{and} \quad \mathrm{cov}(b_i, b_j) = c_{ij}\sigma^2.$$

An important estimator/estimate is the *estimated response*

$$\hat{y}_i = b_0 + b_1 x_{i,1} + \cdots + b_k x_{i,k}$$

and the *residual* obtained from it

$$e_i = y_i - \hat{y}_i.$$

The residual represents that part of the response that couldn't be explained with the estimated model. In the vector form, we correspondingly obtain the estimated response vector

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

and from it the *residual vector*

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = \big(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\big)\mathbf{y}.$$

Here $\mathbf{I}_n$ is a $n \times n$ identity matrix.

The matrices presented above, by the way, have their own customary names and notations:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} \quad \text{a (\textit{hat matrix}) and}$$
$$\mathbf{P} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} = \mathbf{I}_n - \mathbf{H} \quad \text{(a \textit{projection matrix})}.$$

Multiplying with $\mathbf{H}$ projects the data matrix of the response vector into column space of the data matrix, multiplying with $\mathbf{P}$ projects into its orthogonal complement.

By a little calculation we can note that $\mathbf{H}^\mathsf{T} = \mathbf{H}$ and $\mathbf{P}^\mathsf{T} = \mathbf{P}$, and that $\mathbf{H}^2 = \mathbf{H}$ and $\mathbf{P}^2 = \mathbf{P}$. $\mathbf{H}$ and $\mathbf{P}$ are in other words symmetric idempotent matrices. Additionally, $\mathbf{PH}$ is a zero matrix. With these notations then

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \quad \text{and} \quad \mathbf{e} = \mathbf{P}\mathbf{y}.$$

The quantity

$$\|\mathbf{e}\|^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

is the *sum of squares of errors,* denoted often by SSE. By using it we obtain an unbiased estimator for the error variance σ^2. For this, let's expand the SSE. Firstly

$$\mathbf{e} = \mathbf{P}\mathbf{y} = \big(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\big)(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{P}\boldsymbol{\epsilon}.$$

Furthermore

$$\text{SSE} = \mathbf{e}^\mathsf{T}\mathbf{e} = (\mathbf{P}\boldsymbol{\epsilon})^\mathsf{T}\mathbf{P}\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^\mathsf{T}\mathbf{P}^\mathsf{T}\mathbf{P}\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^\mathsf{T}\mathbf{P}\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^\mathsf{T}\mathbf{H}\boldsymbol{\epsilon}.$$

If we denote $\mathbf{H} = (h_{ij})$, then we obtain

$$\text{SSE} = \sum_{i=1}^{n} \epsilon_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} \epsilon_i h_{ij} \epsilon_j.$$

For the expectation of the SSE (unbiased), we should remember that $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \mathrm{E}(\epsilon_i^2) = \sigma^2$. Furthermore, because ϵ_i and ϵ_j are independent when $i \neq j$, then they are also uncorrelated, in other words

$$\mathrm{cov}(\epsilon_i \epsilon_j) = \mathrm{E}(\epsilon_i \epsilon_j) = 0.$$

Thus,

$$\mathrm{E}(\text{SSE}) = \sum_{i=1}^{n} \mathrm{E}(\epsilon_i^2) - \sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}\mathrm{E}(\epsilon_i \epsilon_j) = n\sigma^2 - \sigma^2 \sum_{i=1}^{n} h_{ii}.$$

The sum on the right hand side is the sum of the diagonal elements of the hat matrix or its trace trace($\mathbf{H}$). One nice property about the trace is that it's commutative, in other words trace($\mathbf{AB}$) = trace($\mathbf{BA}$). By using this we may calculate the sum in question

$$\sum_{i=1}^{n} h_{ii} = \text{trace}(\mathbf{H}) = \text{trace}\big(\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\big) = \text{trace}\big((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\big)$$
$$= \text{trace}(\mathbf{I}_{k+1}) = k+1$$

Let's choose $\mathbf{A} = \mathbf{X}$ and $\mathbf{B} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$.

and then

$$\text{E(SSE)} = (n-k-1)\sigma^2.$$

Thus,

$$\text{E}\Big(\frac{\text{SSE}}{n-k-1}\Big) = \sigma^2,$$

and finally we obtain the wanted unbiased estimate/estimator

$$\widehat{\sigma^2} = \frac{\text{SSE}}{n-k-1}.$$

It's often denoted the *mean square error*

$$\text{MSE} = \frac{\text{SSE}}{n-k-1}$$

is almost always available in the printout of a statistical program, as well as the estimated standard deviation $\sqrt{\text{MSE}} = \text{RMSE}$. In the example above we obtain MSE = 0.4197 and RMSE = 0.6478.

"root mean square of error"

There are two other sums of squares that are usually in a printout of statistical programs:

$$\text{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \ , \ \text{where} \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

the *total sum of squares* and

$$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2,$$

the *sum of squares of regression.* These sums of squares, by the way, have a connection, which can be found by a matrix calculation (will be omitted here):

$$\text{SST} = \text{SSE} + \text{SSR}.$$

The corresponding mean squares are

$$\text{MST} = \frac{\text{SST}}{n-1} \quad \text{(the } \textit{total mean square}\text{) and}$$

$$\text{MSR} = \frac{\text{SSR}}{k} \quad \text{(the } \textit{mean square of regression}\text{)}.$$

At least the MSR is usually in the printouts of the programs.

As a matter of fact, there is a whole *analysis of variance table* or *ANOVA-table* in the printouts of the programs:

Source of variation	Degrees of freedom	Sums of squares	Mean squares	F
Regression	k	SSR	MSR	
Residual	$n-k-1$	SSE	$\widehat{\sigma^2} = \text{MSE}$	$F = \dfrac{\text{MSR}}{\text{MSE}}$
Total variation	$n-1$	SST	(MST)	

Note the sum: $n-1 = k+(n-k-1)$.

The quantity F in the table is a test statistic, with which, with some assumptions about normality, the significance of the regression can be tested by using the F-distribution (with k and $n-k-1$ degrees of freedom), as we'll see. There is also usually the realized P-probability of the test in the table. The ANOVA-table of the example above is

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	365.47657	73.0953	174.1791
Error	12	5.03587	0.4197	**Prob > F**
C. Total	17	370.51244		<.0001

and the mentioned estimate $\widehat{\sigma^2} = \text{MSE} = 0.4197$ from it.

6.4 Statistical Consideration of Regression [12.5]

A regression model is considered *insignificant* if all the parameters $\beta_1, \ldots, \beta_k$ are equal to zero. In that case, the chosen regressors have no effect on the response. Similarly, a single regressor x_i is *insignificant* if the corresponding parameter β_i is equal to zero. When testing the significance, there has to be some (sort of) distribution presented in order to calculate the probabilities. Because of this it's assumed that all the random variables ϵ_i have a $\text{N}(0, \sigma^2)$-distribution. In most of the cases, this is a natural assumption.

Note that β_0 isn't included.

When testing the significance of the whole model, the null hypothesis is

$$\text{H}_0 : \beta_1 = \cdots = \beta_k = 0.$$

The alternative hypothesis, for one, claims that at least one of the parameters $\beta_1, \ldots, \beta_k$ is $\neq 0$. It can be shown that if H_0 is true, then the quantity (random variable) in above mentioned ANOVA-table

The presented results concerning distributions are difficult to prove.

$$F = \frac{\text{MSR}}{\text{MSE}}$$

is F-distributed with k and $n-k-1$ degrees of freedom. The critical region is the right tail, for the insignificance of the model decreases the SSR and increases the SSE.

If H_0 isn't rejected, the model isn't too useful, even though the parameters would have been estimated. In the above mentioned example, for F we obtain a value 174.1791 (with 5 and 12 degrees of freedom) and the corresponding P-probability is close to zero. Thus, the model is very significant.

There is a test that uses the t-distribtution to test single parameters. The test is very similar to the t-tests presented earlier. It can be namely shown that if $\beta_i = \beta_{0,i}$, where $\beta_{0,i}$ is known, then the random variable

$$T_i = \frac{b_i - \beta_{0,i}}{\mathrm{RMSE}\sqrt{c_{ii}}}$$

Remember from above $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$ and the matrix $\mathbf{C} = (c_{ij}) = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$.

has the t-distribution with $n - k - 1$ degrees of freedom. Let's set the null hypothesis $\mathrm{H}_0 : \beta_i = 0$ (that is, choose $\beta_{0,i} = 0$), and the alternative hypothesis $\mathrm{H}_1 : \beta_i \neq 0$. The testing is performed in a usual way by using the t-distribution and the realized test statistic t_i, usually two-sided. Statistical programs usually print all these tests and the corresponding P-probabilities. In the example above all the test results are in the parameter estimation section:

Any null hypothesis $\mathrm{H}_0 : \beta_i = \beta_{0,i}$ could be of course tested this way.

We can also calculate the $100(1-\alpha)$ % confidence limits for β_i: $b_i \pm t_{\alpha/2}\mathrm{RMSE}\sqrt{c_{ii}}$.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	56.441111	7.994016	7.06	<.0001
Aika	-2.753	0.550955	-5.00	0.0003
Lämpötila	-0.361933	0.110191	-3.28	0.0065
Aika*Aika	0.0817333	0.012956	6.31	<.0001
Lämpötila*Lämpötila	0.0008133	0.000518	1.57	0.1425
Aika*Lämpötila	0.00314	0.001832	1.71	0.1123

There are also the estimated deviations of the parameter estimators $\mathrm{RMSE}\sqrt{c_{ii}}$ (in the column "Std Error").

We can for example test the hypothesis $\mathrm{H}_0 : \beta_2 = 0$, when the realized value for the test statistic is $t_2 = -3.28$. The corresponding P-probability is obtained from the t-distribution (with 12 degrees of freedom) and it's $P = 0.0065$. Thus, H_0 is rejected and we'll come to a conclusion that the regressor x_2 (temperature) is useful in the model. The regressors x_2^2 and x_1x_2 correspondingly aren't shown to be useful in the tests. The other regressors (including the constant term) are, however, seen to be useful..

We have to note that these tests for different parameters aren't independent, for the parameter estimates aren't (usually) independent. Thus, excluding many regressors as a result of the tests may sometimes lead to unexpected results.

The obtained model with its estimated parameters and error variances can be used to calculate the response with new regressor tuples, with which the experiments haven't been performed. Then we can either include the simulated error term or leave it out. The latter option is useful among other things when the error arises only from the measurements and doesn't exist in the modeled phenomenon. Let's take a new interesting regressor combination under consideration

$$x_1 = x_{0,1}\ , \ldots ,\ x_k = x_{0,k} \quad \text{or} \quad \mathbf{x}_0 = \begin{pmatrix} 1 \\ x_{0,1} \\ \vdots \\ x_{0,k} \end{pmatrix},$$

Note the 1 added for the constant term.

