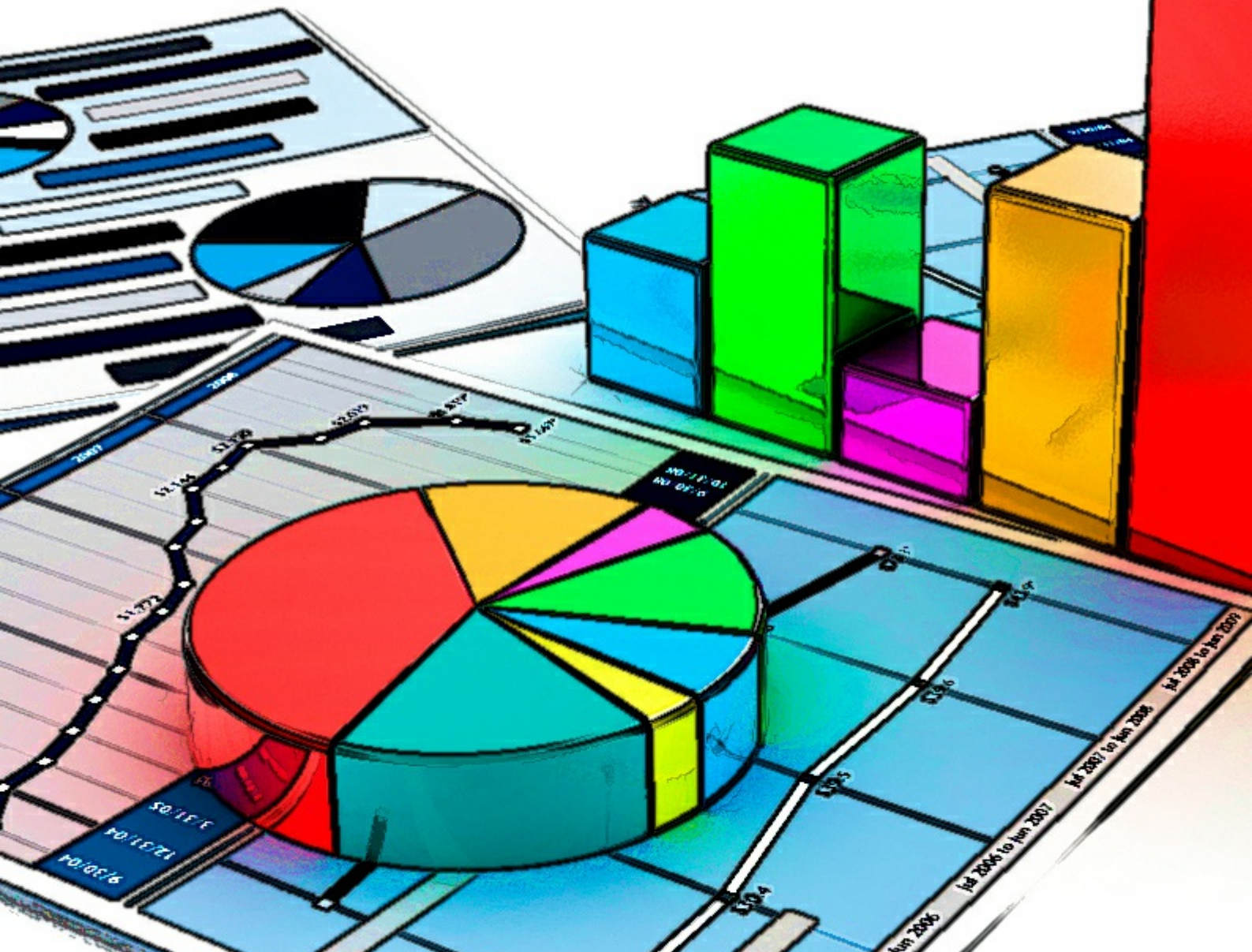


Understanding Statistics

Sture Holm



STURE HOLM

UNDERSTANDING STATISTICS

Understanding Statistics

1st edition

© 2016 Sture Holm & bookboon.com

ISBN 978-87-403-1407-6

Peer reviewed by Professor emeritus Elisabeth Svensson, Örebro university


CONTENTS

	About the author	6
1	Lots of figures – with quality and without	8
2	More or less probable	11
3	Dependence and independence	16
4	My first confidence interval	24
5	Location and dispersion in theory and practice	31
6	A useful root	41
7	Completely normal and almost normal	47
8	Within the error margin	57
9	Pure juridical matters	63
10	Some old classics	70

I joined MITAS because
I wanted **real responsibility**

The Graduate Programme
for Engineers and Geoscientists
www.discovermitas.com







Month 16

I was a construction supervisor in the North Sea advising and helping foremen solve problems

Real work
International opportunities
Three work placements







11	Comparing two cases	77
12	One sided or two sided	89
13	It depends	92
14	Analysis of variance	99
15	And analysis of variances	107
16	Now and then or here and there	110
17	Generalised linear models	118
18	To measure the almost non-measurable	122
19	Like a starry sky	127
20	All these ranks	130
21	Simulation, imputation and elimination	134
	Where is the concept defined?	144

ABOUT THE AUTHOR

Sture Holm, born in 1936, is a retired professor of biostatistics at Göteborg University. After starting his academic career with a Master's degree in Electrical engineering and working with construction of radar systems in the industry for some years, he went back to the university for further study of mathematical statistics. From the beginning the interest was directed towards random processes, but soon it shifted over to statistical inference both theoretically and in application.

He has always had a broad interest within the field. Among the early subfields of his may be mentioned nonparametric statistics and sequential analysis. Always there has been a broad genuine interest of all kinds of application as well as education at all levels. In the job as a senior lecturer at Chalmers Technical University he used to give a basic courses in statistics to almost 500 students per year. During those years he also wrote some Swedish textbooks. All this time, but also later, he has had an interest both to give advanced lectures to specialists and to give introductory lectures to those who did not have a mathematical background, and perhaps even hated mathematics.

Year 1979–1980 Sture Holm was employed at Aalborg University as professor of “Mathematics, in particular mathematical statistics”. It was a good opportunity at this time for two reasons. The first one was to get a better balance between education and research than in Göteborg earlier, and the other one was to be able to work in the interesting education system they had in Aalborg, with good real life projects included in the education all the time and small students groups working together under surveillance mixed freely with lectures in bigger groups when needed.

This year 1979 also appeared his not known paper ‘A Simple Sequentially Rejective Multiple Test Procedure’ (Scandinavian Journal of Statistics, 6, p. 65–70). It met a very big interest in different fields of application and is nowadays in the reference list of more than eleven thousands of scientific papers. It was a pioneering paper on methods to handle several statistical issues simultaneously in a common strictly logical setting. During all times very few papers within statistics has reached that number of citation.

From 1984 to 1993 Sture Holm was the professor in Statistics at the School of Economics in Göteborg which gave insight in a new type of applications. Since it was the only professorship at the department it also meant to give PhD courses in all parts of statistical inference. Bootstrap is a nice idea for statistical applications, which interested him much in those years. He gave seminars on bootstrap at several universities and made also some theoretical contributions to the field.

In 1993 there was created a new professorship in Biostatistics at the faculty of Natural sciences and mathematics at Göteborg University supported by the medical faculty. The job was placed at the Institute for mathematical sciences, which is common for Göteborg University and Chalmers Technical University. It gave a better possibility to get good coworkers and students as well better contact with many important applications, among them also medical.

One type of data appearing often in ‘soft sciences’ is results of scale judgments. Holm started a development of suitable methods for analysis together with a coworker, who has later developed the methods further. Another type of work is design of experiments and analysis of variance dependence in those, starting also with a coworker, who has continued the development. He has also given courses for industry in experimental design, and there is a Swedish book by him on this subject available at bookboon.com. With a colleague he has done some works on models for metal fatigue life analysis. Application work has been done for instance also within odontology on the analysis of oral implants, analysis of weight modules in archeological gold and silver finds and numerous other types of applications.

Of recent interests on application oriented statistical methods may be mentioned development of statistically proper methods for rankings between units concerning some quality, e.g. operation results for different hospitals, and investigation of properties of methods using imputation e.g. in educational studies.

During the years Sture Holm has also had some leading academic positions. He has been the head of the Department of Mathematical sciences at Chalmers and Department of statistics at the School of economics, head of education in Engineering physics at Chalmers, chairman of the Swedish statistical association for two periods and chairman (president) of the Nordic region of Biometric society for one period.

1 LOTS OF FIGURES – WITH QUALITY AND WITHOUT

There are lots of figures that appear in the newspapers, and figures are often mentioned on TV as well. A certain gene may increase the risk of getting a certain disease by a factor of three, a certain fraction of thirteen-year-old girls smoke almost every day, three out of four citizens think that the mayor should resign and the support for the labor party has increased since last month. But this last change is said to be within the error margin.

What is all this now? Figures are figures which may perhaps be understood. But what quality do they have? Error margin, what is that? And what is it useful for?

Everyone ought to understand that statistical estimates may have different qualities depending on how they have been collected, how many units have been included in the estimate and so on. In a certain time period, there were 66.7% women among the professors of biostatistics in Sweden. A clear indication of a change of gender distribution in academic life? Of course, the fact that two out of three professors were women may give some tiny little indication, but we must regard these 67.7% as a poor quality estimate due to the small sample size.

The central statistical office as well as others who conduct studies on people's preferences among political parties, have much bigger sample sizes. They also make their estimates based on random samples. They then get much better precision in their estimates. Further, they often have a quality declaration by reporting an error margin. Even if one does not fully understand the exact mathematical meaning of this concept, it gives quite a good idea of the possible errors in the estimates. A change within the error margin is 'not much to talk about'. Margins of error also appear in other contexts in general life. If we say that the distance to the town centre is four and a half kilometers, there is certainly some error margin in this figure. And if we estimate the cost of a holiday trip at 3000 euros, and the real cost appears to be 3107 euros we think that it is within the error margin. In forthcoming sections we will further discuss error margins and similar things, but for the moment we focus on the basic problem that quality declarations so often miss.

One thing may first be noted. The quality of an estimate depends very much on the sample size, that is the number of units (people, machines, towns or whatever) involved in the investigation. Good design helps, careful data collection helps, but it is unavoidable that a small sample size gives big random variations in the estimates. If the estimate is a relative frequency, i.e. a ratio between the number of units satisfying some condition and the total number of units, a reader with basic statistical knowledge can make an approximate quality measure by himself or herself. It would be a kind of approximate mean error. More on this will feature in a later section. Earlier I had to make such quality calculations myself very often, but nowadays at least in the case of political alignments, party sympathy investigations often declare these quality measures. So fortunately it has become better in this case. In most other investigations there are no reports of quality measures, and the information on design and data collection is not enough to make a self calculation, even for a skilled statistician. This is mostly true even if the sample size happens to be reported.

Medical investigations and results in technical applications often have proper quality measures. For example, one can read that a certain complication can be associated with a 30% higher risk of death, which is however not significant. This is also a kind of quality declaration. It means that the measured increase in risk of death perhaps is just a natural random variation within the patient group in the investigation. We will discuss the concept of significance in a special section in this book. In a technical context one can, for example, read that the investigated material breaks at a load between 47 and 61 units per square inch. It may be said that this interval is a 95% confidence interval. This concept is related to error margin and mean error. It will be discussed further in over the following sections.

Unfortunately, there are some types of investigations where you almost never get any kind of quality declaration. Newspapers may perhaps present a ranking of the climate for entrepreneurship in different communities. Owners of companies would have answered multiple choice questions and the results are weighted together in some kind of index, which is compared across the communities. Then there may be a discussion that this year community A is not as good as community B, because they were placed twenty-fourth with an index value 27.62 while community B was placed thirteenth with an index value 28.31. Very sad, since in the previous investigation two years ago, A was placed ninth and B, seventeenth. This is a completely meaningless discussion without information on mean error or other statistical quality measures related to mean error. If for instance, the mean error was 1.50, pure random effects would be much larger than the observed differences between A and B. Even if the mean error was as small as 0.50 a random difference of 0.69, as in the example, would be quite a natural random variation.

From a psychological point of view it is easy to be trapped in erroneous thinking in this situation. When some communities are quite far from each other on the list, there may truly seem to be a substantial difference between these communities. But without the knowledge of mean errors or an equivalent parameter, the discussion is meaningless. The right place for this so-called investigation is the waste-paper basket.

Unfortunately, such comparisons without quality declarations appear in many different fields. You may for instance see comparisons between services in medical care units, comparisons between shirking in schools and so on.