Let's then consider a case, where the *error term is excluded.* Then

the true response is

$$y_0 = \beta_0 + \sum_{i=1}^{k} \beta_i x_{0,i} = \mathbf{x}_0^\mathsf{T}\boldsymbol{\beta}$$

(a number), whereas the estimated response is

$$\hat{y}_0 = b_0 + \sum_{i=1}^{k} b_i x_{0,i} = \mathbf{x}_0^\mathsf{T}\mathbf{b}.$$

Because apparently (in the random variable interpretation)

$$\mathrm{E}(\hat{y}_0) = \mathrm{E}(b_0) + \sum_{i=1}^{k} \mathrm{E}(b_i) x_{0,i} = \beta_0 + \sum_{i=1}^{k} \beta_i x_{0,i} = y_0,$$

the obtained respond estimator is unbiased. With matrix calculation we may notice that

$$\mathrm{var}(\hat{y}_0) = \sigma^2 \mathbf{x}_0^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_0.$$

Additionally, it can be shown that the random variable

$$T_0 = \frac{\hat{y}_0 - y_0}{\mathrm{RMSE}\sqrt{\mathbf{x}_0^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_0}}$$

has the t-distribution with $n-k-1$ degrees of freedom. Thus, we obtain, in a way familiar from the above, the $100(1-\alpha)$ % confidence limits for y_0

$$\hat{y}_0 \pm t_{\alpha/2}\mathrm{RMSE}\sqrt{\mathbf{x}_0^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_0}.$$

Similarly, if *the error term is included,* then the correct respond is the random variable

Cf. the prediction interval in section 2.3.

$$Y_0 = \beta_0 + \sum_{i=1}^{k} \beta_i x_{0,i} + \epsilon_0 = \mathbf{x}_0^\mathsf{T}\boldsymbol{\beta} + \epsilon_0,$$

A capital letter is used here for clarity.

where ϵ_0 is a $\mathrm{N}(0, \sigma^2)$-distributed random variable independent of $\mathbf{b}$. Apparently, $\mathrm{E}(Y_0) = \mathbf{x}_0^\mathsf{T}\boldsymbol{\beta}$ and $\mathrm{var}(Y_0) = \sigma^2$, and furthermore

$$\mathrm{E}(\hat{y}_0 - Y_0) = \mathrm{E}(\hat{y}_0) - \mathrm{E}(Y_0) = 0$$

Like before, $\hat{y}_0 = \mathbf{x}_0^\mathsf{T}\mathbf{b}$.

and (because of the independence)

$$\mathrm{var}(\hat{y}_0 - Y_0) = \mathrm{var}(\hat{y}_0) + \mathrm{var}(Y_0) = \sigma^2 \mathbf{x}_0^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_0 + \sigma^2.$$

The random variable

$$T_0 = \frac{\hat{y}_0 - Y_0}{\mathrm{RMSE}\sqrt{1 + \mathbf{x}_0^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_0}}$$

has now the t-distribution with $n-k-1$ degrees of freedom and for y_0, the realized value of Y_0, we obtain by using it the $100(1-\alpha)$ % prediction interval

$$\hat{y}_0 - t_{\alpha/2}\mathrm{RMSE}\sqrt{1 + \mathbf{x}_0^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_0} < y_0 < \hat{y}_0 + t_{\alpha/2}\mathrm{RMSE}\sqrt{1 + \mathbf{x}_0^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_0}.$$

6.5 Choice of a Fitted Model Through Hypothesis Testing [12.6]

If the earlier presented F-test finds the model insignificant, in other words the null hypothesis $H_0 : \beta_1 = \cdots = \beta_k = 0$ can't be rejected, there's not much use for the model. On the other hand, even if the F-test would find the model to be significant, it's still not always very good for different reasons:

It would then be of the form "response = constant + deviation".

- Perhaps a good enough collection of regressors wasn't included in the model. This case is tested with the *lack-of-fit-test*. The null hypothesis H_0 is that the model is suitable, in other words it has adequately many regressors and it couldn't be significantly improved in that matter. If this null hypothesis is rejected, there is a reason to examine whether more regressors could be found for the model. The lack-of-fit-testing is usually done only if many tests are performed with the same regressor combinations. In that case, many statistical programs perform the test automatically. The lack-of-fit-test is as well based on the F-distribution and the programs print the test statistic and the realized P-probability of the test.

  It can be done also in other cases.

  In the example above, replicated tests are performed and JMP does the lack-of-fit-test:

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	3	0.9211722	0.307057	0.6716
Pure Error	9	4.1147000	0.457189	Prob > F
Total Error	12	5.0358722		0.5906
				Max RSq
				0.9889

  In the test, the P-probability obtained was 0.5906, which is so large that H_0 isn't rejected, and thus we may consider the model to have adequately many regressors.

- On the other hand, *not too many regressors should be included in the model.* An over-fitted model namely explains a part of its error, which of course can't be the purpose.

  In an extreme case, even completely!

- A method widely used to measure how much the model explains the examined phenomenon is to calculate the *coefficient of (multiple) determination*

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

  The square root of the coefficient R is often called the *multiple correlation coefficient.*

  This name arises from the fact that R is the Pearson sample correlation coefficient of the observed $y_1, \ldots, y_n$ and the predicted $\hat{y}_1, \ldots, \hat{y}_n$ responses. See section 7.5.

  A value of R^2 close to 1 tells that the model can explain a great deal of the variation of the response. This is especially important if the response is, in one way or another, related to energy or power.

  On the other hand, if the model is significant, even a small coefficient of determination (like 0.1 – 0.2) may be useful, if for example

there is a cheap method to partly remove an expensive fault. Such a case would be encountered if a lot of tests are performed. If the model explains even a little the respond, the F-test finds the model significant, even if the coefficient of determination was small.

On the other hand, if there are few experiments, the coefficient of determination can be relatively large, although the F-test finds the model insignificant. The F-test isn't very strong if there are only a few experiments and/or they aren't planned well.

- Many people prefer the *adjusted coefficient of determination* over R^2

  $$R^2_{\text{adj}} = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{n-1}{n-k-1}\frac{\text{SSE}}{\text{SST}},$$

  with which the effect of degrees of freedom is tried to be taken into account better.

  The choice between these two coefficients is somewhat a matter of opinion, statistical programs usually print both of them.

- In the example above we obtained the coefficient of determination to be $R^2 = 0.9864$, which is very good:

Summary of Fit	
RSquare	0.986408
RSquare Adj	0.980745
Root Mean Square Error	0.647809
Mean of Response	13.99556
Observations (or Sum Wgts)	18

  With a coefficient this good there is a danger of over-fitting and there maybe might be reason to exclude some regressors or increase the number of tests.

6.6 Categorical Regressors [12.8]

In above, the regressors are considered to be continuous or at least their values are numerical. *Categorical regressors* are classification variables. Their "values" or *levels* are classes (for example names, colors, or something like that), which have no numerical scale.

The categorical regressors $z_1, \ldots, z_l$ can be included in the regression model in addition to or instead of the "ordinary" continuous regressors $x_1, \ldots, x_k$ in the following manner. If the m_i levels of the regressor z_i are $\text{A}_{i,1}, \ldots, \text{A}_{i,m_i}$, then we introduce $m_i - 1$ "ordinary" regressors $z_{i,1}, \ldots, z_{i,m_i-1}$. In the data matrix the levels of z_i and the values obtained by the new regressors are connected as follows:

In fact, continuous regressors aren't necessarily needed at all.

z_i	$z_{i,1}$	$z_{i,2}$	$\cdots$	z_{i,m_i-1}
$\text{A}_{i,1}$	1	0	$\cdots$	0
$\text{A}_{i,2}$	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$
A_{i,m_i-1}	0	0	$\cdots$	1
A_{i,m_i}	0	0	$\cdots$	0

The values of the new regressors $z_{i,1}, \ldots, z_{i,m_i-1}$ are then either $= 0$ or $= 1$. The whole regression model is thus

They are *dichotomy variables.*

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \sum_{i=1}^{l} (\beta_{i,1} z_{i,1} + \cdots + \beta_{i,m_i-1} z_{i,m_i-1}) + \epsilon$$

Note the indexing of the new variables!

and it's fitted in the familiar way. The used levels of categorical regressors are of course recorded while performing tests, and they are encoded into a data matrix in the presented way.

The encoding method presented earlier is just one of the many possible. For example JMP-program uses a different encoding:

z_i	$z_{i,1}$	$z_{i,2}$	$\cdots$	z_{i,m_i-1}
$\mathrm{A}_{i,1}$	1	0	$\cdots$	0
$\mathrm{A}_{i,2}$	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$
A_{i,m_i-1}	0	0	$\cdots$	1
A_{i,m_i}	-1	-1	$\cdots$	-1

This can be seen from the estimated parameters.

Example. *Here the response y is the number of particles after cleaning. In the model there are included one continuous regressor x_1, the pH of the system, and one three-leveled categorical regressor z_1, the used polymer (P_1, P_2 or P_3). The model is* **[12.9]**

The encoding used here is

$$y = \beta_0 + \beta_1 x_1 + \beta_{1,1} z_{1,1} + \beta_{1,2} z_{1,2} + \epsilon.$$

z_1	$z_{1,1}$	$z_{1,2}$
P_1	1	0
P_2	0	1
P_3	0	0

$n = 18$ tests were performed, six for each level of z_1. Estimation gives then the values

$$b_0 = -161.8973\ ,\ b_1 = 54.2940\ ,\ b_{1,1} = 89.9981\ ,\ b_{1,2} = 27.1657,$$

to the parameters, from which it can be concluded that the polymer P_1 has the greatest effect and the polymer P_3 the second greatest. The obtained estimate for error variance is MSE $= 362.7652$. *The F-test (with 3 and 14 degrees of freedom) gives the P-probability, which is nearly zero, thus, the model is very significant. The coefficient of determination is $R^2 = 0.9404$, which is very good. The P-probabilities of the t-tests of parameter estimates (with 14 degrees of freedom) are small and all the regressors are necessary in the model:*

Because of the encoding, the level of polymer P_3 is a reference level.

$$0.0007 \quad , \quad \cong 0 \quad , \quad \cong 0 \quad , \quad 0.0271.$$

The data is input into the JMP program in the form

Rows	pH	Polymer	Response
1	6.5	P1	292
2	6.9	P1	329
3	7.8	P1	352
4	8.4	P1	378
5	8.8	P1	392
6	9.2	P1	410
7	6.7	P2	198
8	6.9	P2	227
9	7.5	P2	277
10	7.9	P2	297
11	8.7	P2	364
12	9.2	P2	375
13	6.5	P3	167
14	7	P3	225
15	7.2	P3	247
16	7.6	P3	268
17	8.7	P3	288
18	9.2	P3	342

The encoding that JMP uses here is

z_1	$z_{1,1}$	$z_{1,2}$
P_1	1	0
P_2	0	1
P_3	-1	-1

The encoding in JMP is different, as it was noted. On the other hand, the user need not do the encoding, for the program makes the encoding automatically after obtaining the information about the types of variables. The obtained (a bit trimmed) printout is

Summary of Fit

RSquare	0.940433
RSquare Adj	0.927669
Root Mean Square Error	19.0464
Mean of Response	301.5556
Observations (or Sum Wgts)	18

There are no replications, so the lack-of-fit-test isn't printed.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	80181.731	26727.2	73.6764
Error	14	5078.713	362.8	Prob > F
C. Total	17	85260.444		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-122.8427	37.44157	-3.28	0.0055
pH	54.294026	4.755411	11.42	<.0001
Polymer[P1]	50.943475	6.372994	7.99	<.0001
Polymer[P2]	-11.88889	6.348799	-1.87	0.0822

The parameter estimates are now

$$b_0 = -122.8427\ ,\ b_1 = 54.2940\ ,\ b_{1,1} = 50.9435\ ,\ b_{1,2} = -11.8889.$$

The comparing between different polymers can be done in that case as well. This doesn't affect on the F-test or the coefficient of determination or the MSE*-value. Instead, the t-tests change, their P-probabilities are now*

$$0.0055 \quad , \quad \cong 0 \quad , \quad \cong 0 \quad , \quad 0.0822.$$

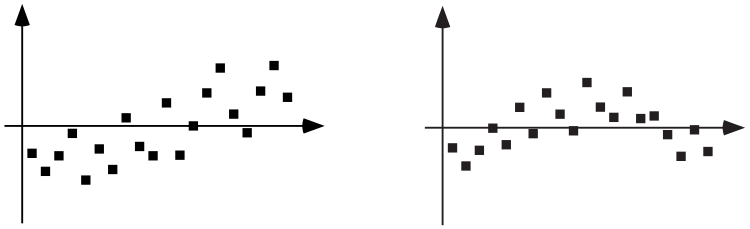
There might be some product-form interaction terms between the new regressors obtained from the categorical regressors, as well between them and the ”old” regressors, or some other calculated regressors.