Why are such statistics produced then? Quite often it would not be so very difficult to give at least approximate quality measures. If the producers had at least a basic knowledge of statistical theory, they would be able to do that. There may be two reasons. One is that perhaps they do not have this basic knowledge. Another reason may be that perhaps they could make some proper calculations, but do not want to, because that would reveal how bad their data are. It would then be more difficult to sell this type of work to newspapers and other organisations in the future.

The ambition of this book is to explain statistical principles to those who are not specialists in the field. I will discuss different types of common statistical methods and concepts. Those who have studied a little more of statistics, have quite often studied these in a technical computational way, sometimes without a basic understanding. They may perhaps also make good use of a little book which concentrates on the understanding of statistics. And thirdly, it might be good to have an accompanying text which concentrates on the basic principles, while you study any course in statistics. So I hope that this little book may serve all these cases well.

2 MORE OR LESS PROBABLE

Statistics has a lot to do with probabilities. We will not go into the theory of probability in any depth, but with respect to the following sections it may be helpful to first have a look at the most elementary probability concepts.

If we toss a dice, the probability of getting a six is equal to a sixth, or about 17%. Everyone knows that. And that a lottery with 2 million lottery-tickets and 492624 winning lottery-tickets has a winning chance of $\frac{g}{m} = \frac{492624}{2000000} = 0,246 = 24,6\%$ is also rather self evident. Almost one fourth of the lottery-tickets are winning ones. Well, in this lottery, 400007 of the prizes have the value which is the same as the price of the lottery-ticket, so there are only 92617 lottery-tickets which give a gain. Thus the probability of gaining when you buy a lottery-ticket is only $\frac{g}{m} = \frac{92617}{2000000} = 0,046 = 4,6\%$.

These trivial calculations follow the simplest model for determining probabilities, the so-called classical probability model, where you determine the probability p for an event as the ratio $p = \frac{g}{m}$, between the number g of cases favorable for the event and the total number m of possible cases.

Probably people have thought this way earlier too, but a formal definition and slightly more advanced calculations of this type were first used in the seventeen hundreds, in connection with interest in some game problems. To calculate how many cases of different kinds there are, is a topic of mathematics which is called combinatorial analysis. The combinatorial problems, although often easily stated, are sometimes very tricky to solve. We will not go into these things. Now we turn to a very simple example. The calculations are trivial, but in a quality control situation, this type of calculation is just what is needed.

Example

Let us suppose that some kind of units are manufactured for sale, and that the producers want to keep track of the units' ability to function. Perhaps a certain amount of time goes into a control function for this. Can this time be decreased by checking only a sample of units in each batch? We study this problem in a numerical example.

The batch size is 100. Consider first the case where 10 randomly chosen units from each batch are checked. If at least one non-functioning unit is found in the sample, further action is undertaken. The whole batch is then checked and a general control of the production process is undertaken.

What is the probability that a sample gives such a general control of the batch, if there are in fact 20 non-functioning units in the batch? When using the classical probability model, we have to consider the total number of possibilities of choosing 10 units in a batch of 100 and the number of the ‘favorable’ cases, where there is at least one non-functioning unit in the sample.

The first unit in the sample may be chosen in 100 ways. For each of these ways there are then 99 ways to choose the next unit in the sample. For each combination of the choice of the first two units, the third unit may be chosen in 98 ways. And so on. The total number of ways to chose 10 units equals

$$m = 100 \cdot 99 \cdot 98 \cdot 97 \cdot 96 \cdot 95 \cdot 94 \cdot 93 \cdot 92 \cdot 91.$$



www.job.oticon.dk

oticon
PEOPLE FIRST

This number is extremely large. In mathematical terms, it can be written as $62.82 \cdot 10^{18}$. The number 10^{18} is, simply put, millions of millions of millions (which is certainly true here, since 10^6 is a million).

The simplest way to calculate the number of favourable cases is to take all cases minus the unfavourable ones. This latter number here equals $80 \cdot 79 \cdot 78 \cdot 77 \cdot 76 \cdot 75 \cdot 74 \cdot 73 \cdot 72 \cdot 71$ or $5.97 \cdot 10^{18}$. The number of favorable cases is now $62.82 \cdot 10^{18} - 5.97 \cdot 10^{18} = 56.85 \cdot 10^{18}$,

and the probability we want to calculate is

$$\frac{g}{m} = \frac{56,85 \cdot 10^{18}}{62,82 \cdot 10^{18}} = 0,905 = 90,5\%$$

After this journey in the world of big numbers, we arrive at the statement that if there are in fact 20 non-functioning units in the batch, the chance is 90.5% that we will discover it and make a general control of the whole batch.

Enough of calculations for the moment. I think that you now understand the principles well enough so you could make your own calculations for this type as also for other numbers of real non-functioning units in the batch. I have run calculations for all cases from 1 to 30 non- functioning units in the batch, and exhibited the result in the following figure.

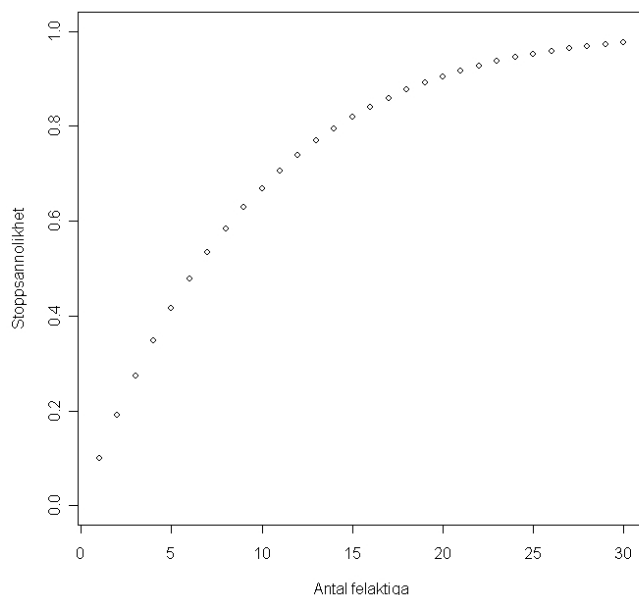
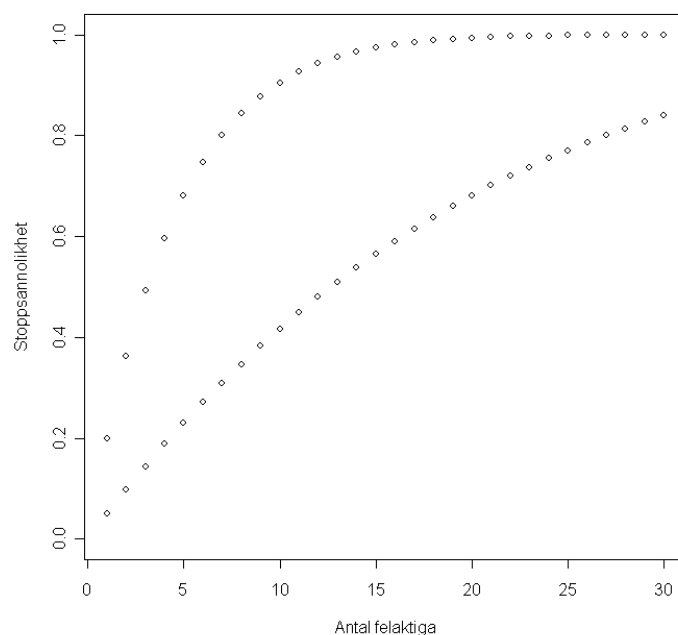


Figure 2.1. Probability (y-axis) of discovering problems in the batch with 100 units as a function of the true number of non-functioning units (x-axis) in the whole batch, for a quality control based on a sample size of 10.

We have used a very simple probability model here and made very elementary calculations, even if the numbers are big. Yet we have found results which may be useful in practice. We can for instance, see in the figure that if there are about 30 non-functioning units in the whole batch, we are almost sure to discover that there are problems with the production. On the other hand if there are 5 or less non-functioning units in the batch, we have rather a small probability of going ahead with a general control of the whole batch.

You can also gain some general understanding from this simple problem. You see that due to random influence there is always a risk of wrong conclusions in a statistical investigation. But you can also learn that with suitable calculations you can find the size of that risk.

We finish the section by looking at how we could suitably change the control procedure by changing the sample size. In the figure below, I show calculated probabilities of discovering problems as a function of the real number of non-functioning units in the batch, both, for double the sample size 20 and for the half the sample size 5, instead of 10 as before. Some calculations have to be done, but they all have the same elementary character as before.



Figur 2.2. Probability (y-axis) of discovering problems as a function of the real number of non-functioning units (x-axis) in quality controls with sample sizes 20 (upper curve) and 5 (lower curve).

In the figure we can clearly see that if we take a sample of size 20 from each batch, we have a high chance (90% or more) of discovering production problems if there are more than 10 non-functioning units. If we take a small sample of size 5 only, we have only a limited chance (up to 80%) of the same, even if we have as many as 25 non-functioning units in the whole batch.

These calculations clearly show that the quality of a statistical method is highly dependent on the sample size. A more detailed discussion of the importance of sample size will feature in a later section.

Now we leave this introductory quality control problem. All calculations could be made by the simple classical probability model in this case. But I want to finally point out that the simple classical model should be used in practice only for situations where the cases are of equal type, i.e. the possible outcomes can be assumed to have the same basic probability.



In the past four years we have drilled

81,000 km

That's more than **twice** around the world.

Who are we?
We are the world's leading oilfield services company. Working globally—often in remote and challenging locations—we invent, design, engineer, manufacture, apply, and maintain technology to help customers find and produce oil and gas safely.

Who are we looking for?
We offer countless opportunities in the following domains:

- **Engineering, Research, and Operations**
- **Geoscience and Petrotechnical**
- **Commercial and Business**

If you are a self-motivated graduate looking for a dynamic career, apply to join our team.

careers.slb.com

What will you be?

Schlumberger

3 DEPENDENCE AND INDEPENDENCE

In the previous section we considered probabilities only according to the classical definition, as the ratio of favorable and possible cases. However, this simple definition is not enough for most application situations. In this section we will take a look at another simple form of probability calculation and its practical applications. We start with the principle coupling between the theoretical probability model and the empirical reality, where the model is used.

What does it mean if a medical paper declares that there is an 8% probability of a mild adverse effect? It ought to mean that this adverse effect appears in 8% of a large population or that this percentage has been estimated in a smaller sample from the population. We talk here of the relative frequency of 8% in the population. This relative frequency in the sample is an empirical estimate of the theoretical probability of the adverse effect, which can be thought of as the relative frequency in the whole considered population. This is the coupling we have between the empirical and theoretical world, and which should be there for all kinds of situations and for all kinds of events.

If an item of a mathematical test for some grade in school has a degree of difficulty with a chance of 40% for the pupils to get it right, this ought to mean that one has either observed that 40% of the pupils in a representative big population have got the test item right or that some authority has made the judgment that this is the case. We take this figure as a basis for a simple numerical discussion of a very important concept in statistics, the independence concept.

Now think of two randomly chosen pupils A and B, who have to solve the above mentioned item. When we consider the two pupils at the same time, there are four possible combined outcomes: both A and B get it right, A gets it right but not B, B gets it right but not A, and neither A nor B gets it right. What value is reasonable for the probability for the combined event that both A and B get it right?

One randomly chosen pupil has the probability 40% to get it right. Either this event has happened or not; there is 40% probability that the other randomly chosen pupil should get it right. The chance that both pupils A and B get it right, is 40% of 40%, which is $0,40 \cdot 0,40 = 0,16 = 16\%$. In a similar manner we find that the reasonable value of the probability that A gets it right but B does not, is $0,40 \cdot 0,60 = 0,24 = 24\%$. The probability that B gets it right but A does not is also 24% and finally the event that none of them get it right, is $0,60 \cdot 0,60 = 0,36 = 36\%$. Observe that the sum of the probabilities for the four cases adds up to $16\% + 24\% + 24\% + 36\% = 100\%$.

When two events in this way have a probability that both should happen, which is equal to the product of the probability of the individual events, we say that the events are independent.

The most important statistical independence concept is independent sub-trials. Suppose a trial consists of two sub-trials. If any event in the first sub-trial is independent of any event in the second sub-trial, we say that the sub-trials are independent. Observe that it should hold for all possible combinations of events in the two sub-trials.

Independent sub-trials is usually not something you make a calculation to find. It is usually an assumption that you have reason to make when there are sub-trials, whose random results do not influence each other. When we can make this assumption of independence we can in principle calculate the probability for all combined events. Here is a table for a simple numerical example.

0,1		0,03	0,05	0,02
0,7		0,21	0,35	0,14
0,2		0,06	0,10	0,04
Sub-trial 2 itself				
	Sub-trial 1 itself	0,3	0,5	0,2

Table. Probabilities in a trial which consists of two independent sub-trials with three possible outcomes and different probability distributions.

A trial which consists of more than two sub-trials whose results do not influence each other randomly, can also be supposed to be independent in a generalised definition. There the probability of any combined event is the product of all the events in the individual sub-trials. We take a simple example.

Example


Consider four randomly chosen pupils, who are to work with the item we had in the example above. In the example we had the probability of 40% for success. We now generalise this slightly by using instead a general notation p for any value of that probability. The results for the four pupils are supposed to be independent.

The probability that all the four pupils should get it right is now p^4 and the probability that none of them gets it right is $(1 - p)^4$. The probability that pupil number 1 gets it right, but none of the others do, has a probability $p(1 - p)^3$. It is the same probability that only pupil number 2 has right (but none of the others has). In all there are four scenarios with this probability in the combined full trial.

The probability that pupil number 1 misses, but all the other gets it right is $(1 - p)p^3$. Also, here there are in all four probabilities for one pupil missing and the others getting it right.

The probability that pupils 1 and 2 get it right and the other two miss equals $p^2(1 - p)^2$. Thinking of all cases you find that there are six cases in all with this probability. Thus now finally we have the following table of the results.



 Sweden
Sverige

Linköping University –
innovative, highly ranked,
European

Interested in Engineering and its various branches? Kick-start your career with an English-taught master's degree.

→ [Click here!](#)

li.u LINKÖPING
UNIVERSITY

Number getting right	0	1	2	3	4
Probability for one case	$(1 - p)^4$	$p(1 - p)^3$	$p^2(1 - p)^2$	$p^3(1 - p)$	p^4
Number of cases	1	4	6	4	1

Going back to our numerical example with p -value 0.4, for instance, the probability that exactly two pupils get the item right equals $6 \cdot 0,4^2 \cdot 0,6^2 = 0,3456$ and the whole distribution of the number of pupils getting the item right can be mapped as in the following figure.

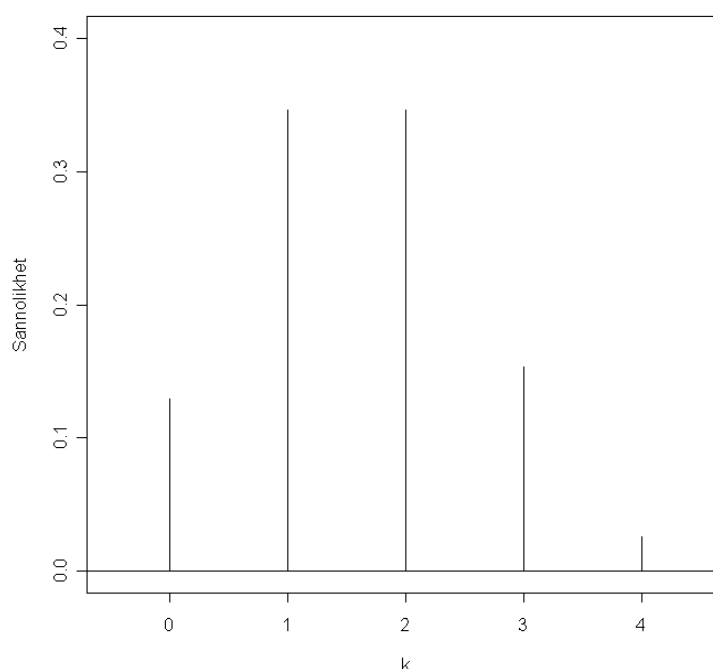


Figure 3.1. Probability distribution for the number of pupils getting the item right.

What am I doing here? I have just introduced you to the most important probability distribution for random variables, with outcomes in form of countable numbers. It is called the binomial distribution. The motivation for its use is just that it fits well as a distribution for the number of times a given event occurs in a number of independent sub-trials of the same kind.

The binomial distribution has two parameters, the size parameter, often denoted by n and the probability parameter, often denoted by p . Thus in our introductory example, the parameters are $n = 4$ and $p = 0.4$. In a mathematical description the probability of outcome k equals

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Here the factor $\binom{n}{k}$, the so called binomial coefficient, is the number of ways to choose k units out of n units, without respect to the order of units.

The motivation for the use of binomial distribution can be generally deduced mathematically. It follows, in principle, our simple motivation for the case $n = 4$ above. We do not care very much about the mathematical technique here, but I hope that you understand the importance of the motivation for the use of the distribution. This simple type of situation appears in many applications. To get the numerical values of probabilities in the distribution is a job for a computer. It's rather cumbersome to make it by hand.

All reasonably big mathematical or statistical computer programs can handle the necessary calculations. If you do not already have a program available, you can always download the statistics program R from the internet, free of cost. One such url is <http://ftp.sunet.se/pub/lang/CRAN/>. You can also google, for instance, 'statistics program R'. There is also an instruction booklet for the program. I have used that program for compiling all calculations and figures in this book. It is a good program with a lot of possibilities. One drawback is that it is operated by commands. However these are listed in the instruction booklet. There are no menus with alternatives or other click alternatives.

Here are some examples of binomial distributions. The first one may, for instance, illustrate the distribution of the number of patients with mild adverse effects in a group of 50 patients, when the adverse effect has a probability of 8%. This is an example we had in the beginning of this section.

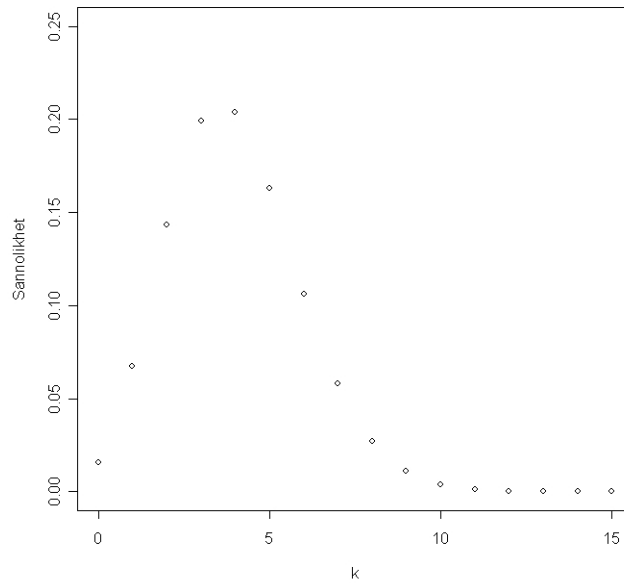


Figure 3.2. Binomial distribution with parameters $n = 50$ and $p = 0,08$.

Here we have a skewness for natural reasons. There is a ‘tail’ on the right side, but none on the left side. There is no room for a tail on that side because there cannot be any negative outcomes.

STUDY FOR YOUR MASTER'S DEGREE
IN THE CRADLE OF SWEDISH ENGINEERING

Chalmers University of Technology conducts research and education in engineering and natural sciences, architecture, technology-related mathematical sciences and nautical sciences. Behind all that Chalmers accomplishes, the aim persists for contributing to a sustainable future – both nationally and globally.

Visit us on **Chalmers.se** or **Next Stop Chalmers** on facebook.

CHALMERS
UNIVERSITY OF TECHNOLOGY



Let us take another example. The number of pupils in a class of 20, who get the right answer for a puzzle, could have a binomial distribution with parameters $n=20$ and $p = 0.4$, if the probability of getting the right answer is 0.4. This binomial distribution has the following shape.

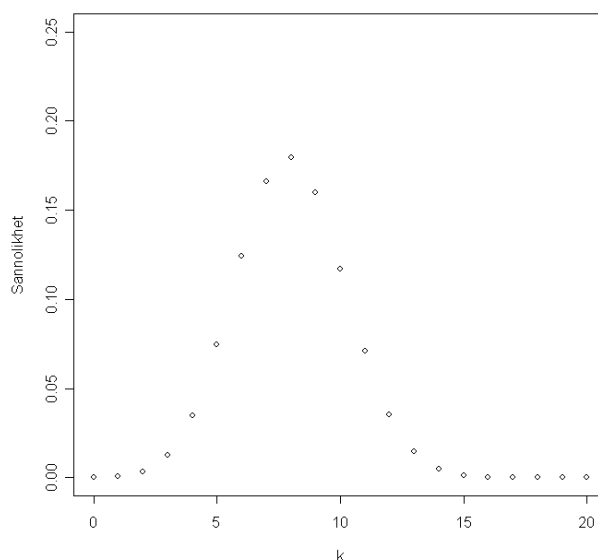


Figure 3.3. Binomial distribution with parameters $n = 20$ and $p = 0,4$

One could expect that among 20 randomly chosen pupils, there would be approximately $20 \cdot 0,4 = 8$ with the right answer. It is not always exactly this number because of the random variation. But there is a great chance of getting between 3 and 14 pupils with the right answer, according to the figure. Are you surprised that the variation is so huge? I can understand if you are, but the variations are really that large. We will come back to the size of the random variations several times in subsequent sections. I hope that by the time you have read the whole book, you will have got a good idea of the size of random variation in different situations.

The numbers n and p in the binomial distribution are called the parameters of the distribution. In mathematics and statistics, as well as often in natural sciences and technique, the word parameter means something which determines ‘which case we have here’. The parameters in the equation of the straight line determine which line it is (of all possible ones), and so on. In recent decades I have noticed that both in medicine and in social sciences, it has become common to use the name parameter for observations. My aim in writing for a general audience is to only use words that are understandable by everyone, as far as possible. But in this case I must stick to the mathematical convention and use the word parameter only in its original meaning in order to not confuse the reader completely. I will use the words observations, variables and measurement values for what we see in the real world, and use the word parameter only for the abstract numbers behind, which determine which case we have at hand. But in order to express myself clearly, I will often attach the prefix empirical for observations in the real world and the prefix theoretical for parameters in the abstract world.

Can we always use the binomial distribution as a distribution of counts? No! A very important assumption in the deduction of the binomial distribution is that the sub-trials can be considered to be independent. If that assumption is not satisfied, it does not work. If for instance, a zoologist studies the breeding success for pied flycatchers, it does not work. What is then so special about a zoologist? Nothing! But there is something about pied flycatchers! Now let me explain. A natural way to measure the breeding success is to estimate the proportion of laid eggs that hatch to a fine young bird. In nature there are, however, some risks that counteract a successful result. Birds of prey and pollution influence the result locally. Often there is a negative result for all or a number of eggs in the same nest. This means that the breeding results for eggs in the same nest are dependent. The independence assumption for the deduction of the binomial distribution does not hold.

Another example where the assumptions do not hold fully, is found in pedagogical studies. If one studies test results, where whole classes or parts of classes are included, the teacher has an influence on the result of his or her whole class, which gives a dependence between students in the same class. From a strictly mathematical point of view there is not much of a difference between a bird nest and a teacher.

4 MY FIRST CONFIDENCE INTERVAL

The probability distribution for a discrete random variable which can get outcomes only at distinct points, can generally be described by a probability mass attached to each possible outcome. The sum of all these masses is equal to one. The probability distribution of a continuous random variable, which can get outcomes in all points in an interval, cannot be described in that way. In this case we must work with a continuous distribution of mass instead. This density of probability mass is called a frequency function. For a one-dimensional continuous random variable the probability for outcome in an interval is given by the area between the frequency function and the x axis between the end points of the interval. Here is an example of a frequency function.