6.7 Study of Residuals [12.10]

By using the residual, there are many ways to study after the model-fitting the goodness of the model or if the assumptions used when formulating the model are true. Clearly exceptional or failed experimental situations turn up as residuals with large absolute values, outliers.

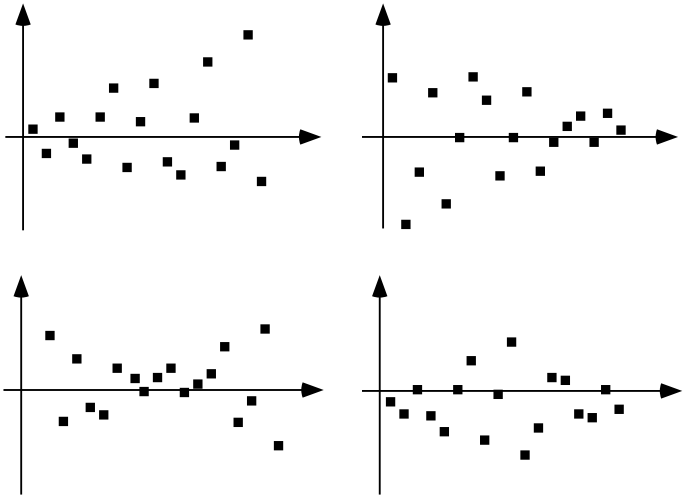
Cf. the example in section 1.3.

The most simple way is to plot the residuals for example as a function of the predicted response, in other words the points $(\hat{y}_i, e_i)$ $(i = 1, \ldots, n)$. If the obtained point plot is somewhat ”curved”, then there is clearly an unexplained part in the response and more regressors are needed:



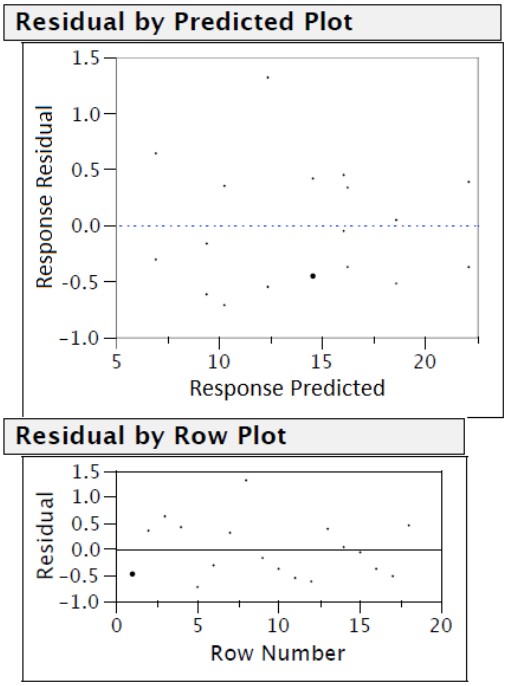
If again, the plot is somewhat ”necked” or ”bulged” or ”wedge-shaped”, then the assumption concerning the similarity of the distribution of the error term concerning variance isn't true, and a bigger change is required in modeling:

heteroscedasticity



The realized residuals can also be plotted as a function of order of experiments, in other words the points (i, e_i) $(i = 1, \ldots, n)$, and examine the plot similarly as before.

In the example in section 6.2 the residual vs. the predicted respond is quite usual (the upper plot), as well is the residual vs. the order of tests (the lower picture):



Here one of the residuals is exceptionally large, maybe it's an outlier?

There is some suspicious regularity here.

6.8 Logistical Regression [12.12]

In above, the response y has always been continuous. *Logistical Regression* allows a multileveled categorical response. The model doesn't then predict the response to the given regressor values, but gives the probabilities of the different alternatives. Let's begin with a case, where the respond is two-leveled or a *binary response*. Let's denote the two different levels of response by A and B and the probability of A by p (, which depends on the values of the regressors).

Accordingly to its name, the logistical regression uses a *logistical distribution*, whose cumulative distribution function is

$$F(z) = \frac{1}{1+e^{-z}}.$$

The idea is that the parameters $\beta_0, \beta_1, \ldots, \beta_k$ of the formula

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

A *logit*.

are estimated so that the probability obtained from the logistical distribution

$$F(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) = \frac{1}{1+e^{-\beta_0-\beta_1 x_1-\cdots-\beta_k x_k}}$$

is the probability p of the level A of the respond y for the used regressor combination.

Experiments are performed (n of them) for different regressor combinations (data matrix $\mathbf{X}$) and the obtained responses $y_1, \ldots, y_n$ (levels A and B) are recorded. The pooled probability of the realized levels is then, because of the independence of the to experiments, the product

$$L(\beta_0, \ldots, \beta_k) = L_1(\beta_0, \ldots, \beta_k) \cdots L_n(\beta_0, \ldots, \beta_k),$$

where

$$L_i(\beta_0, \ldots, \beta_k) = \begin{cases} p_i = \dfrac{1}{1 + e^{-\beta_0 - \beta_1 x_{i,1} - \cdots - \beta_k x_{i,k}}}, \text{ if } y_i = \text{A} \\ 1 - p_i = \dfrac{e^{-\beta_0 - \beta_1 x_{i,1} - \cdots - \beta_k x_{i,k}}}{1 + e^{-\beta_0 - \beta_1 x_{i,1} - \cdots - \beta_k x_{i,k}}}, \text{ if } y_i = \text{B} \end{cases} \qquad (i = 1, \ldots, n).$$

As it can be noted already from the notation, the maximum likelihood estimate is going to be used and $L(\beta_0, \ldots, \beta_k)$ is the likelihood function. The estimates of the parameter values $b_0, b_1, \ldots, b_k$ are chosen so that $L(\beta_0, \ldots, \beta_k)$ or the corresponding loglikelihood function

See chapter 5.

Other estimation methods than MLE can be used and the results may then sometimes be different.

$$l(\beta_0, \ldots, \beta_k) = \ln L(\beta_0, \ldots, \beta_k)$$

obtains its largest value when $\beta_0 = b_0, \beta_1 = b_1, \ldots, \beta_k = b_k$. By setting the partial derivatives equal to zero we obtain a system of equations, whose solution usually requires a lot of numerical computation. The number of tests performed is usually large as well. Statistical programs are needed then, and there are also web-calculators for the most simple cases.

As a result of estimation we obtain the probability $\hat{p}_0$ for A to happen when regressors have the values $x_1 = x_{0,1}, \ldots, x_k = x_{0,k}$:

$$\hat{p}_0 = \frac{1}{1 + e^{-b_0 - b_1 x_{0,1} - \cdots - b_k x_{0,k}}}.$$

The data obtained from the tests is often given in the following form. If there are l pcs. of different regressor combinations (that is, different rows of $\mathbf{X}$), then the number of tests performed $n_1, \ldots, n_l$ are given to each combination and the numbers $v_1, \ldots, v_l$ of the realized response values A as well (or the realized numbers of both realized response values).

Example. *Here the effect of the level of a certain toxin x_1 on insects is being studied. In the test, the numbers of all insects and died insects are recorded for each tested level of toxin. The results are the following:* **[12.15]**

Test	*Level of toxin x_1*	*Number of all insects*	*Number of died insects*
1	0.10	47	8
2	0.15	53	14
3	0.20	55	24
4	0.30	52	32
5	0.50	46	38
6	0.70	54	50
7	0.95	52	50

Statistical programs (i.a. JMP) can usually handle the data in this form, certain variables just have to be marked as frequency variables. The JMP-print is

In fact, this would become a data matrix with $n = 359$ rows.

Nominal Logistic Fit for Died

Iteration History

Iter	LogLikelihood	Step	Delta-Criterion	Obj-Criterion
1	-248.8398378	Initial	3700555860	.
2	-180.2962958	Newton	0.45434522	0.38015057
3	-172.2325127	Newton	0.14444273	0.04681645
4	-171.3239135	Newton	0.0215696	0.00530309
5	-171.3046844	Newton	0.00052041	0.00011224
6	-171.3046733	Newton	3.082e-7	6.493e-8

Freq: Lkm

The progress of the numerical solution of the system of equations with Newton's method can be seen here.

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	70.06115	1	140.1223	<.0001
Full	171.30467			
Reduced	241.36582			

RSquare (U)	0.2903
Observations (or Sum Wgts)	359

Converged by Gradient

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	5	2.94976	5.899523
Saturated	6	168.35491	Prob>ChiSq
Fitted	1	171.30467	0.3161

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	1.73610651	0.2420424	51.45	<.0001
Toxin	-6.2953873	0.7422285	71.94	<.0001

For log odds of E/K

The estimated parameters are

$$b_0 = -1.7361 \quad ja \quad b_1 = 6.2954$$

(JPM gives these with opposite signs). The probability of an insect to die $\hat{p}_0$ for the given level $x_1 = x_{0,1}$ is obtained (estimated) from the formula

There's $p = \dfrac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}$ in the JMP-model.

$$\hat{p}_0 = \frac{1}{1 + e^{1.7361 - 6.2954 x_{0,1}}}.$$

The significance of the estimated model can be tested with an approximative χ^2-test, the *likelihood-ratio test.* The significance of the estimated parameters, in particular is tested often with the *Wald's χ^2-test.* In the preceding example the χ^2 test statistic of the estimated model given by the significance test is 140.1223 (with 1 degree of freedom), for which the corresponding P-probability is very nearly $= 0$. Thus, the model is very significant. The parameter testing with the Wald's χ^2-test additionally shows that both of them are very significant.

Abraham Wald (1902–1950)

$P \cong 10^{-32}$

An interesting quantity is often the *odds ratio* of the response level A

The logarithm of the odds ratio is the above mentioned logit.

$$\frac{p}{1-p},$$

which is predicted to be $e^{b_0 + b_1 x_{0,1} + \cdots + b_k x_{0,k}}$.

multinomial logistical regression

A multileveled response is considered similarly. If the levels of a response are $\mathrm{A}_1, \ldots, \mathrm{A}_m$, then the corresponding probabilities are obtained from the parameters in the following way:

$$\mathrm{P}(y = \mathrm{A}_1) = \frac{1}{1 + \sum_{j=2}^{m} e^{-\beta_0^{(j)} - \beta_1^{(j)} x_1 - \cdots - \beta_k^{(j)} x_k}} \quad \text{and}$$

$$\mathrm{P}(y = \mathrm{A}_h) = \frac{e^{-\beta_0^{(h)} - \beta_1^{(h)} x_1 - \cdots - \beta_k^{(h)} x_k}}{1 + \sum_{j=2}^{m} e^{-\beta_0^{(j)} - \beta_1^{(j)} x_1 - \cdots - \beta_k^{(j)} x_k}} \quad (h = 2, \ldots, m).$$

There are in total $(m-1)(k+1)$ parameters $\beta_i^{(j)}$. The estimation is customarily done with the maximum likelihood estimation method by forming a likelihood function as a product of these probabilities.

This idea has many variants. Instead of the logistical distribution other distributions can be used as well, for example the standard normal distribution. Furthermore, logistical models may include categorical regressors (when encoded properly), interaction terms and so on.

A probit model.