MÄLARDALEN UNIVERSITY SWEDEN

WELCOME TO OUR WORLD OF TEACHING!
INNOVATION, FLAT HIERARCHIES AND OPEN-MINDED PROFESSORS

STUDY IN SWEDEN - CLOSE COLLABORATION WITH FUTURE EMPLOYERS
MÄLARDALEN UNIVERSITY COLLABORATES WITH MANY EMPLOYERS SUCH AS ABB, VOLVO AND ERICSSON

TAKE THE RIGHT TRACK
GIVE YOUR CAREER A HEADSTART AT MÄLARDALEN UNIVERSITY
www.mdh.se

DEBAJYOTI NAG
SWEDEN, AND PARTICULARLY MDH, HAS A VERY IMPRESSIVE REPUTATION IN THE FIELD OF EMBEDDED SYSTEMS RESEARCH, AND THE COURSE DESIGN IS VERY CLOSE TO THE INDUSTRY REQUIREMENTS.
HE'LL TELL YOU ALL ABOUT IT AND ANSWER YOUR QUESTIONS AT MDUSTUDENT.COM

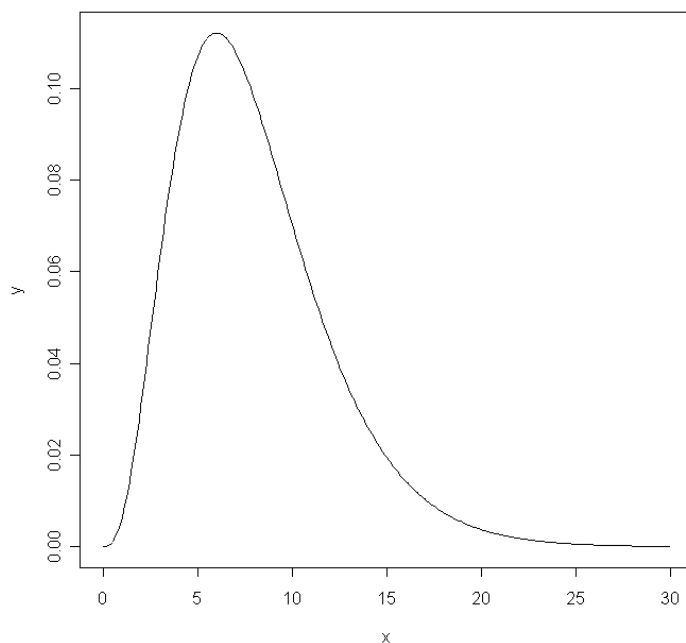


Figure 4.1. An example of a frequency function for a continuous random variable. In this case the random variable can only have non-negative outcomes since the density is 0 for negative values. The most common outcomes are in the parts where the frequency function has large values.

For a continuous random variable one can define a general position measure, the so called median. It is defined as a value, such that the probability of outcome on the two sides of this value are equal, that is, are equal to 0.50 each. In the above figure the median is in fact, equal to 7.34. The areas under the frequency function to the left and right of this value are both equal to 0.50.

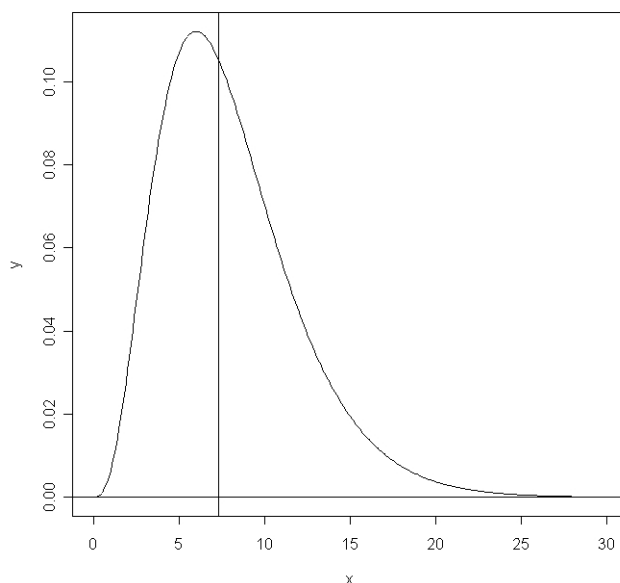


Figure 4.2. The same frequency function as in Figure 4.1 completed with an axis $y=0$ and the median line $x=7.34$. The area between the curve and the x axis is the same to the left and to the right of the line $x=7.34$ which is the median here.

Example

If we have a number of independent observations from any continuous distribution with an unknown median m , it is possible to make an interval, which with high probability catches the unknown theoretical median. This interval is called a confidence interval for the (theoretical) median, and now we will see how it can be constructed.

Consider a set of 6 observations of service times, which are continuous random variables. We denote the unknown median in the distribution by m as before. Suppose the outcomes of the 6 service times are

0,83; 1,13; 0,13; 0,94; 0,97; 1,22.

Can we now calculate the probability that the interval from the smallest observation (outcome 0.13) to the largest observation (outcome 1.22) should hit the true unknown median. Yes we can! The risk that such an interval misses the median by moving too much to the right is equal to the probability that all observations happen to get outcomes above the median. This probability is $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{64} \approx 0,0156 = 1,56\%$. In a similar way we find that the risk of missing the median by getting an interval to the left is the same $\frac{1}{64} \approx 1,56\%$. Thus the probability that the interval hits the median equals

$$1 - 2 \cdot \frac{1}{64} \approx 1 - 2 \cdot 0,0156 = 0,9688 = 96,88\%.$$

The hitting probability, which is a kind of security, is called the confidence degree of the confidence interval. We can express our numerical result in the following way

$$0,13 \leq m \leq 1,22 \quad 96,88\%$$


We can never know for sure if our interval has hit or missed the unknown median, but the meaning of the confidence degree is that we have used a method which has a probability of 96.88 percent security of hitting the median.

**LIFE SCIENCE IN UMEÅ, SWEDEN
- YOUR CHOICE!**

- 32 000 students • world class research • top class teachers
- modern campus • ranked nr 1 in Sweden by international students
- study in English

- Bachelor's programme in Life Science
- Master's programme in Chemistry
- Master's programme in Molecular Biology

Download brochure here!


UMEÅ UNIVERSITY
 FACULTY OF SCIENCE & TECHNOLOGY

If we have a much larger series of observations, the confidence degree of the interval from the smallest observation to the largest one will have a very large confidence degree – perhaps too large. Then we may construct the confidence interval instead, for example from the third smallest observation to the third largest observation. The confidence degree for such a confidence interval can be calculated with the help of the binomial distribution. We will see in the following example how it is done.

Example

Suppose that we have got the following 12 observations on the service times studied before

0.97; 0.61; 1.02; 0.63; 0.83; 2.23; 0.81; 1.42; 0.72; 1.38; 0.95; 1.75

The third smallest observation is 0.72 and the third largest observation is 1.38. What is the probability that the interval from the third smallest observation to the third largest observation misses the true median to the right? This event may also be described as the event that there are at the most two observations below the median. And the probability of this event may be calculated for a binomial distribution with parameters $n=12$ och $p=0.5$. From the statistical program we find that the event of outcome at the most 2 equals $0.0193=1.93\%$. The probability of missing to the left is the same. Thus the confidence degree for such an interval equals $1-2\cdot 0.0193=0.9614=96.14\%$.

In order to further illustrate how confidence intervals work, I have generated on the computer, 100 series with 12 observations in each series. I have calculated the confidence intervals for the median in each series. In the following figure you can see the outcomes for the limits of the 96.4% confidence intervals. In all, there were 6 intervals missing the true median, which in the simulation was known to be 0.918. Three cases got an interval to the left and three cases got an interval to the right. In real life you can never know if an interval has missed or hit, but the chance that it hits is high if the confidence degree chosen is big. Since the confidence degree here is about 96%, there ought to be in the mean 4 intervals out of 100 missing. We got 6, but that is just a normal random variation. It could just as well have been less than 4 instead.

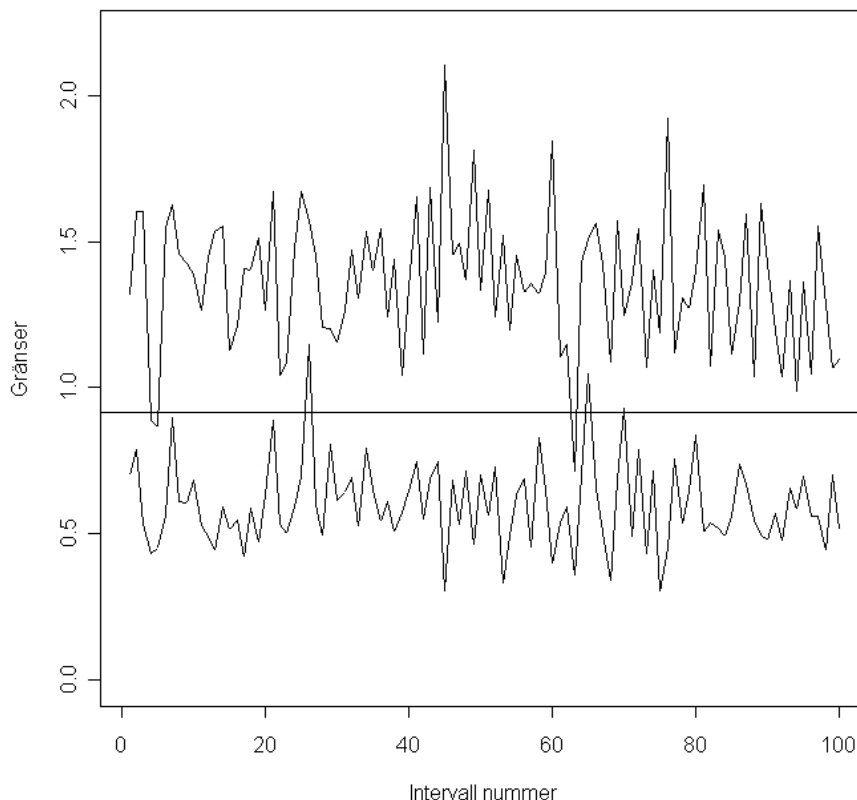


Figure 4.3. Lower and upper limits (y-axis) of 100 confidence intervals (number in x-axis) for the median with a confidence degree of 96.14% in a generated series of 12 observations.

The type of confidence intervals I have presented here are often called sign intervals. It is worth noting that the very simple method I have now described is not always efficient. If one can assume a more precise distribution for the observations, there may exist some special methods, which are more efficient, i.e. which generally gives shorter intervals. The following table is a short one of suitable choices for the order of observations to use for an approximate confidence degree of 95%. This value of the confidence degree is some kind of standard which is very much used. It is often considered to give enough safety. Of course it is good to have as high hitting a probability as possible, but very high hitting probability will also give very long intervals.

Number of observations	8	10	12	14	16	18	20
Order for lower limit	2	2	3	4	4	5	6
Order for upper limit	7	9	10	11	13	14	15
Degree of confidence	0,930	0,978	0,961	0,943	0,979	0,969	0,958

Table. Choice of ordered variables for a simple sign confidence interval.

In the description of the confidence intervals I have assumed that the random variables are continuous with a probability distribution determined by a frequency function. The method also works for discrete distributions, but the real confidence degrees will then be higher. With respect to the confidence degree being a safety declaration, the deviation goes in the correct direction. So the same type of intervals can also be used for discrete observations.

In later sections I will discuss confidence intervals for different, more specific situations. As I have already pointed out, the sign intervals are perhaps not always so efficient. But in all simplicity they work well as an introduction to the principles of confidence intervals. I hope that you now grasp the idea of a confidence interval as a kind of estimate with a built-in safety margin.

A confidence interval is a kind of
interval parameter estimate
which is constructed to have a given
high probability to catch the parameter.



Lnu.se

 *Scholarships*

**Open your mind to
new opportunities**

With 31,000 students, Linnaeus University is one of the larger universities in Sweden. We are a modern university, known for our strong international profile. Every year more than 1,600 international students from all over the world choose to enjoy the friendly atmosphere and active student life at Linnaeus University. Welcome to join us!

Linnæus University
Sweden

Bachelor programmes in
*Business & Economics | Computer Science/IT |
Design | Mathematics*

Master programmes in
*Business & Economics | Behavioural Sciences | Computer
Science/IT | Cultural Studies & Social Sciences | Design |
Mathematics | Natural Sciences | Technology & Engineering*

Summer Academy courses

5 LOCATION AND DISPERSION IN THEORY AND PRACTICE

If you want to characterise a series of observations or a probability distribution, there are two kinds of measures you think of first, a location measure and a dispersion measure. Of course there are other more detailed measures too, but these two types of measures are the most important ones.

We came across the first theoretical location measure in the previous section, the median in a distribution, which we could estimate with a sign interval. There is also an empirical point measure in the observation series corresponding to the median in the theoretical distribution. The empirical median in a series of observations is the middle observation in the order, if the total number of observations is odd. If the number of observations is even, the empirical median consists of all values between the two middle observations, limits included. Or the common value if the two middle observations are equal. But now we will consider another location measure, which is used more often than the median.

Everyone knows that the mean of an observation series is the sum of the observations divided by the number of observations. This is a very simple and easily understood location measure. And it is an empirical measure in the sense that it is determined by observed quantities in the real world.

There is also a correspondence to this measure in the theoretical world of distributions. Think first of a discrete distribution with possible outcomes $x_1, x_2, x_3, \dots, x_n$ and the corresponding probabilities $p_1, p_2, p_3, \dots, p_n$ for these outcomes. Then we define the expectation in the distribution, which is a location measure μ defined by

$$\mu = x_1 \cdot p_1 + x_2 \cdot p_2 + x_3 \cdot p_3 + \dots + x_n \cdot p_n = \sum_{k=1}^n x_k \cdot p_k$$

If you are interested in mechanics, it might help to consider this to be the gravitation centre of the mass distribution with weights p_k in the points x_k . If the distribution should be symmetric, the expectation equals the value in this symmetry point.

One could perhaps think that median and mean are equal. This is not always the case. It is true for symmetric distributions, but in general the two location measures differ a little.

For a continuous distribution, we need to use a little more advanced mathematics in order to define the mean. With help of integrals we define it as,

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

Even if you are not used to integrals, you need not worry. In practice you can think of a rounding off to some small basic unit and consider the mean in the resulting discrete case. It is then a sum of a great number of very small contributions of outcome values times the small probabilities for rounding off intervals. If you are used to integrals, you may recognise this explanation from the common definition of integrals. So do not worry because I may write a few integrals in the following. It is essentially a discrete case with many possible outcomes with extremely small probability for each of them.

Let us consider two examples of expectations. We start with a discrete one. Below is a figure of a discrete distribution. It has its highest probability in the point 5. This is not a symmetry point, but in fact the expectation also happens to be equal to 5, with the probabilities I have chosen in the example.

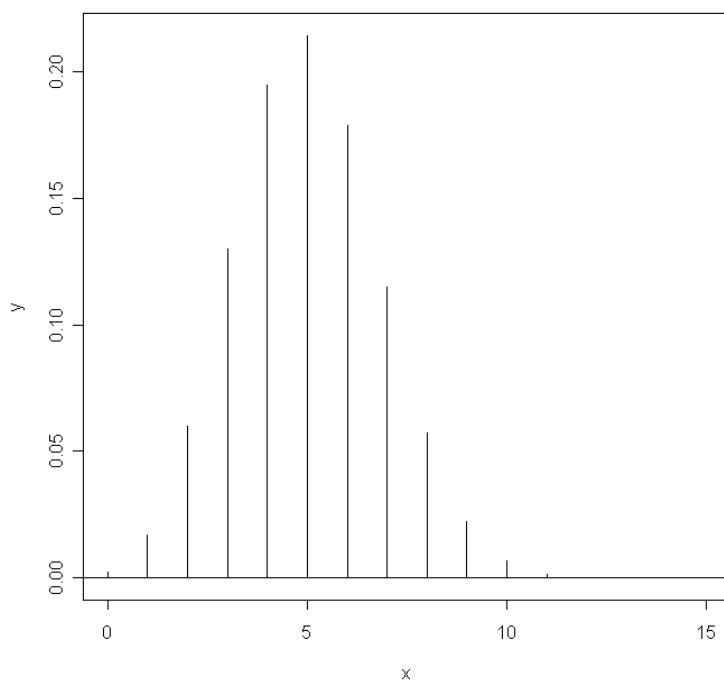


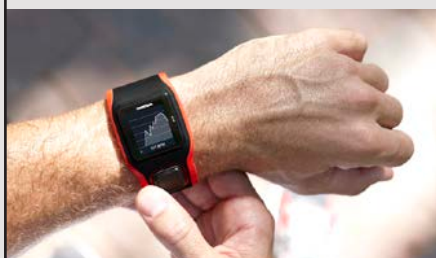
Figure 5.1. Discrete probability distribution with expectation equal to 5.0.

With the help of a computer, 10 random outcomes were generated from this distribution. The outcomes were 3, 0, 4, 4, 1, 5, 8, 5, 7, 4. These observations have the mean 4.1. This is a natural deviation of an empirical mean from the theoretical expectation, when there are (only!) 10 observations. Exactly how these deviances vary with the sample size, will be studied in the next section. But already now I will try to give you an intuitive feeling for these variations. Thus I have generated 1000 observations from this distribution, and then calculated the empirical mean for the first 10 observations, the first 20 observations and so on. In the following figure these means are given as functions of the sample sizes. The points are connected by lines in order to make the picture clearer. You can see how the deviances from the theoretical expectation 5.0 are smaller for the bigger sample sizes. If I had generated a series with an extremely big sample size, the empirical mean would differ just a little from the theoretical expectation.

**YOUR WORK AT TOMTOM WILL
BE TOUCHED BY MILLIONS.
AROUND THE WORLD. EVERYDAY.**

Join us now on www.TomTom.jobs

follow us on **LinkedIn**



#ACHIEVEMORE

TOMTOM 

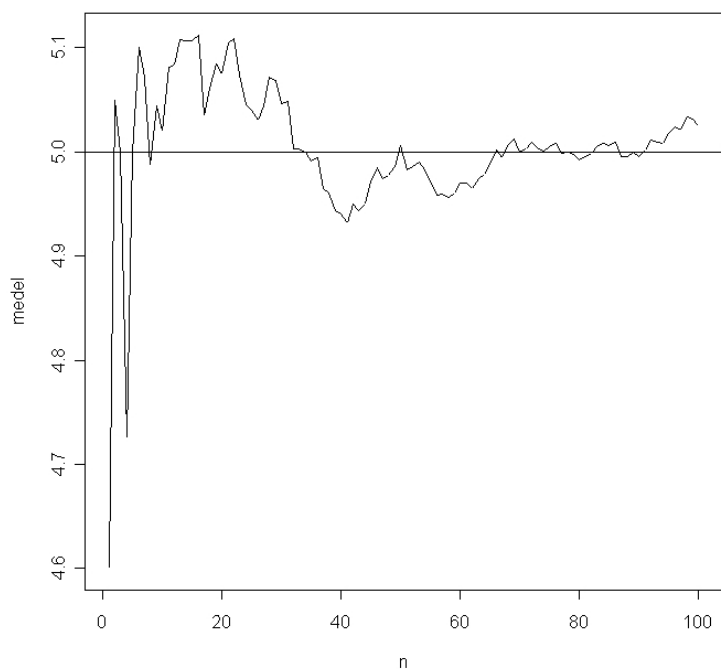


Figure 5.2. Successive empirical means (y-axis) for observation series of sizes (x-axis) 10, 20, 30, ..., 1000. The theoretical expectation 5.0 is indicated by a horizontal line.

The empirical mean works in the same way in discrete and continuous distributions. As an illustration, I have chosen a continuous distribution with expectation equal to 4.00. The form of the distribution is seen in the following figure. I generated 10 independent observations from this distribution and I got the results 2.49, 4.18, 5.59, 4.86, 3.18, 4.99, 3.12, 2.05, 5.80, 3.96. Those values, which vary between just above 2 up to almost 6 are included in the figure. Their empirical mean happened to be 4.02, which thus by pure chance came very close to the theoretical expectation.

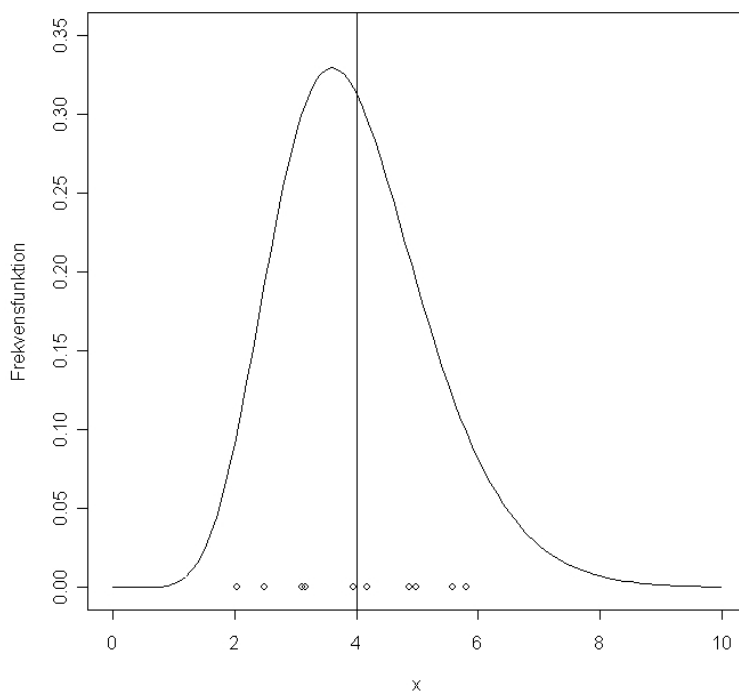


Figure 5.3. A continuous distribution and an example of a series of 10 observations from the same distribution.

The other important characterisation beside location measure is some measure of dispersion. As for the location measures, it is required to have dispersion measures both on the empirical side (for the observations) and the theoretical side (for the distribution). Some kind of mean deviation would do. How do we then make a suitable definition?

We start with the discrete distributions. Suppose the possible outcomes are $x_1, x_2, x_3, \dots, x_n$ and that their respective probabilities are $p_1, p_2, p_3, \dots, p_n$. We consider the expectation,

$$\mu = x_1 \cdot p_1 + x_2 \cdot p_2 + x_3 \cdot p_3 + \dots + x_n \cdot p_n = \sum_{k=1}^n x_k \cdot p_k$$

already calculated. It acts as a centre in the distribution. Now we can define what could be called the mean quadratic deviance σ^2 from the expectation by,

$$\sigma^2 = (x_1 - \mu)^2 \cdot p_1 + (x_2 - \mu)^2 \cdot p_2 + \dots + (x_n - \mu)^2 \cdot p_n = \sum_{k=1}^n (x_k - \mu)^2 \cdot p_k$$

The terms in this sum are squares of deviations from the expectation (centre point) μ multiplied by the probability for the corresponding possible outcomes. It may seem strange that we should square the deviances before we weight them with the probabilities. This is a smart way of getting rid of the signs of the deviations. Now however, the dimension of the calculated measure is the square of the dimension of the observations themselves. If the observations, for instance, have the unit as centimeter, the calculated measure will have the unit square centimeters. This interest in dimension is the reason that we used the notation σ^2 for this measure, which is called the (theoretical) variance in the distribution. The dispersion measure we use in practice is the (theoretical) standard deviation in the distribution, which is defined as the square root $\sigma = \sqrt{\sigma^2}$ of the variance σ^2 . We call the parameter σ the (theoretical) standard deviation in the distribution. It will have the same unit as the observations and it works well as a dispersion measure in the distribution.

Even if a formal definition of the variance in a continuous distribution includes an integral, it is in practice defined in complete analogy with the definition for a discrete distribution. You may again think of a continuous distribution as a discrete one with many possible outcomes with a very small probability for each of them.

.....Alcatel-Lucent 

www.alcatel-lucent.com/careers

What if you could build your future and create the future?

One generation's transformation is the next's status quo. In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".



Now we will take a look at some pictures of distributions with expectations and standard deviations included. Hopefully they will give you an idea of the size of the standard deviations in distributions.

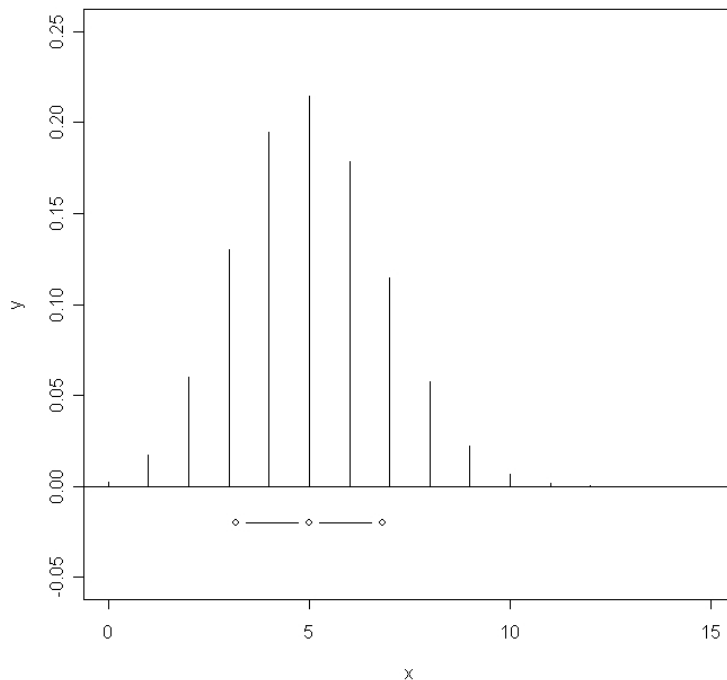


Figure 5.4. A discrete probability distribution with its expectation 5.00 under the distribution together with one standard deviation 1.82 from the expectation on both sides.