Chapter 7

NONPARAMETRIC STATISTICS

Nonparametric tests are tests that don't assume a certain form of the population distributions and are focused on the probabilities concerning the distribution. Because the (approximative) normality required by the t-tests isn't always true or provable, it's recommendable to use the corresponding nonparametric tests instead. Please however note that these tests measure slightly different quantities.

Such methods were already the χ^2-tests considered in chapter 4.

7.1 Sign Test [16.1]

By a *sign test*, the quantiles $q(f)$ of a continuous distribution are being tested. Recall that if X is the corresponding random variable, then $q(f)$ is a number such that $\mathrm{P}\big(X \leq q(f)\big) = f$, in other words the population cumulation in the quantile $q(f)$ is f. The null hypothesis is then of the form

See section 1.3.

$$\mathrm{H}_0 : q(f_0) = q_0,$$

where f_0 and q_0 are given values. The alternative hypothesis is then one of the three following:

$$\mathrm{H}_1 : q(f_0) < q_0 \quad , \quad \mathrm{H}_1 : q(f_0) > q_0 \quad \text{or} \quad \mathrm{H}_1 : q(f_0) \neq q_0.$$

Let's denote by f a value such that exactly $q(f) = q_0$. The null hypothesis can then be written in the form $\mathrm{H}_0 : f = f_0$ and the above mentioned alternative hypotheses are correspondingly of the form

$$\mathrm{H}_1 : f_0 < f \quad , \quad \mathrm{H}_1 : f_0 > f \quad \text{or} \quad \mathrm{H}_1 : f_0 \neq f.$$

In order to test hypothesis, let's take a random sample $x_1, \ldots, x_n$. Let's form a corresponding sign sequence $s_1, \ldots, s_n$, where

$$s_i = \operatorname{sign}(x_i) = \begin{cases} +, \text{ if } x_i > q_0 \\ 0\,, \text{ if } x_i = q_0 \\ -, \text{ if } x_i < q_0. \end{cases}$$

Because the sample data is often, in one way or another, rounded, let's leave elements x_i, for which $s_i = 0$, outside the sample and continue

with the rest of them. After that, s_i is always either + or −. Let's now denote the sample size by n. When considered to be random variables, the sample is $X_1, \ldots, X_n$ and the signs are $S_1, \ldots, S_n$. The number of minus signs Y has then, if H_0 is true, the binomial distribution $\mathrm{Bin}(n, f_0)$ and the testing of the null hypothesis can be done similarly as in section 3.4.

Theoretically, the probability, that exactly $X_i = q_0$, is zero.

There are also web-calculators, but they mostly test only the median.

Example. *The recharging time (in hours) of a battery-powered hedge trimmer was studied. The sample consists of* 11 *times:* [**16.1**]

$$1.5\ ,\ 2.2\ ,\ 0.9\ ,\ 1.3\ ,\ 2.0\ ,\ 1.6\ ,\ 1.8\ ,\ 1.5\ ,\ 2.0\ ,\ 1.2\ ,\ 1.7.$$

The distribution of recharging time is unknown, except that it's continuous. We want to test, could the median of recharging time be $q_0 = 1.8$ *h. The hypothesis pair to be tested is then* $\mathrm{H}_0 : q(0.5) = 1.8$ *h vs.* $\mathrm{H}_1 : q(0.5) \neq 1.8$ *h, in other words* $\mathrm{H}_0 : f = 0.5$ *vs.* $\mathrm{H}_1 : f \neq 0.5$*, where* $q(f) = 1.8$ *h (and* $f_0 = 0.5$*).*

Because one of the realized sample elements is exactly 1.8 *h, it's left out and we continue with the remaining* $n = 10$ *elements. The sign sequence* $s_1, \ldots, s_{10}$ *is now*

$$-\ ,\ +\ ,\ -\ ,\ -\ ,\ +\ ,\ -\ ,\ -\ ,\ +\ ,\ -\ ,\ -.$$

The realized number of the minus signs is $y = 7$*. The P-probability of the binomial distribution test is the smaller of the numbers*

$$\sum_{i=0}^{7} \binom{10}{i} 0.5^i (1-0.5)^{10-i} \quad \textit{and} \quad \sum_{i=7}^{10} \binom{10}{i} 0.5^i (1-0.5)^{10-i}$$

(it's the latter) multiplied by two, that is $P = 0.3438$*. The null hypothesis isn't rejected in this case. The calculations on MATLAB:*

```
>> X=[1.5,2.2,0.9,1.3,2.0,1.6,1.8,1.5,2.0,1.2,1.7];

>> P=signtest(X,1.8)

P =
    0.3438
```

Example. 16 *drivers tested two different types of tires R and B. The gasoline consumptions, in kilometers per liter, were measured for each car and the results were:* [**16.2**]

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
R	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0	7.4	4.9	6.1	5.2	5.7	6.9	6.8	4.9
B	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.9	6.0	4.9	5.3	6.5	7.1	4.8
s_i	+	−	+	+	−	+	0	+	+	0	+	+	+	+	−	+

The sign sequence calculated from the difference of the consumptions is included. In two cases the results were equal and these are left out. Then there are $n = 14$ *cases and the realized number of minus signs is* $y = 3$*. Thus, the population consists of the differences of gasoline consumption. The null hypothesis is* $\mathrm{H}_0 : q(0.5) = 0$*, in other words that the median*

difference of consumption is $= 0$*, and the alternative hypothesis is* $\mathrm{H}_1 : q(0.5) > 0$*. In other words, the hypothesis pair* $\mathrm{H}_0 : f = 0.5$ *vs.* $\mathrm{H}_1 : f < 0.5$*, where* $q(f) = 0$ *(and* $f_0 = 0.5$*), is being tested with the binomial test. The obtained P-probability of the test is the tail probability of the binomial distribution*

$$\sum_{i=0}^{3} \binom{14}{i} 0.5^i (1-0.5)^{14-i} = 0.0287.$$

At the risk $\alpha = 0.05$*, the null hypothesis has to be rejected then, and conclude that when considering the median of the differences of consumptions, the tire type R is better. The calculations on MATLAB:*

```
>> D=[4.2 4.7 6.6 7.0 6.7 4.5 5.7 6.0 7.4 4.9 6.1 5.2 5.7 6.9 6.8 4.9;
      4.1 4.9 6.2 6.9 6.8 4.4 5.7 5.8 6.9 4.9 6.0 4.9 5.3 6.5 7.1 4.8];

>> P=signtest(D(1,:),D(2,:))

P =
    0.0574

>> P/2

ans =
    0.0287
```

7.2 Signed-Rank Test [16.2]

If we confine ourselves to certain kinds of distributions and certain quantiles, we may perform stronger tests. One of such tests is the *(Wilcoxon) signed-rank test.* There, in addition to the assumption that the population distribution is continuous, the population distribution is assumed to be symmetric as well. Furthermore, we can only test the median.

Frank Wilcoxon (1892–1965), a pioneer in nonparametric statistics

In the following, let's denote the median of the population distribution by $\tilde{\mu}$. By the above mentioned symmetry it's meant that the population density function f fulfills the condition $f(\tilde{\mu} + x) = f(\tilde{\mu} - x)$. The null hypothesis is $\mathrm{H}_0 : \tilde{\mu} = \tilde{\mu}_0$, where $\tilde{\mu}_0$ is a given value. If the obtained sample is $x_1, \ldots, x_n$, we proceed as follows:

1. Let's subtract $\tilde{\mu}_0$ from the sample elements and obtain the numbers

$$d_i = x_i - \tilde{\mu}_0 \quad (i = 1, \ldots, n).$$

   If some $d_i = 0$, the sample value x_i is left out.

2. Let's order the numbers $d_1, \ldots, d_n$ in *increasing order due to their absolute values* and give each number k_i a corresponding sequence number. If there are numbers equal by their absolute values in the list, their sequence number will be the mean of the original consecutive sequence numbers. If for example exactly four of the numbers $d_0, \ldots, d_n$ have a certain same absolute value and their original sequence numbers are $6, 7, 8$ and 9, the sequence number $(6 + 7 + 8 + 9)/4 = 7.5$ is given to them all.

3. Let's calculate the sum of the sequence numbers of all the positive numbers d_i. Thus we obtain the value w_+. Similarly, let's calculate the sum of the sequence numbers of all the negative numbers d_i, and we obtain the value w_-.

4. Let's denote $w = \min(w_+, w_-)$.

In the random variable consideration we would correspondingly obtain W_+, W_- and W.

In testing, the different alternatives are:

- If actually $\tilde{\mu} < \tilde{\mu}_0$, w_+ tends to be small and w_- large. This case then leads to rejecting H_0 in favor of the alternative hypothesis $\mathrm{H}_1 : \tilde{\mu} < \tilde{\mu}_0$.

- Similarly, if actually $\tilde{\mu} > \tilde{\mu}_0$, w_+ tends to be large and w_- small and H_0 is rejected in favor of the alternative hypothesis $\mathrm{H}_1 : \tilde{\mu} > \tilde{\mu}_0$.

- Furthermore, if either of the values w_+ and w_- is small, when w is small, it suggests that $\tilde{\mu} \neq \tilde{\mu}_0$ and H_0 should be rejected in favor of the alternative hypothesis $\mathrm{H}_1 : \tilde{\mu} \neq \tilde{\mu}_0$.

It's laborious to calculate the exact critical values for different risk levels (when H_0 is true) and they are even nowadays often read from tables. For large values of n however, the distribution(s) of W_+ (and W_-) are nearly normal, in other words

There are web-calculators for this test. Note however that different programs announce the signed-rank sum a bit differently.

$$W_+ \approx \mathrm{N}\Big(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\Big).$$

Because of symmetry reasons, it's probably quite clear that $\mathrm{E}(W_+) = n(n+1)/4$, for the sum of all sequence numbers is as a sum of an arithmetic series $1 + 2 + \cdots + n = n(n+1)/2$. The variance is more difficult to work out.

Example. *Let's return to the previous example concerning recharging time, but let's now do it using the signed-rank test. The obtained numbers d_i and their sequence numbers r_i are*

[**16.3**]

Now we have to assume that the distribution is symmetric.

i	1	2	3	4	5	6	7	8	9	10
x_i	1.5	2.2	0.9	1.3	2.0	1.6	1.5	2.0	1.2	1.7
d_i	−0.3	0.4	−0.9	−0.5	0.2	−0.2	−0.3	0.2	−0.6	−0.1
r_i	5.5	7	10	8	3	3	5.5	3	9	1

By summing from these, we obtain the realized values $w_+ = 13$ and $w_- = 42$, so $w = 13$. The corresponding P-probability is $P = 0.1562$ (MATLAB) and thus, null hypothesis isn't rejected in this case either. The print of JMP is:

MATLAB-command
`P=signrank(X,1.8)`

Distributions

Time

Test Mean=value

Hypothesized Value	1.8
Actual Estimate	1.60909
df	10
Std Dev	0.38589

	t Test	Signed-Rank
Test Statistic	-1.6408	-14.500
Prob > |t|	0.1319	0.156
Prob > t	0.9341	0.922
Prob < t	0.0659	0.078

The t-test result is here similar to the signed-rank test result.