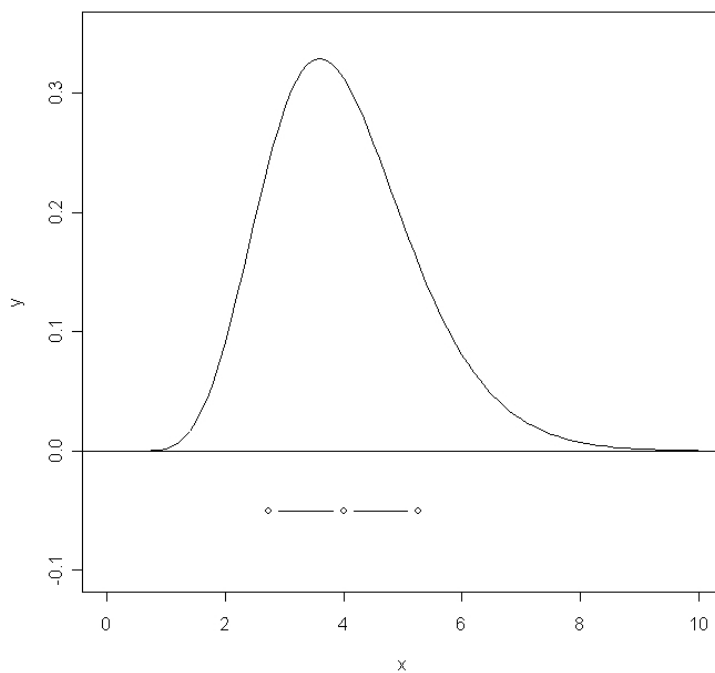


Figure 5.5. Frequency function for a continuous random variable with the expectation 4.00 together with standard deviation 1.27 on both sides of the expectation.

For an empirical (real) series of observation we have location and dispersion measures, which are similar to the theoretical ones. If the observations are measure is the ordinary mean $x_1, x_2, x_3, \dots, x_n$, a suitable location

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{k=1}^n x_k.$$

As you can see I have denoted it by an x with a bar above it, which is a common notation for a mean. For dispersion measure, we start with a definition of the empirical variance

$$s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2,$$

which is a kind of square mean deviation for the real observations. It may seem curious to have the denominator $n - 1$ and not the full number n . An explanation for this will appear later.

Even the smallest pocket calculators have a simple direct calculation for empirical means and variances. Then one also gets the empirical standard deviation $s = \sqrt{s^2}$, which is thus the square root of the empirical variance. As in the theoretical case, this standard deviation is of the same dimension as the observations.

So what does a typical case look like? In the figure below, there are three observation series. Under each of them are marked the mean and the mean plus and minus one empirical standard deviation.

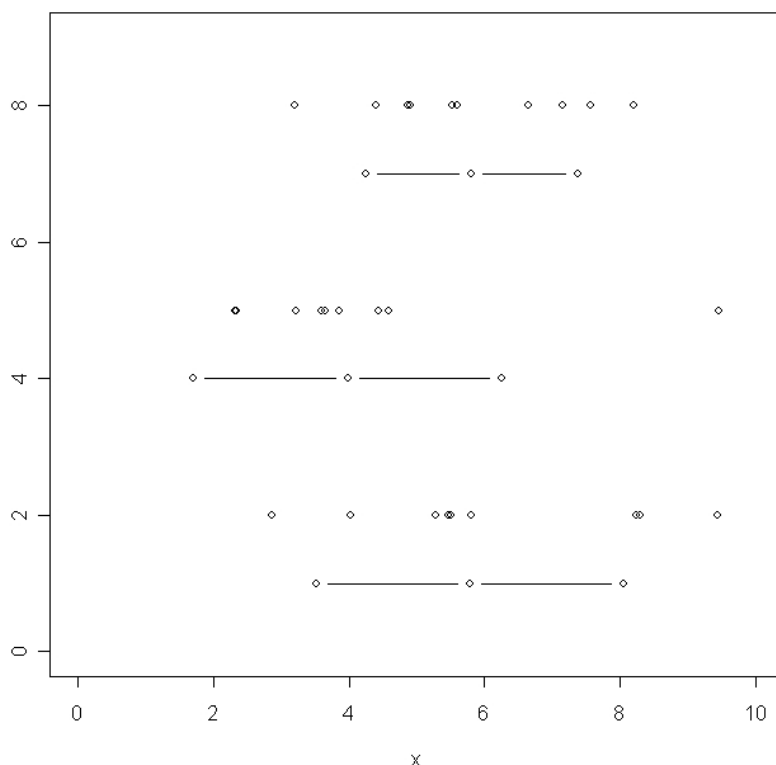


Figure 5.6. Three different observation series completed with mean and mean plus minus empirical standard deviations under the data sets.

Nido

- Luxurious accommodation
- Central zone 1 & 2 locations
- Meet hundreds of international students
- BOOK NOW** and get a £100 voucher from voucherexpress

Nido Student Living - London

Visit www.NidoStudentLiving.com/Bookboon for more info.

+44 (0)20 3102 1060



The first series is a typical normal series, which is rather symmetric and well kept together. The middle series is skewed to the right with one very distant observation on that side. This has implied a big standard deviation too. The third series is more symmetric than number two, but has a different character from number one. There are both some scattered observations with tails in both directions and some concentration in the middle.

It is on purpose that I have added the word empirical to variance and standard deviation quite often in the text here. You may think that it is too much, but it has been done in order to really point out the difference between empirical and theoretical measures. It is unfortunate that variance and standard deviation are used as names for deviance measures in both cases. The reason for this is historical. They were already named a hundred years ago.

Now we will dwell on the randomness in this connection. Suppose we have n independent observations, which are due to random variation and have the same distribution. We denote n those observations by big letters $X_1, X_2, X_3, \dots, X_n$. Their mean is denoted with $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$ and their empirical variance with $S^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2$

Expectations in trials with random variations can not only be defined for the basic observations but also for functions of these. If we denote expectation by E , it is possible to calculate mathematically that for any distributions of the basic observations we have generally

$$E[\bar{X}] = \mu \quad \text{and} \quad E[S^2] = \sigma^2,$$

where μ and σ^2 are the expectation and the theoretical variance in the distribution of the basic observations.

This means that our empirical quantities determined in the series of observation do not have any systematic errors. Randomly there are deviations from the theoretical parameters. Both \bar{X} and S^2 may get values both below and above their parameters μ and σ^2 , but in the mean they have the correct position. In statistical theory it is said that they are unbiased.

This is one of the motivations for the denominator $n-1$ in the definition of empirical variance S^2 . If we had used the denominator n instead, there would be a systematic tendency of getting a value that was too small. But with the denominator $n-1$ the empirical variance is an unbiased estimate of the theoretical variance σ^2 .

6 A USEFUL ROOT

In the previous section, we saw in a simulated example how the results in samples could vary, depending on chance. The variations in that example were quite large. But there were only 10 observations, which is rather a small sample size. Intuitively you probably understand that there would be less variation if we take a larger sample. In this section we will study the importance of sample size. You will get a very simple useful role, which shows how the size of the variations depends on sample size.

Let us start with a very simple example. Suppose one has studied the risk of getting a mild adverse effect in a medication by asking 500 randomly chosen patients taking this medicine, if they have got this adverse effect. In investigations involving patients, it is natural to suppose that results for different patients are independent, and we make that assumption here. As we have mentioned before, this means in practice that the results for different patients do not influence the random variation of each other. We can consider the results for the different patients as being independent sub-trials. According to what we have said in an earlier section, we get here a binomial distribution which has two parameters, the sample size n and the probability p .

The expectation in the binomial distribution is equal to np and the theoretical standard deviation is equal to $\sqrt{np(1-p)}$. These are general theoretical results which can be mathematically deduced from the probability distribution. If, for instance, the adverse effect should have the probability 3.5%, the expected number of patients who get the adverse effect would be $np = 500 \cdot 0,035 = 17,5$ and the theoretical standard deviation would be $\sqrt{np(1-p)} = \sqrt{500 \cdot 0,035 \cdot (1-0,035)} = 4,11$. The probability distribution would have the following shape.

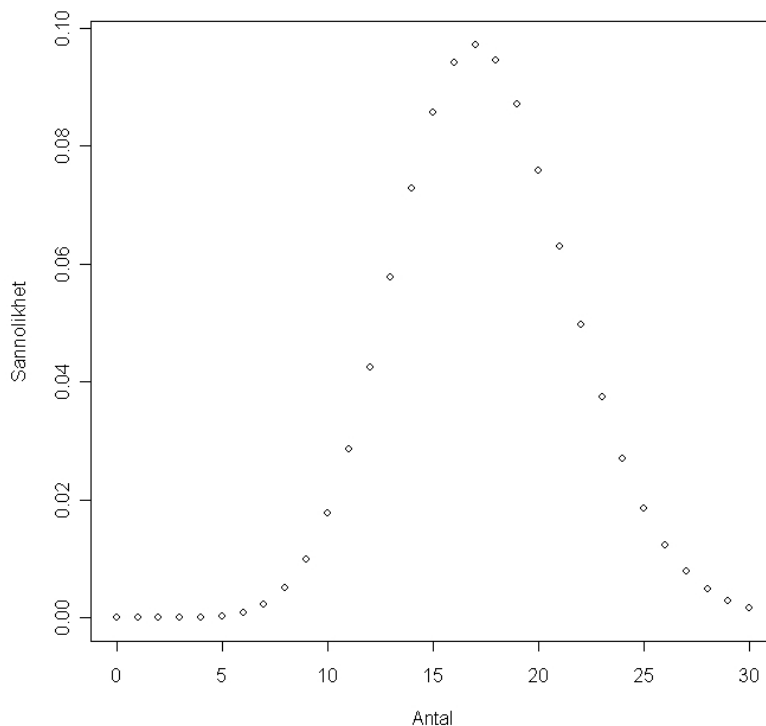




Figure 6.1. Binomial distribution with parameters $n = 500$ and $p = 0,035$. The probability for outcome is biggest in the neighbourhood of the expectation $n \cdot p = 500 \cdot 0,035 = 17,5$.

SIMPLY CLEVER


WE WILL TURN YOUR CV INTO AN OPPORTUNITY OF A LIFETIME



Do you like cars? Would you like to be a part of a successful brand?
 As a constructor at ŠKODA AUTO you will put great things in motion. Things that will ease everyday lives of people all around Send us your CV. We will give it an entirely new new dimension.

Send us your CV on
www.employerforlife.com



Now consider the problem of estimating the probability of the adverse effect, in our example. Such a probability is estimated by the relative frequency, where we divide the number of patients who got the adverse effect by the total number of patients in the investigation. Thus we have a binomial random number divided by a fixed number. Division by the fixed number gives only a change of scale. This means that since the binomial number has the expectation np and the theoretical standard deviation $\sqrt{np(1-p)}$, the relative frequency will have the expectation $\frac{1}{n}np = p$ and the theoretical standard deviation $\frac{1}{n}\sqrt{np(1-p)} = \frac{1}{\sqrt{n}}\sqrt{p(1-p)}$ after this change of scale.

We have just seen the first example of a very fundamental statistical property. When we estimate the unknown probability p , the theoretical standard deviation for the estimate is inversely proportional to the square root of the number of observations we have made. The standard deviation of the estimate, of course, decreases with the sample size. But here we have got a first example of exactly how it decreases. And such a behavioral pattern, that the estimate of an unknown parameter has a standard deviation inversely proportional to the square root of the sample size, is not limited to the estimation of probabilities. It is much more general. We will soon come back to this, but first we take a further look at our introductory example.

If the adverse effect should have the probability of 3.5%, an estimate of this probability in a sample of n individuals would have a theoretical standard deviation

$$\frac{1}{\sqrt{n}}\sqrt{p(1-p)} = \frac{1}{\sqrt{n}}\sqrt{0,035 \cdot (1-0,035)} = 0,184 \frac{1}{\sqrt{n}}.$$

For example, a sample size of 100 would give the standard deviation 0.0184 and a sample size of 400 would give a standard deviation of 0.0092. In order to make the standard deviation half as big, we must have four times as big a sample size. Intuitively you might think that it is enough to make it double, but that is wrong. In the same way, if we want to decrease the standard deviation of an estimate by a factor 10, we must increase the sample size by a factor of 100.

In real life you never know the probability p which you want to estimate, and the standard deviation depends on this unknown p . But when you have got an estimate you may get an approximate standard deviation by using this estimated p in the formula $\frac{1}{\sqrt{n}}\sqrt{p(1-p)}$.