Example. *Certain psychology test results are being compared. We want to know if the result is better when the test subject is allowed to practice beforehand with similar exercises, or not. In order to study the matter,* $n = 10$ *pairs of test subjects were chosen, and one of the pair was given a few similar exercises, the other wasn't. The following results (scores) were obtained:* [**16.4**]

i	1	2	3	4	5	6	7	8	9	10
Training	531	621	663	579	451	660	591	719	543	575
No training	509	540	688	502	424	683	568	748	530	524

According to the chosen null hypothesis H_0*, the median of the differences is* $\tilde{\mu}_0 = 50$*. The alternative hypothesis* H_1 *is chosen to be the claim that the median is* < 50*. This is a one-sided test we consider here then. For the test, let's calculate the table*

Note that the medians of the scores aren't tested here! Usually the median of difference isn't the same as the difference of medians.

i	1	2	3	4	5	6	7	8	9	10
d_i	22	81	-25	77	27	-23	23	-29	13	51
$d_i - \tilde{\mu}_0$	-28	31	-75	27	-23	-73	-27	-79	-37	1
r_i	5	6	9	3.5	2	8	3.5	10	7	1

from which we can see that $w_+ = 10.5$*. The corresponding P-probability is* $P = 0.0449$ *(MATLAB). Thus,* H_0 *can be rejected at the risk* $\alpha = 0.05$ *and it can be concluded that practicing beforehand doesn't increase the test result by (at least)* 50*, when concerning the median of the differences. The print of JMP is:*

MATLAB-command

```
P=signrank(D(1,:)-50,
D(2,:))/2
```

Matched Pairs

Difference: Training - 50 - No_training

Training - 50	543.3	t-Ratio	-2.24606
No_training	571.6	DF	9
Mean Difference	-28.3	Prob > |t|	0.0513
Std Error	12.5999	Prob > t	0.9743
Upper95%	0.20288	Prob < t	0.0257
Lower95%	-56.803		
N	10		
Correlation	0.93713		

Wilcoxon Sign-Rank

	Training - 50 - No_training
Test Statistic	-17.000
Prob > |z|	0.090
Prob > z	0.955
Prob < z	0.045

Here, the t-test result is somewhat different to the signed-rank test result.

7.3 Mann–Whitney test [16.3]

Henry Mann (1905–2000)
Ransom Whitney (1915–2001)

The *Mann–Whitney test* compares the medians of two continuous population distributions. The test is called also *U-test* or *(Wilcoxon) rank-sum test* or just *Wilcoxon test.* Let's denote the population medians by $\tilde{\mu}_1$ and $\tilde{\mu}_2$. The null hypothesis is then $\mathrm{H}_0 : \tilde{\mu}_1 = \tilde{\mu}_2$. Actually the null hypothesis is that the population distributions are the same—when they of course have the same median—because with this assumption the critical limits etc. are calculated.

Thus, the test doesn't finally solve the Behrens–Fisher-problem, although it's often claimed do so.

The Mann–Whitney test reacts sensitively to the difference of the population medians, but much more weakly to many other differences in population distributions. For this reason, it's not quite suitable to test the similarity of two populations, although it's often recommended. Many people think that the test has to be considered a location test, when the distributions, according to the hypotheses H_0 and H_1, are of the same form, only in different locations.

In order to perform the test, let's take two samples from a population

$$x_{1,1}, \ldots, x_{1,n_1} \quad \text{and} \quad x_{2,1}, \ldots, x_{2,n_2}.$$

Let the sample size n_1 be the smaller one. Let's now proceed as follows:

If they are unequal—only to make the calculations easier.

1. Let's combine the samples as a pooled sample

$$x_{1,1}, \ldots, x_{1,n_1}, x_{2,1}, \ldots, x_{2,n_2}.$$

2. Let's order the elements in the pooled sample in increasing order and give them the corresponding sequence numbers

$$r_{1,1}, \ldots, r_{1,n_1}, r_{2,1}, \ldots, r_{2,n_2}.$$

   If there are duplicate numbers in the pooled sample, when their sequence numbers are consecutive, let's give all those numbers a sequence number, which is the mean of the original consecutive sequence numbers. If for example exactly three elements of the pooled sample have a certain same value and their original sequence numbers are $6, 7$ and 8, let's then give to all of them the sequence number $(6 + 7 + 8)/3 = 7$.

3. Let's sum the n_1 sequence numbers of the first sample. Thus we obtain the value $w_1 = r_{1,1} + \cdots + r_{1,n_1}$.

4. Correspondingly, by summing the n_2 sequence numbers of the second sample we obtain the value $w_2 = r_{2,1} + \cdots + r_{2,n_2}$. Note that as a sum of an arithmetic series we have

$$w_1 + w_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2},$$

   from which w_2 can easily be calculated, when w_1 is already obtained.

5. Let's denote $w = \min(w_1, w_2)$.

In random variable consideration we would obtain the corresponding random variables W_1, W_2 and W. Often, instead of these, values

$$u_1 = w_1 - \frac{n_1(n_1+1)}{2} \quad , \quad u_2 = w_2 - \frac{n_2(n_2+1)}{2} \quad \text{and} \quad u = \min(u_1, u_2),$$

are used and the corresponding random variables are U_1, U_2 and U.

The name "U-test" arises from here.

In testing, the following cases may occur:

- If actually $\tilde{\mu}_1 < \tilde{\mu}_2$, w_1 tends to be small and w_2 large. This case often leads to rejecting H_0 in favor of the alternative hypothesis $\mathrm{H}_1 : \tilde{\mu}_1 < \tilde{\mu}_2$.
- Similarly, if actually $\tilde{\mu}_1 > \tilde{\mu}_2$, w_1 tends to be large and w_2 small and H_0 is rejected in favor of the alternative hypothesis $\mathrm{H}_1 : \tilde{\mu}_1 > \tilde{\mu}_2$.
- Furthermore, if either of the values w_1 and w_2 is small, when w is small, it suggests that $\tilde{\mu}_1 \neq \tilde{\mu}_2$ and H_0 should be rejected in favor of the alternative hypothesis $\mathrm{H}_1 : \tilde{\mu}_1 \neq \tilde{\mu}_2$.

In a similar way, the values u_1, u_2 and u could be used in the test.

It's laborious to calculate the exact values for different risk probabilities (when H_0 is true) and they are even nowadays often read from tables. For large values of n_1 and n_2 the distribution(s) of W_1 (and W_2) are nearly normal, in other words

$$W_1 \approx \mathrm{N}\Big(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\Big).$$

There are web-calculators for this test as well.

Example. *The nicotine contents of two brands of cigarettes A and B were measured (mg). The hypothesis pair to be tested is* $\mathrm{H}_0 : \tilde{\mu}_\mathrm{A} = \tilde{\mu}_\mathrm{B}$ *vs.* $\mathrm{H}_1 : \tilde{\mu}_\mathrm{A} \neq \tilde{\mu}_\mathrm{B}$. *The following results were obtained, also the sequence numbers of the pooled sample are included:* **[16.5]**

i	1	2	3	4	5	6	7	8	9	10
$x_{\mathrm{A},i}$	2.1	4.0	6.3	5.4	4.8	3.7	6.1	3.3	–	–
$r_{\mathrm{A},i}$	4	10.5	18	14.5	13	9	16	8	–	–
$x_{\mathrm{B},i}$	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4
$r_{\mathrm{B},i}$	12	1	7	6	10.5	17	2	5	3	14.5

The sample sizes are $n_\mathrm{A} = 8$ *and* $n_\mathrm{B} = 10$. *By calculating we obtain* $w_\mathrm{A} = 93$ *and* $w_\mathrm{B} = 78$, *so* $w = 78$. *(Similarly we would obtain* $u_\mathrm{A} = 57$, $u_\mathrm{B} = 23$ *and* $u = 23$.*) From this, the obtained P-probability is* $P = 0.1392$ *(MATLAB) and there is no reason to reject* H_0. *The print of JMP is:*

MATLAB-command
`P=ranksum(X_A,X_B)`

Oneway Analysis of Nicotine By Brand

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Score Mean	(Mean-Mean0)/Std0
A	8	93	11.6250	1.468
B	10	78	7.8000	-1.468

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
93	1.46758	0.1422

These are approximations.

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
2.2863	1	0.1305

7.4 Kruskal–Wallis test [16.4]

William Kruskal (1919–2005), Allen Wallis (1912–1998)

The *Kruskal–Wallis test* is a generalization of the Mann–Whitney test for the case, where there can be more than two populations to be compared. Let's denote the medians of the populations (k of them) similarly as before by $\tilde{\mu}_1, \ldots, \tilde{\mu}_k$. Like the Mann–Whitney test, the Kruskal–Wallis test compares population distributions according to their medians yet when calculating critical values, it's assumed that the population distributions are the same. The essential null hypothesis is

$$\mathrm{H}_0 : \tilde{\mu}_1 = \cdots = \tilde{\mu}_k.$$

In order to perform the test, let's take a sample from each of the populations. These samples are then combined as a pooled sample and its elements are ordered in increasing order, just like in the Mann–Whitney test. Especially, duplicate values are handled similarly as before. By calculating the sums of sequence numbers of the elements of each population, we obtain the rank sums $w_1, \ldots, w_k$ and the corresponding random variables $W_1, \ldots, W_k$. Let's denote the sample size of the j:th population by n_j and $n = n_1 + \cdots + n_k$.

It's very laborious to calculate the exact critical value of the test, at least for greater values of k. The test is usually performed with the information that (when H_0 is true) the random variable

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{W_j^2}{n_j} - 3(n+1)$$

is approximately χ^2-distributed with $k-1$ degrees of freedom. This approximation can also be used in the Mann–Whitney test (where $k = 2$) The (approximative) P-probability of the test corresponding the realized value of H

JMP did this in the previous example.

$$h = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{w_j^2}{n_j} - 3(n+1)$$

is then obtained from the tail probability of the χ^2-distribution (with $k-1$ degrees of freedom that is). Again, there are web-calculators for this test, at least for smaller values of k.

Example. *The propellant burning rates of three different types of missiles A, B and C were studied. The results (coded) are presented below, there are also the sequence numbers included:* [16.6]

i	1	2	3	4	5	6	7	8	w
$x_{\mathrm{A},i}$	24.0	16.7	22.8	19.8	18.9	–	–	–	
$r_{\mathrm{A},i}$	19	1	17	14.5	9.5	–	–	–	61
$x_{\mathrm{B},i}$	23.2	19.8	18.1	17.6	20.2	17.8	–	–	
$r_{\mathrm{B},i}$	18	14.5	6	4	16	5	–	–	63.5
$x_{\mathrm{C},i}$	18.4	19.1	17.3	17.3	19.7	18.9	18.8	19.3	
$r_{\mathrm{C},i}$	7	11	2.5	2.5	13	9.5	8	12	65.5

Here the calculated test statistic is $h = 1.6586$ *and the corresponding P-probability obtained from the* χ^2 *distribution (with* 2 *degrees of freedom) and* H_0 *isn't rejected. Thus, the missile types are similar in propellant burning rates when measuring with medians. The print of JMP is:*

Oneway Analysis of Burning Rate By Type

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Score Mean	(Mean-Mean0)/Std0
A	5	61	12.2000	0.973
B	6	63.5	10.5833	0.263
C	8	65.5	8.1875	-1.158

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
1.6630	2	0.4354

Note the slight difference compared to the previous one! JMP calculates a *fixed test variable*. It's advantageous if there are many duplicate values.

The calculations with MATLAB are:

```
>> X=[24.0 16.7 22.8 19.8 18.9];
>> Y=[ 23.2 19.8 18.1 17.6 20.2 17.8];
>> Z=[18.4 19.1 17.3 17.3 19.7 18.9 18.8 19.3];
>> group=[ones(1,length(X)) 2*ones(1,length(Y)) 3*ones(1,length(Z))];
>> P=kruskalwallis([X Y Z],group)

P =
    0.4354
```

So does MATLAB!

7.5 Rank Correlation Coefficient [16.5]

If two populations are connected element by element, their relation is often represented by a value obtained from the sample, the *(Pearson) correlation coefficient* r. In order to calculate this, let's take an n-element random sample from both populations counterpart by counterpart:

$$x_{1,1}, \ldots, x_{1,n} \quad \text{and} \quad x_{2,1}, \ldots, x_{2,n}.$$

In order to calculate r, let's first calculate the *sample variance*

$$q = \frac{1}{n-1} \sum_{i=1}^{n} (x_{1,i} - \overline{x}_1)(x_{2,i} - \overline{x}_2),$$

which is an (unbiased) estimate of the population distributions' covariance. Here $\overline{x}_1$ is the sample mean of the first sample and $\overline{x}_2$ of the second. From this we obtain the mentioned sample correlation coefficient

$$r = \frac{q}{s_1 s_2},$$

where s_1^2 is the sample variance of the first sample s_2^2 of the second. This is used when studying the (linear) dependence of population distributions similarly as the actual correlation coefficient $\mathrm{corr}(X, Y)$. Also the values of r belong to the interval $[-1, 1]$.

An additional assumption is of course that $s_1, s_2 \neq 0$.

See the course Probability Calculus.

The rank correlation coefficient of two populations is a similar non-parametric quantity. For it, let's order the elements of both populations

separately in increasing order and give them sequence numbers like before:

$$r_{1,1}, \ldots, r_{1,n} \quad \text{and} \quad r_{2,1}, \ldots, r_{2,n}.$$

Possible duplicate values are handled as before. For both samples, the mean of the sequence numbers is

Cf. an arithmetic series.

$$\overline{r} = \frac{1}{n}(1 + 2 + \cdots + n) = \frac{n+1}{2}.$$

Furthermore, we obtain the sum of squares of the sequence numbers, supposing that there are no duplicate values:

$$\sum_{i=1}^{n} r_{1,i}^2 = \sum_{i=1}^{n} r_{2,i}^2 = 1^2 + 2^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1).$$

The Spearman rank correlation coefficient is then simply the sample correlation coefficient obtained from the sequence numbers, in other words

An additional assumption is that all the sequence numbers in either population are not all the same.

Charles Spearman (1863–1945)

$$r_{\mathrm{S}} = \frac{\displaystyle\sum_{i=1}^{n}(r_{1,i} - \overline{r})(r_{2,i} - \overline{r})}{\displaystyle\sqrt{\sum_{i=1}^{n}(r_{1,i} - \overline{r})^2}\,\sqrt{\sum_{i=1}^{n}(r_{2,i} - \overline{r})^2}}\,.$$

This is easier to calculate if (as it's now assumed) there are no duplicate numbers in the samples. By proceeding similarly as with the sample variances, we see that

$$\sum_{i=1}^{n}(r_{1,i} - \overline{r})(r_{2,i} - \overline{r}) = \sum_{i=1}^{n} r_{1,i}r_{2,i} - n\,\overline{r}^{\,2} = \sum_{i=1}^{n} r_{1,i}r_{2,i} - \frac{1}{4}n(n+1)^2$$

and

$$\begin{aligned}\sum_{i=1}^{n}(r_{1,i} - \overline{r})^2 &= \sum_{i=1}^{n} r_{1,i}^2 - \frac{1}{4}n(n+1)^2 = (1^2 + 2^2 + \cdots + n^2) - \frac{1}{4}n(n+1)^2 \\ &= \frac{1}{6}n(n+1)(2n+1) - \frac{1}{4}n(n+1)^2 = \frac{1}{12}n(n^2-1),\end{aligned}$$

similarly to the other sample. By using these and with a little calculation, we obtain a simpler formula for the rank correlation coefficient

$$r_{\mathrm{S}} = \frac{12}{n(n^2-1)}\sum_{i=1}^{n} r_{1,i}r_{2,i} - 3\,\frac{n+1}{n-1}.$$

The sum of squares of differences of the sequence numbers $d_i = r_{1,i} - r_{2,i}$ can be unified to the sum $\sum_{i=1}^{n} r_{1,i}r_{2,i}$ included in the formula:

$$\sum_{i=1} d_i^2 = \sum_{i=1}^{n}(r_{1,i}^2 - 2r_{1,i}r_{2,i} + r_{2,i}^2) = -2\sum_{i=1}^{n} r_{1,i}r_{2,i} + \frac{1}{3}n(n+1)(2n+1).$$

thus, with a little further calculation and by using these differences, we can formulate r_S in a simpler way:

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2.$$

This "easy" formula holds exactly only when there are no duplicate sample values.

Oddly enough, it's often used even when there are duplicate values. The result isn't necessarily very exact then.

Unlike the Pearson correlation coefficient, the Spearman correlation coefficient is able to measure also nonlinear correlation between populations, at least at some level. It can be used for *ordinal-valued* population distributions (discrete categorical distribution, whose levels can be ordered.)

Example. *In an earlier example the rank correlation coefficient of the two types of tires A and B* $r_S = 0.9638$ *is high as it should be, for the cars and drivers are the same in one test pair. Also the (Pearson) sample correlation coefficient* $r = 0.9743$ *is high. This is calculated with MATLAB as follows:*

```
>> D=[4.2 4.7 6.6 7.0 6.7 4.5 5.7 6.0 7.4 4.9 6.1 5.2 5.7 6.9 6.8 4.9;
      4.1 4.9 6.2 6.9 6.8 4.4 5.7 5.8 6.9 4.9 6.0 4.9 5.3 6.5 7.1 4.8];

>> corr(D(1,:)',D(2,:)','type','Spearman')

ans =
    0.9638

>> corr(D(1,:)',D(2,:)','type','Pearson')

ans =
    0.9743
```

Another widely used rank correlation coefficient is the *Kendall correlation coefficient.*

Chapter 8

STOCHASTIC SIMULATION

Stochastic simulation and the generation of random numbers are topics that are not considered in WMMY. In the following there is a brief overview of some basic methods.

8.1 Generating Random Numbers

Stochastic simulation is a term used to describe methods that, at one point or another, involve the use of generated random variables. These random variables may come from different distributions, but usually they are independent. The generation of random variables — especially fast and exact generation — is a challenging field of numerical analysis. The methods to be presented here are simple but not necessarily fast or precise enough for advanced applications. Practically all statistical programs, including MATLAB, have random number generators for the most common distributions. There are also web-based generators, but they aren't always suitable for solving "real" simulation problems.

8.1.1 Generating Uniform Distributions

Independent random variables uniformly distributed over the interval $[0, 1)$ are generated with methods involving number theory. In the following it is assumed that such random numbers are available. We have to note that these random number generators are completely deterministic programs that have no contingency. However, generated sequences of numbers have most of the properties of "real" random numbers

"pseudo-random numbers"

Random variables uniformly distributed over the open interval $(0, 1)$ are obtained by rejecting the generated 0-values. Samples in $[0, 1]$ can be obtained by for example rejecting all the values that are > 0.5 and by multiplying the result by two. Furthermore, if U is uniformly distributed over the interval $[0, 1)$, then $1-U$ is uniformly distributed over the interval $(0, 1]$. Thus, the type of the interval doesn't matter.

It's quite easy to obtain uniformly distributed random variables over half-open intervals other than $[0, 1)$. If namely U is uniformly distributed

over the interval $[0, 1)$, then $(b - a)U + a$ is uniformly distributed over the interval $[a, b)$. Other kinds of intervals are considered similarly.

8.1.2 Generating Discrete Distributions

Finite distributions can be easily generated. If the possible cases of a finite distribution are $T_1, \ldots, T_m$ and their probabilities are correspondingly $p_1, \ldots, p_m$ (where $p_1, \ldots, p_m > 0$ and $p_1 + \cdots + p_m = 1$), then the following procedure generates a random sample from the desired distribution:

1. Generate random number u from the uniform distribution over the interval $[0, 1)$.
2. Find an index i such that $p_0 + \cdots + p_i \leq u < p_0 + \cdots + p_{i+1}$, with the convention that $p_0 = 0$.
3. Output T_{i+1}.

This method works well in particular when generating a *discrete uniform distribution*, for which $p_1 = \cdots = p_n = 1/n$. This way we can for example take a random sample from a finite population by numbering its elements.

A *binomial distribution* $\mathrm{Bin}(p, n)$ can basically be generated as a finite distribution using the above mentioned method, but this is usually computationally too heavy. It's easier to generate n cases of a finite distribution such that the possible cases are T_1 and T_2 and $\mathrm{P}(T_1) = p$. The realization of the binomially distributed random number x is then the realization of the number of T_1 cases.

Bernoulli distribution

The *Poisson distribution* is more difficult to generate. With the parameter λ the possible values x of the Poisson-distributed random variable X are the integers $0, 1, 2, \ldots$ and

$$\mathrm{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

On way to generate the values x of X is to use the exponential distribution (whose generation will be considered later). If the random variable Y has the exponential distribution with the parameter λ, then its density function is $\lambda e^{-\lambda y}$ (when $y \geq 0$ and $= 0$ elsewhere). With a simple calculation we note that

$$\mathrm{P}(Y \leq 1) = 1 - e^{-\lambda} = 1 - \mathrm{P}(X = 0) = \mathrm{P}(X \geq 1).$$

It's more difficult to show a more general result (the proof is omitted here) that if $Y_1, \ldots, Y_k$ are independent exponentially distributed random variables (each of them with the parameter λ) and $W_k = Y_1 + \cdots + Y_k$, then

$$\mathrm{P}(W_k \leq 1) = 1 - \sum_{i=0}^{k-1} \frac{\lambda^i}{i!} e^{-\lambda} = 1 - \mathrm{P}(X \leq k - 1) = \mathrm{P}(X \geq k).$$

Thus,

$$\mathrm{P}(X = k-1) = \mathrm{P}(X \geq k-1) - \mathrm{P}(X \geq k) = \mathrm{P}(W_{k-1} \leq 1) - \mathrm{P}(W_k \leq 1).$$

From this we may conclude that the following procedure produces a random number x from the Poisson distribution with parameter λ:

1. Generate independent exponentially distributed random variables with the parameter λ until their sum is ≤ 1.
2. When the sum first time exceeds 1, look at the number k of generated exponentially distributed random variables.
3. Output $x = k - 1$.

8.1.3 Generating Continuous Distributions with the Inverse Transform Method

If the cumulative distribution function F of the continuous random variable X has an inverse F^{-1} (in a domain where its density function is $\neq 0$), then the values x of X can be generated starting from an uniform distribution. This method is attractive provided that the values of the inverse function in question can be computed quickly. This *Inverse transform method* is:

1. Generate random number u from the uniform distribution over the interval $[0, 1)$. (The corresponding random variable is U).
2. Calculate $x = F^{-1}(u)$ (i.e. $u = F(x)$ and for random variables $U = F(X)$).
3. Output x.

The procedure is based on the following observation. Being a cumulative distribution function, the function F is non-decreasing. Let G denote the cumulative distribution function of U in the interval $[0, 1)$, that is, $G(u) = u$. Then

$$\mathrm{P}(X \leq x) = \mathrm{P}\big(F(X) \leq F(x)\big) = \mathrm{P}\big(U \leq F(x)\big) = G\big(F(x)\big) = F(x).$$

The method can also be used to generate random numbers for an empirical cumulative distribution function obtained from a large sample, by linearly interpolating between the cdf values. That is, by using an *ogive*.