For planning purposes you may first estimate the unknown parameter in a preliminary smaller trial, and then determine a suitable sample size in the bigger main trial, from this same formula.

I can never see a relative frequency without making a rough estimate of the standard deviation in my head. When I hear that there is an investigation of political party preferences involving 1000 citizens, a large party with 30–40% preferences will have a standard deviation of approximately

$$\frac{\sqrt{0,35 \cdot (1-0,35)}}{\sqrt{1000}} \approx 0,015 = 1,5\%$$

And a smaller party with 5–10% preferences would have a standard deviation of the order

$$\frac{\sqrt{0,075 \cdot (1-0,075)}}{\sqrt{1000}} \approx 0,008 = 0,8\%$$

Deviances between the estimate and the true value (proportion in the whole population) can easily approach two times the standard deviation. We must bear in mind that the standard deviation is a kind of mean deviation and occasional small deviations must have a compensation of occasional big deviations.

Are you surprised that there is a smaller standard deviation for smaller parties than for the larger ones? As you see above, this is true if we consider absolute deviations. But if we consider standard deviations relative to the estimates, it is the opposite. A standard deviation of 1.5% in the region of 30–40% is a relative variation of 3–5%. And a standard deviation of 0.8% in the region of 7–8% is a relative variation of around 10%.

I mentioned that one has to expect that variations of two times the standard deviation may occur for an estimate. This can be formalised. One can make confidence intervals for the parameters. How these are made in a proper way for this situation, is discussed further in a later section. For the moment we will be content with being able to estimate the standard deviation for the estimate, which is important in order to judge the quality of various presented estimates of probabilities.

As I mentioned earlier, the role of the ‘square root of sample size’ is valid in many more scenarios other than estimation of probabilities. By methods in mathematical statistics, one can show that a mean of n independent observations of the same type always has a standard deviation which is equal to the standard deviation for one observation divided by the square root of n . Often, the standard deviation for a single observation is not known, and in such a case we get an approximate standard deviation for the mean by using an empirical estimate of this basic standard deviation. Empirical standard deviations are treated earlier, as you will recall.

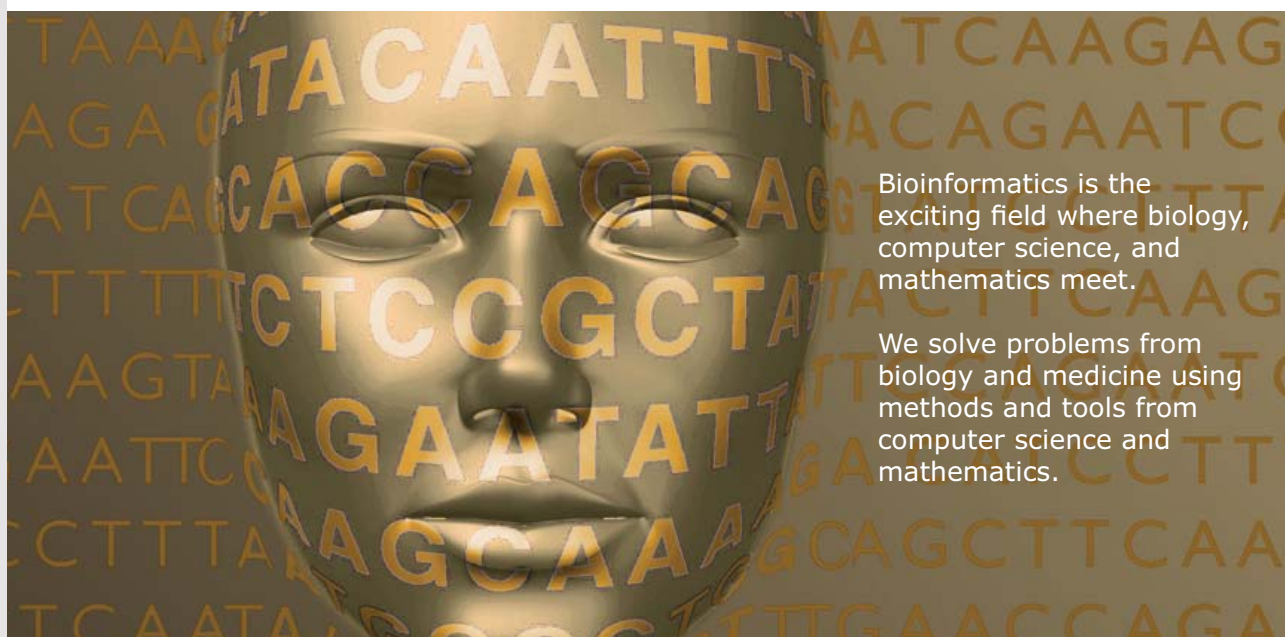
What we have studied in this section is a completely general and simple property. A mean in a series of n independent observations has a standard deviation which is equal to $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation for a single observation. And now some rules of thumb.

In order to get half standard deviation
in a mean, you must take
four times as many observations
And if you want to decrease it
ten times, you must increase the
sample size by a factor of one hundred.



UPPSALA
UNIVERSITET

Develop the tools we need for Life Science Masters Degree in Bioinformatics



Bioinformatics is the exciting field where biology, computer science, and mathematics meet.

We solve problems from biology and medicine using methods and tools from computer science and mathematics.

Read more about this and our other international masters degree programmes at www.uu.se/master

The standard deviation of an estimate is usually called standard error. It is often shortened to s.e., while the standard deviation for an observation is shortened to s.d. It is very important to distinguish between those two types of standard deviations since they differ by a factor square root of n as explained above.

This very simple ‘square root of n rule’ is exactly valid only for mean values. For more complicated cases where the parameters are not estimated by plain means, but by more complicated calculations, the simple ‘square root of n rule’ is most often approximately true. The few exceptions are some really funny cases of no practical interest. So the standard deviation of an estimate may generally be supposed to approximately have the form τ/\sqrt{n} , when n is big. The constant τ varies from one estimation situation to another. Often it may be estimated approximately in the data material, but that requires a deeper study for each separate case.

In Section 5 we had a figure which showed how a mean ‘tuned in’ towards the expectation when the sample size increased from 10 up to 1000. On the basis of what we have learnt in this section, we may understand this behavior much better. The smallest sample of size 10 has a standard error factor $\frac{1}{\sqrt{10}} \approx 0,32$ and the largest sample of size 1000 has a standard error factor $\frac{1}{\sqrt{1000}} \approx 0,032$. As you can see it fits the simulated data in the aforementioned figure very well.

7 COMPLETELY NORMAL AND ALMOST NORMAL

I hope that you have got a basic idea of the size of the standard deviation of estimates, from the previous section. Now we will study the distributions of estimates. This part is based on advanced mathematics, but I think that you will understand the basic ideas without digging deep into the mathematical theory. At least I will try to explain it in a more intuitive, non-technical way.

Carl Friedrich Gauss was a German mathematician, who lived in 1777–1855. He is one of the pioneers of the theory of probability, which forms the basis of the concepts we will now discuss. There is a probability distribution which is often called the Gaussian distribution, but is also called the normal distribution. I will use the latter name. An interesting snippet here – there was a figure of a normal frequency function as well as a picture of Gauss on old German Mark-bills of a certain value. There, one could see the well rounded church-bell-formed shape of the normal frequency function. However, we will not start with this form, but with the story of how it may appear in applications.

We may wonder what happens if we add a number of independent random results of the same type. That is what we do, for example, when we form a mean in order to estimate an expectation in a distribution. Will the sum get a distribution of the same form as the individual observations? No, it will not! I will now show you some figures of the distributions we get for the distributions of the sum of the points in two dice tosses, and the sum of the points in three dice tosses. If you are interested, you may check the distributions yourself by counting the number of cases leading to the different results.

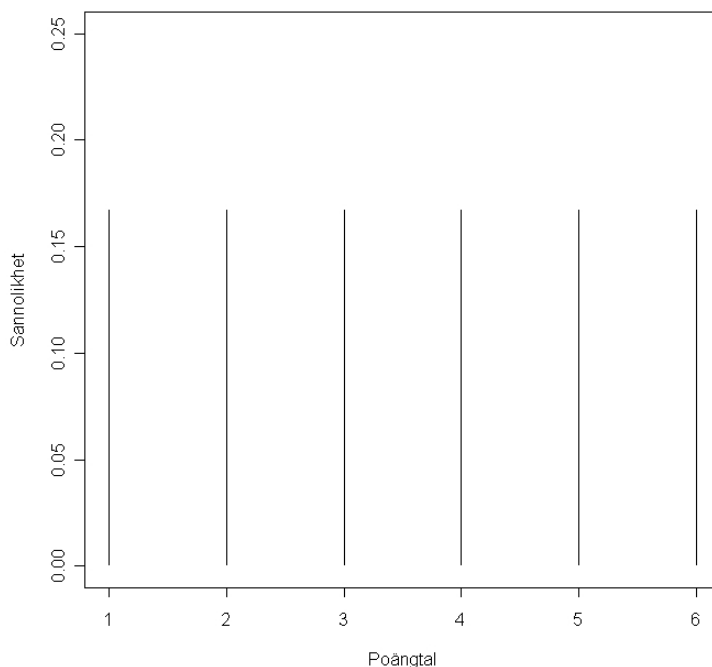


Figure 7.1. Probability distribution for the point in a dice toss.

UNIVERSITY OF COPENHAGEN



Copenhagen Master of Excellence

Copenhagen Master of Excellence are two-year master degrees taught in English at one of Europe's leading universities

Come to Copenhagen - *and aspire!*

Apply now at
www.come.ku.dk



cultural studies

religious studies

science



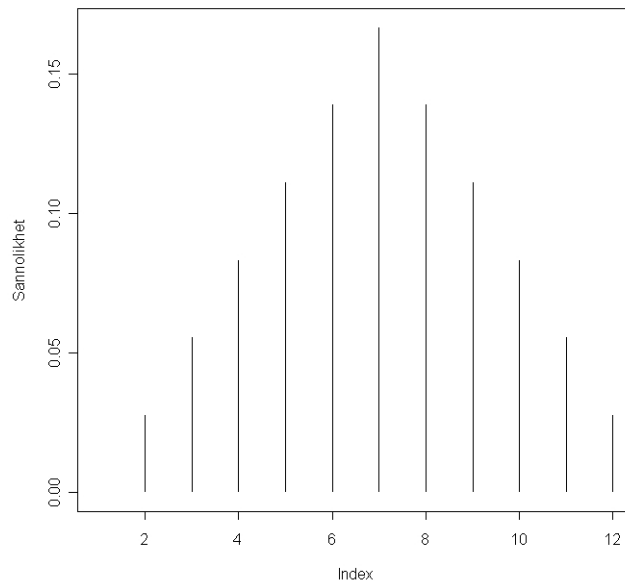


Figure 7.2. Probability distribution for the sum of the points in two dice tosses.

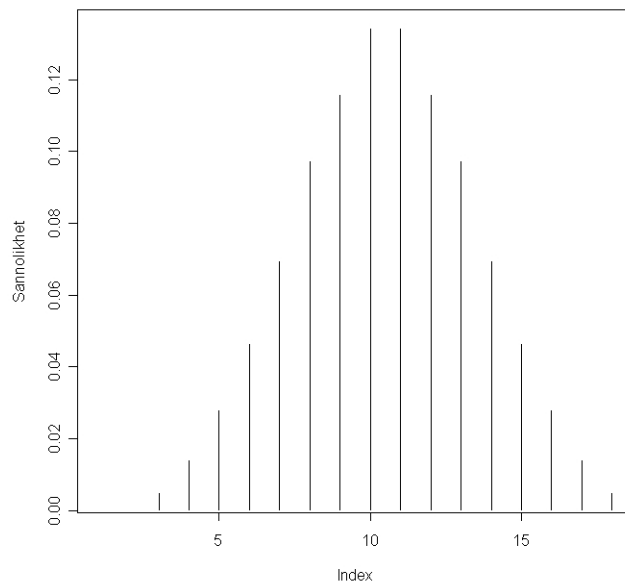


Figure 7.3. Probability distribution of the sum of points in three dice tosses.

In the figures you can clearly see how the probability distribution gets a more rounded form when we add the points in more dice tosses. Starting with the distribution for the points in one coin toss, the form of the distribution of the sum of two coin tosses is more rounded. And the form of the distribution of the sum of the points in three coin tosses is even more rounded. Then one could naturally pose the question on whether it eventually approaches a final form. And it does. The mathematical proof for this statement is deep and difficult, but the final result is not so complicated. In order to describe the final form, we need to align the cases by adjusting for position and scaling for the different sample sizes. What we do is to align the cases to a common expectation of, for example, 0 and a common standard deviation of, say 1. Then the form of the distributions will approach the following form, which has a normal frequency function where the expectation equals 0 and the standard deviation equals 1. Here is a figure with the final form.