Let's consider as an example the *exponential distribution* that was used earlier when generating the Poisson distribution. If X has the exponential distribution with the parameter λ, then its cumulative distribution function is $F(x) = 1 - e^{-\lambda x}$ (when $x \geq 0$). The inverse function F^{-1} can be easily found: If $y = 1 - e^{-\lambda x}$, then

$$x = F^{-1}(y) = -\frac{1}{\lambda} \ln(1 - y).$$

Thus, for every random number u uniformly distributed over the interval $[0, 1)$ we obtain an exponentially distributed random number x with the parameter λ by the transformation

$$x = -\frac{1}{\lambda}\ln(1-u).$$

In order to generate a *normal distribution* $\mathrm{N}(\mu, \sigma^2)$, it's enough to generate the standard normal distribution. If namely the random variable Z has the standard normal distribution, then the random variable $X = \sigma Z + \mu$ has the $\mathrm{N}(\mu, \sigma^2)$-distribution. The cumulative distribution function of the standard normal distribution is

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\,\mathrm{d}t$$

Its inverse Φ^{-1} (the quantile function) cannot be formulated using the "familiar" functions nor it is easy to calculate numerically. The result mentioned in section 1.3,

$$\Phi^{-1}(y) = q_{0,1}(y) \cong 4.91\big(y^{0.14} - (1-y)^{0.14}\big),$$

gives some sort of approximation. A much better approximation is for example

$$\Phi^{-1}(y) \cong \begin{cases} w - v, \text{ when } 0 < y \leq 0.5 \\ v - w, \text{ when } 0.5 \leq y < 1, \end{cases}$$

where

$$w = \frac{2.515517 + 0.802853v + 0.010328v^2}{1 + 1.432788v + 0.189269v^2 + 0.001308v^3}$$

and

$$v = \sqrt{-2\ln\big(\min(y, 1-y)\big)}\ .$$

Distributions obtained from the normal distribution can be generated in the way they are obtained from the normal distribution. For example, for the χ^2-*distribution* with n degrees of freedom we can generate n independent standard normal random numbers $z_1, \ldots, z_n$ and calculate

$$v = z_1^2 + \cdots + z_n^2.$$

For the *t-distribution* with n degrees of freedom, we can generate $n+1$ independent standard normal random numbers $z_1, \ldots, z_{n+1}$ and calculate

$$t = \frac{z_{n+1}\sqrt{n}}{\sqrt{z_1^2 + \cdots + z_n^2}}.$$

For the *F-distribution* with n_1 and n_2 degrees of freedom, we can generate $n_1 + n_2$ independent standard normal random numbers $z_1, \ldots, z_{n_1+n_2}$ and calculate

$$f = \frac{z_1^2 + \cdots + z_{n_1}^2}{z_{n_1+1}^2 + \cdots + z_{n_1+n_2}^2}\,\frac{n_2}{n_1}.$$

8.1.4 Generating Continuous Distributions with the Accept–Reject Method

The *accept–reject method* can be used when generating a random number x such that the density function f of the corresponding distribution is $\neq 0$ only in a certain finite interval $[a, b]$ (not necessarily in the whole interval) and is in this interval limited by the number c. The procedure is:

1. Generate a random number u that is uniformly distributed over the interval $[a, b]$
2. Generate independently a random number v that is uniformly distributed over the interval $[0, c]$.
3. Repeat step 2, if necessary, until $v \leq f(u)$. (Recall that f is bounded above by c, that is, $f(u) \leq c$.)
4. Output $x = u$.

The method works because of the following reasons:

- The generated pairs (u, v) of random numbers are uniformly distributed over the rectangle $a \leq u \leq b$, $0 < v \leq c$.
- The algorithm retains only pairs (u, v) such that $v \leq f(x)$, and they are uniformly distributed over the region $\mathcal{A} : a \leq u \leq b$, $0 < v \leq f(u)$.
- Because f is a probability density function, the area of the region $\mathcal{A}$ is
$$\int_a^b f(u)\,\mathrm{d}u = 1,$$
so the density function of the retained pairs has the value $= 1$ inside the region $\mathcal{A}$ (and $= 0$ outside of it). (Recall that the density function f was $= 0$ outside the interval $[a, b]$.)
- The distribution of the random number u is a marginal distribution of the distribution of the pairs (u, v). The density function of u is thus obtained by integrating out the variable v, i.e.

See the course Probability Statistics.

$$\int_0^{f(u)} 1\,\mathrm{d}v = f(u).$$

- Thus, the output random number x has the correct distribution.

The accept–reject method can be used also when the domain of the density function is not a finite interval. In that case, we have to choose an interval $[a, b]$ outside of which the probability is very small.

There are also other variants of the method. A problem with the above mentioned basic version often is that the density function f of X has one or more narrow and high peaks. Then there will be many rejections in the third phase and the method is slow. This can be fixed with the following idea. Let's find a random variable U whose density function g is $= 0$ outside the interval $[a, b]$, whose values we can rapidly generate, and for which

$$f(x) \leq Mg(x)$$

for some constant M. By choosing a g that "imitates" the shape of f better than a straight horizontal line, there will be fewer rejections. The procedure itself is after this the same as before, except that the first two steps are replaced with

In the basic version above U has a uniform distribution over the interval $[a, b]$ and $M = c(b - a)$.

1'. Generate a random number u that is distributed over the interval $[a, b]$ according to the density g.

Here the interval $[a, b]$ could be an infinite interval, $(-\infty, \infty)$ for example.

2'. Generate independently a random number w that is uniformly distributed over the interval $[0, 1]$, and set $v = wMg(u)$.

The justification of the method is almost the same, the generated pairs of random numbers (u, v) are uniformly distributed over the region $a \leq u \leq b$, $0 < v \leq Mg(u)$ and so on, but the proof requires the concept of a conditional distribution, and is omitted.

The density function is $1/M$ in that region.

8.2 Resampling

Resampling refers to a whole set of methods whose purpose is, by simulation sampling, to study statistical properties of a population that would otherwise be difficult to access.

The basic idea is the following: Let's first take a comprehensive large enough sample of the population to be studied. This is done thoroughly and with adequate funding. After that, let's take a very large number of smaller samples from this base sample, treating it as a population. Because the whole base sample is saved on a computer, this can be done very rapidly. Nevertheless, resampling is usually computationally very intensive. Thus we may obtain a very large number of samples from a statistic (sample quantile, sample median, estimated proportion, sample correlation coefficient and so on) corresponding to a certain sample size. By using the samples we can actually obtain quite a good approximation for the whole distribution of the statistic in question in the original population as quite accurate empirical density and cumulative distribution functions. A more modest goal would for example be just a confidence interval for the statistic.

In many cases, the distribution of such a statistic would be impossible to derive with analytical methods.

8.3 Monte Carlo Integration

Nowadays stochastic simulation is often called Monte Carlo simulation, although the actual *Monte Carlo method* is a numerical integration method.

Let's consider a case where a function of three variables $f(x, y, z)$ should be integrated possibly over a complicated bounded three-dimensional body $\mathcal{K}$, in other words we should numerically calculate the integral

$$\int_{\mathcal{K}} f(x, y, z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$$

with a reasonable precision. Three-dimensional integration with, say, Simpson's method would be computationally very slow.

A Monte Carlo method for this problem would be the following. It's assumed that there is a fast way to determine whether or not a given point (x, y, z) lies inside the body $\mathcal{K}$ and that the body $\mathcal{K}$ lies entirely inside a given rectangle $\mathcal{P} : a_1 \le x \le a_2$, $b_1 \le y \le b_2$, $c_1 \le z \le c_2$. Let's denote the volume of $\mathcal{K}$ by V. Then the method is

1. The sample that is gathered in the method is denoted by $\mathcal{O}$. Initially it's empty.

2. Generate a random point $\mathbf{r} = (x, y, z)$ from the rectangle $\mathcal{P}$. This is simply done by generating three independent uniformly distributed random numbers x, y and z over the intervals $[a_1, a_2]$, $[b_1, b_2]$ and $[c_1, c_2]$ respectively.

3. Repeat step 2. until the point $\mathbf{r}$ lies inside the body $\mathcal{K}$. (The test for belonging to the body was supposed to be fast.)

4. Calculate $f(\mathbf{r})$ and add it to the sample $\mathcal{O}$.

5. Calculate the sample mean $\overline{x}$ of the current sample $\mathcal{O}$. If it has remained relatively unchanged (within the desired accuracy tolerance) in the past few iterations, stop and output $V\,\overline{x}$. Otherwise return to step 2. and continue.

The procedure works because after many iterations the sample mean $\overline{x}$ approximates fairly well the expectation of the random variable $f(X, Y, Z)$ when the triplet (X, Y, Z) is uniformly distributed over the body $\mathcal{K}$. The corresponding density function is then $= 1/V$ inside the body $\mathcal{K}$ (and $= 0$ outside of it), and the expectation of $f(X, Y, Z)$ is

$$\mathrm{E}\big(f(X, Y, Z)\big) = \int_{\mathcal{K}} f(x, y, z)\,\frac{1}{V}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z,$$

so by multiplying by V the desired integral is obtained.

Example. *Let's calculate the integral of the function* $f(x, y, z) = e^{x^3+y^3+2z^3}$ *over the unit sphere* $x^2 + y^2 + z^2 \le 1$. *The exact value is* 4.8418 *(Maple), the result MATLAB gives after a million iterations of Monte Carlo approximation is* 4.8429.

In fact, the volume V can also be obtained with the Monte Carlo method. This procedure is:

1. There are two counters n and l in the method. Initially $n = l = 0$.

2. Generate a random point $\mathbf{r}$ from the rectangle $\mathcal{P}$ and increment counter n by one.

3. If the point $\mathbf{r}$ lies inside the body $\mathcal{K}$, increment counter l by one.

4. If $p = l/n$ hasn't changed significantly within the last few iterations, stop and output $p \cdot (a_2 - a_1)(b_2 - b_1)(c_2 - c_1)$. Otherwise return to step 2. and continue.

Note that $(a_2 - a_1)(b_2 - b_1)(c_2 - c_1)$ is the volume of the rectangle $\mathcal{P}$.

There are many variations of this basic method, such as generalisation to higher dimensions and so on. In general, Monte Carlo integration requires a large number of iterations in order to achieve reasonable precision.

Appendix

TOLERANCE INTERVALS

The tables are calculated with the Maple program. The table gives the value to the coefficient k. First for the two-sided tolerance interval:

k:	$\gamma = 0.1$			$\gamma = 0.05$			$\gamma = 0.01$		
n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
5	3.4993	4.1424	5.3868	4.2906	5.0767	6.5977	6.6563	7.8711	10.222
6	3.1407	3.7225	4.8498	3.7325	4.4223	5.7581	5.3833	6.3656	8.2910
7	2.9129	3.4558	4.5087	3.3895	4.0196	5.2409	4.6570	5.5198	7.1907
8	2.7542	3.2699	4.2707	3.1560	3.7454	4.8892	4.1883	4.9694	6.4812
9	2.6367	3.1322	4.0945	2.9864	3.5459	4.6328	3.8596	4.5810	5.9803
10	2.5459	3.0257	3.9579	2.8563	3.3935	4.4370	3.6162	4.2952	5.6106
11	2.4734	2.9407	3.8488	2.7536	3.2727	4.2818	3.4286	4.0725	5.3243
12	2.4139	2.8706	3.7591	2.6701	3.1748	4.1555	3.2793	3.8954	5.0956
13	2.3643	2.8122	3.6841	2.6011	3.0932	4.0505	3.1557	3.7509	4.9091
14	2.3219	2.7624	3.6200	2.5424	3.0241	3.9616	3.0537	3.6310	4.7532
15	2.2855	2.7196	3.5648	2.4923	2.9648	3.8852	2.9669	3.5285	4.6212
16	2.2536	2.6822	3.5166	2.4485	2.9135	3.8189	2.8926	3.4406	4.5078
17	2.2257	2.6491	3.4740	2.4102	2.8685	3.7605	2.8277	3.3637	4.4084
18	2.2007	2.6197	3.4361	2.3762	2.8283	3.7088	2.7711	3.2966	4.3213
19	2.1784	2.5934	3.4022	2.3460	2.7925	3.6627	2.7202	3.2361	4.2433
20	2.1583	2.5697	3.3715	2.3188	2.7603	3.6210	2.6758	3.1838	4.1747
21	2.1401	2.5482	3.3437	2.2941	2.7312	3.5832	2.6346	3.1360	4.1125
22	2.1234	2.5285	3.3183	2.2718	2.7047	3.5490	2.5979	3.0924	4.0562
23	2.1083	2.5105	3.2951	2.2513	2.6805	3.5176	2.5641	3.0528	4.0044
24	2.0943	2.4940	3.2735	2.2325	2.6582	3.4888	2.5342	3.0169	3.9580
25	2.0813	2.4786	3.2538	2.2151	2.6378	3.4622	2.5060	2.9836	3.9147
26	2.0693	2.4644	3.2354	2.1990	2.6187	3.4375	2.4797	2.9533	3.8751
27	2.0581	2.4512	3.2182	2.1842	2.6012	3.4145	2.4560	2.9247	3.8385
28	2.0477	2.4389	3.2023	2.1703	2.5846	3.3933	2.4340	2.8983	3.8048
29	2.0380	2.4274	3.1873	2.1573	2.5693	3.3733	2.4133	2.8737	3.7721
30	2.0289	2.4166	3.1732	2.1450	2.5548	3.3546	2.3940	2.8509	3.7426
31	2.0203	2.4065	3.1601	2.1337	2.5414	3.3369	2.3758	2.8299	3.7148
32	2.0122	2.3969	3.1477	2.1230	2.5285	3.3205	2.3590	2.8095	3.6885
33	2.0045	2.3878	3.1360	2.1128	2.5167	3.3048	2.3430	2.7900	3.6638
34	1.9973	2.3793	3.1248	2.1033	2.5053	3.2901	2.3279	2.7727	3.6405
35	1.9905	2.3712	3.1143	2.0942	2.4945	3.2761	2.3139	2.7557	3.6185
36	1.9840	2.3635	3.1043	2.0857	2.4844	3.2628	2.3003	2.7396	3.5976
37	1.9779	2.3561	3.0948	2.0775	2.4748	3.2503	2.2875	2.7246	3.5782
38	1.9720	2.3492	3.0857	2.0697	2.4655	3.2382	2.2753	2.7105	3.5593
39	1.9664	2.3425	3.0771	2.0623	2.4568	3.2268	2.2638	2.6966	3.5414
40	1.9611	2.3362	3.0688	2.0552	2.4484	3.2158	2.2527	2.6839	3.5244
41	1.9560	2.3301	3.0609	2.0485	2.4404	3.2055	2.2424	2.6711	3.5085
42	1.9511	2.3244	3.0533	2.0421	2.4327	3.1955	2.2324	2.6593	3.4927
43	1.9464	2.3188	3.0461	2.0359	2.4254	3.1860	2.2228	2.6481	3.4780
44	1.9419	2.3134	3.0391	2.0300	2.4183	3.1768	2.2137	2.6371	3.4638
45	1.9376	2.3083	3.0324	2.0243	2.4117	3.1679	2.2049	2.6268	3.4502
46	1.9334	2.3034	3.0260	2.0188	2.4051	3.1595	2.1964	2.6167	3.4370
47	1.9294	2.2987	3.0199	2.0136	2.3989	3.1515	2.1884	2.6071	3.4245
48	1.9256	2.2941	3.0139	2.0086	2.3929	3.1435	2.1806	2.5979	3.4125
49	1.9218	2.2897	3.0081	2.0037	2.3871	3.1360	2.1734	2.5890	3.4008
50	1.9183	2.2855	3.0026	1.9990	2.3816	3.1287	2.1660	2.5805	3.3899
55	1.9022	2.2663	2.9776	1.9779	2.3564	3.0960	2.1338	2.5421	3.3395
60	1.8885	2.2500	2.9563	1.9599	2.3351	3.0680	2.1063	2.5094	3.2968
65	1.8766	2.2359	2.9378	1.9444	2.3166	3.0439	2.0827	2.4813	3.2604
70	1.8662	2.2235	2.9217	1.9308	2.3005	3.0228	2.0623	2.4571	3.2282
75	1.8570	2.2126	2.9074	1.9188	2.2862	3.0041	2.0442	2.4355	3.2002
80	1.8488	2.2029	2.8947	1.9082	2.2735	2.9875	2.0282	2.4165	3.1753
85	1.8415	2.1941	2.8832	1.8986	2.2621	2.9726	2.0139	2.3994	3.1529
90	1.8348	2.1862	2.8728	1.8899	2.2519	2.9591	2.0008	2.3839	3.1327
95	1.8287	2.1790	2.8634	1.8820	2.2425	2.9468	1.9891	2.3700	3.1143
100	1.8232	2.1723	2.8548	1.8748	2.2338	2.9356	1.9784	2.3571	3.0977

And then for the one-sided tolerance interval

k:	$\gamma = 0.1$			$\gamma = 0.05$			$\gamma = 0.01$		
n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
5	2.7423	3.3998	4.6660	3.4066	4.2027	5.7411	5.3617	6.5783	8.9390
6	2.4937	3.0919	4.2425	3.0063	3.7077	5.0620	4.4111	5.4055	7.3346
7	2.3327	2.8938	3.9720	2.7554	3.3994	4.6417	3.8591	4.7279	6.4120
8	2.2186	2.7543	3.7826	2.5819	3.1873	4.3539	3.4972	4.2852	5.8118
9	2.1329	2.6499	3.6414	2.4538	3.0312	4.1430	3.2404	3.9723	5.3889
10	2.0656	2.5684	3.5316	2.3546	2.9110	3.9811	3.0479	3.7383	5.0737
11	2.0113	2.5026	3.4434	2.2753	2.8150	3.8523	2.8977	3.5562	4.8290
12	1.9662	2.4483	3.3707	2.2101	2.7364	3.7471	2.7767	3.4099	4.6330
13	1.9281	2.4024	3.3095	2.1554	2.6705	3.6592	2.6770	3.2896	4.4720
14	1.8954	2.3631	3.2572	2.1088	2.6144	3.5845	2.5931	3.1886	4.3372
15	1.8669	2.3289	3.2118	2.0684	2.5660	3.5201	2.5215	3.1024	4.2224
16	1.8418	2.2990	3.1720	2.0330	2.5237	3.4640	2.4594	3.0279	4.1233
17	1.8195	2.2724	3.1369	2.0017	2.4862	3.4144	2.4051	2.9627	4.0367
18	1.7995	2.2486	3.1054	1.9738	2.4530	3.3703	2.3570	2.9051	3.9604
19	1.7815	2.2272	3.0771	1.9487	2.4231	3.3308	2.3142	2.8539	3.8924
20	1.7652	2.2078	3.0515	1.9260	2.3960	3.2951	2.2757	2.8079	3.8316
21	1.7503	2.1901	3.0282	1.9053	2.3714	3.2628	2.2408	2.7663	3.7766
22	1.7366	2.1739	3.0069	1.8864	2.3490	3.2332	2.2091	2.7285	3.7268
23	1.7240	2.1589	2.9873	1.8690	2.3283	3.2061	2.1801	2.6940	3.6812
24	1.7124	2.1451	2.9691	1.8530	2.3093	3.1811	2.1535	2.6623	3.6395
25	1.7015	2.1323	2.9524	1.8381	2.2917	3.1579	2.1290	2.6331	3.6011
26	1.6914	2.1204	2.9367	1.8242	2.2753	3.1365	2.1063	2.6062	3.5656
27	1.6820	2.1092	2.9221	1.8114	2.2600	3.1165	2.0852	2.5811	3.5326
28	1.6732	2.0988	2.9085	1.7993	2.2458	3.0978	2.0655	2.5577	3.5019
29	1.6649	2.0890	2.8958	1.7880	2.2324	3.0804	2.0471	2.5359	3.4733
30	1.6571	2.0798	2.8837	1.7773	2.2198	3.0639	2.0298	2.5155	3.4465
31	1.6497	2.0711	2.8724	1.7673	2.2080	3.0484	2.0136	2.4963	3.4214
32	1.6427	2.0629	2.8617	1.7578	2.1968	3.0338	1.9984	2.4782	3.3977
33	1.6361	2.0551	2.8515	1.7489	2.1862	3.0200	1.9840	2.4612	3.3754
34	1.6299	2.0478	2.8419	1.7403	2.1762	3.0070	1.9703	2.4451	3.3543
35	1.6239	2.0407	2.8328	1.7323	2.1667	2.9946	1.9574	2.4298	3.3343
36	1.6182	2.0341	2.8241	1.7246	2.1577	2.9828	1.9452	2.4154	3.3155
37	1.6128	2.0277	2.8158	1.7173	2.1491	2.9716	1.9335	2.4016	3.2975
38	1.6076	2.0216	2.8080	1.7102	2.1408	2.9609	1.9224	2.3885	3.2804
39	1.6026	2.0158	2.8004	1.7036	2.1330	2.9507	1.9118	2.3760	3.2641
40	1.5979	2.0103	2.7932	1.6972	2.1255	2.9409	1.9017	2.3641	3.2486
41	1.5934	2.0050	2.7863	1.6911	2.1183	2.9316	1.8921	2.3528	3.2337
42	1.5890	1.9998	2.7796	1.6852	2.1114	2.9226	1.8828	2.3418	3.2195
43	1.5848	1.9949	2.7733	1.6795	2.1048	2.9141	1.8739	2.3314	3.2059
44	1.5808	1.9902	2.7672	1.6742	2.0985	2.9059	1.8654	2.3214	3.1929
45	1.5769	1.9857	2.7613	1.6689	2.0924	2.8979	1.8573	2.3118	3.1804
46	1.5732	1.9813	2.7556	1.6639	2.0865	2.8903	1.8495	2.3025	3.1684
47	1.5695	1.9771	2.7502	1.6591	2.0808	2.8830	1.8419	2.2937	3.1568
48	1.5661	1.9730	2.7449	1.6544	2.0753	2.8759	1.8346	2.2851	3.1457
49	1.5627	1.9691	2.7398	1.6499	2.0701	2.8690	1.8275	2.2768	3.1349
50	1.5595	1.9653	2.7349	1.6455	2.0650	2.8625	1.8208	2.2689	3.1246
55	1.5447	1.9481	2.7126	1.6258	2.0419	2.8326	1.7902	2.2330	3.0780
60	1.5320	1.9333	2.6935	1.6089	2.0222	2.8070	1.7641	2.2024	3.0382
65	1.5210	1.9204	2.6769	1.5942	2.0050	2.7849	1.7414	2.1759	3.0039
70	1.5112	1.9090	2.6623	1.5812	1.9898	2.7654	1.7216	2.1526	2.9739
75	1.5025	1.8990	2.6493	1.5697	1.9765	2.7481	1.7040	2.1321	2.9474
80	1.4947	1.8899	2.6377	1.5594	1.9644	2.7326	1.6883	2.1137	2.9237
85	1.4877	1.8817	2.6272	1.5501	1.9536	2.7187	1.6742	2.0973	2.9024
90	1.4813	1.8743	2.6176	1.5416	1.9438	2.7061	1.6613	2.0824	2.8832
95	1.4754	1.8675	2.6089	1.5338	1.9348	2.6945	1.6497	2.0688	2.8657
100	1.4701	1.8612	2.6009	1.5268	1.9265	2.6839	1.6390	2.0563	2.8496