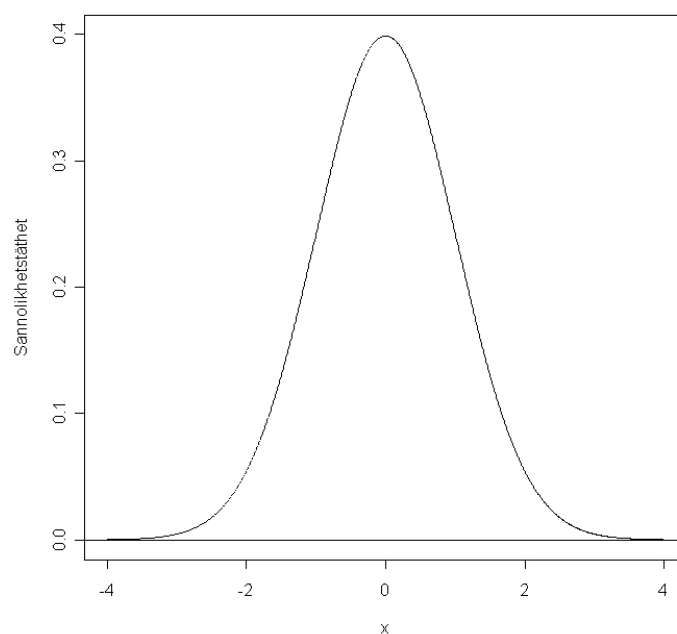


Figure 7.4. Frequency function for a normal distribution with expectation 0 and standard deviation 1.

The probabilities for outcomes in different intervals are given by the area under the curve between the limits of the interval. The mathematical form of the curve is not so essential since there exist tables both for the frequency function and the cumulative distribution function, which is the area under the curve to the left of different points. Modern statistical programs also easily produce probabilities for outcomes in any intervals. But for the sake of completeness, I now give you the mathematical form of the curve:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

The normalised normal distribution is a special case of normal distribution, which has the expectation 0 and the standard deviation 1. A normal distribution with arbitrary parameters μ and σ has the same form but with the centre point in the expectation μ and a scale factor equal to the standard deviation σ .

Now we have got a very nice situation. For a sample of size n , the mean of independent observations of the same type will have an approximate normal distribution with expectation parameter μ equal to the expectation in the basic distribution for the observations and a standard deviation $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation for a single observation. The n only assumption that we have independent observations of the same type is enough to get this important result which gives significant information on the properties of the mean. Some of these properties will be used when we study interval estimation methods in a later section.



Brain power

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations. Therefore we need the best employees who can meet this challenge!

The Power of Knowledge Engineering

Plug into The Power of Knowledge Engineering.
Visit us at www.skf.com/knowledge

SKF

How good is the normal distribution as an approximation of the distribution of a sum of independent random contributions? That depends very much on the sample size and the distribution of the single terms in the sum. If they have a rounded and reasonable symmetric distribution, a rather small number of terms are needed in order to get a good approximation. But if the single terms have a very skewed distribution, many more terms may be required.

If we consider, for instance, the sum of six independent random numbers which individually are uniformly distributed in the interval $[0;1]$, it would not be possible to distinguish between the exact and the approximate frequency functions if I were to put them in the same figure in the book here. Neither on the screen nor in a paper printout.

On the other hand, if the terms had a frequency function similar to that in the first of the following figures, there will be quite a difference between the exact distribution and the normal approximation, as you can see in the second figure.

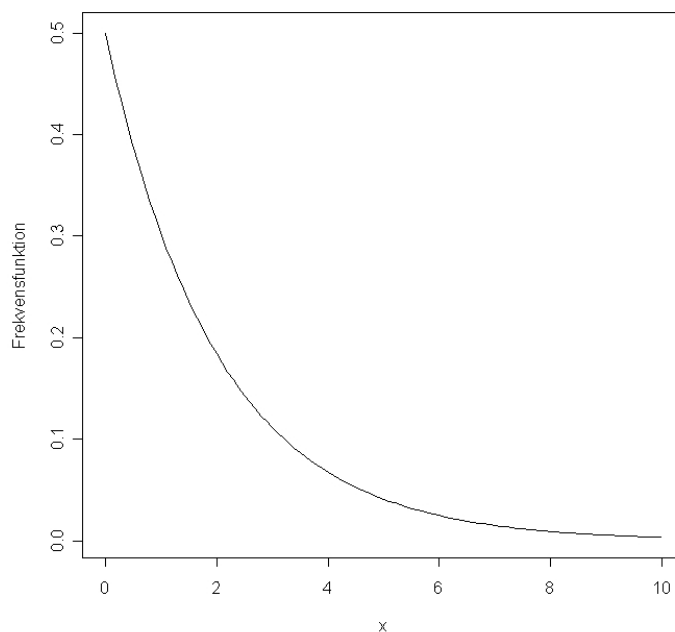


Figure 7.5. Example of a skewed frequency function for a single term.

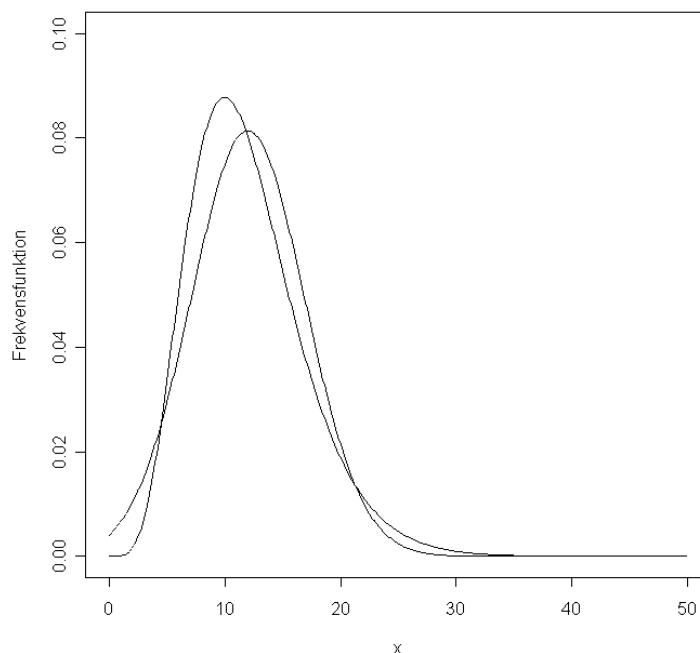


Figure 7.6. Exact frequency function for the sum of six independent terms with the above skewed distribution (the top a little higher and to the left) and the approximating normal frequency function (the top a little lower and to the right).

Thus we can say that the normal distribution may be a good approximation for the distribution of a sum of random contributions of the same.

When it concerns the mathematical form of the normal distribution you do not need to worry. Using statistical programs you can get values of the cumulative distribution function, which for an argument x gives the probability to the left of x . It can not only be obtained for the standard case with parameters 0 and 1, but also for arbitrary parameters. For instance, with the program R the probability to the left of x in a normal distribution with parameters a and b will be given by the command `pnorm(x, mean=a, sd=b)`.

In earlier times, numerical values for the probabilities were available only as tables for the standard normal distribution with parameters 0 and 1. If we denote this cumulative distribution function by Φ , and if we denote by F , the cumulative distribution function in a general normal distribution with expectation μ and standard deviation σ , we would use the transformation

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

Suppose we have a sample of 50 persons, where the result for each person is the number of yes-answers to five related yes-no-questions, and the result is the following

1 0 2 2 0 3 5 3 4 2 3 5 4 2 3 2 3 0 3 2 4 2 3 2 3 3 3 2 4 4 4 2 3 1 3 3 2 12 2 3 0
4 2 4 4 0 4 4 3

The material has a mean of 2.60 and a standard deviation 1.29. With our present knowledge, we may now estimate the distribution of the mean of the results. It ought to be an approximate normal distribution with the parameters 2.60 and $\frac{1,29}{\sqrt{50}} = 0,182$. Its approximate frequency function has the following shape.

Trust and responsibility

NNE and Pharmaplan have joined forces to create NNE Pharmaplan, the world's leading engineering and consultancy company focused entirely on the pharma and biotech industries.

Inés Aréizaga Esteva (Spain), 25 years old
Education: Chemical Engineer

– You have to be proactive and open-minded as a newcomer and make it clear to your colleagues what you are able to cope. The pharmaceutical field is new to me. But busy as they are, most of my colleagues find the time to teach me, and they also trust me. Even though it was a bit hard at first, I can feel over time that I am beginning to be taken seriously and that my contribution is appreciated.



NNE Pharmaplan is the world's leading engineering and consultancy company focused entirely on the pharma and biotech industries. We employ more than 1500 people worldwide and offer global reach and local knowledge along with our all-encompassing list of services.
nnepharmaplan.com

nne pharmaplan®

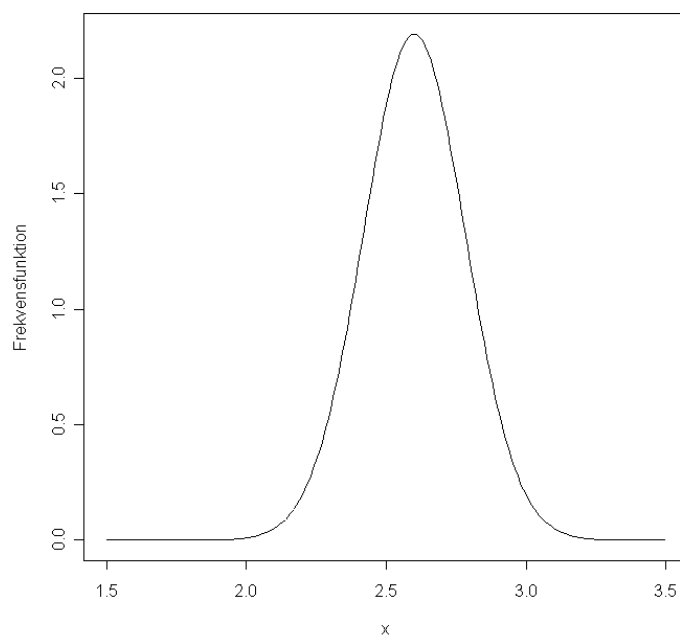


Figure 7.7. Estimated frequency function for the mean in an empirical material with 50 observations.

We ought to remember here, that this is just an estimate based on empirical observations in a sample of moderate size. There are possible random errors in location, dispersion and form of the distribution. However, it is possible to get reasonably good approximations of the probabilities for outcomes in different intervals.

Consider for example, the interval 2.40 to 2.80. If we transform this down to the standard normal distribution function with parameters 0 and 1, these values correspond to $-0.20/0.182 = -1.10$ and $0.20/0.182 = 1.10$. The left interval end point is 1.10 standard deviations below the expectation and the right interval end point is 1.10 standard deviations above the expectation. From a table or a statistical program, we get the value of $\Phi(x)$ for the argument $x = 1.10$ equal to 0.864. Above the point $x = 1.10$ we thus have the probability $1 - 0.864 = 0.136 = 13.6\%$. There is the same probability 13.6% below the left interval end point. Thus we estimate the probability outside the interval from 2.40 to 2.80 to be $2 \cdot 13.6\% = 27.2\%$, and we estimate the probability of outcome inside the interval to be 62.8%. But please remember...these probabilities are just estimates and the sample size here is not overwhelmingly large.

When I introduced the normal distribution, I projected it as the limit form of distribution for sums of independent random contributions of the same kind. In mathematical statistics, this is called the central limit theorem. It can be generalised in two respects for practical importance.

Firstly, there are generalisations meaning that the terms in the sum do not need to have exactly the same distribution. In practice it is enough that the different terms shall be so similar that a single one or a few of them do not dominate the sum too much. Secondly, there are generalisations meaning that in practice we will get an approximate normal distribution of the terms, even if there is some dependence, which is not dominating the whole series. For instance this may be a time series with dependence only between units at a short time distance. Thus to conclude:

When a random variable consists of a
number of additive effects of roughly the
same size and small dependence,
it is reasonable to suppose
that it is normally distributed.

But one cannot just assume that all observations are normally distributed. One must always consider what is a reasonable model and think of the assumptions in the approach. If I look back at statistical practice in the last century, I would rather say that the normal distribution has been used too much than say that it has been used too little. In the following chapters, you will now and then be reminded of situations where it is not suitable to use the assumption that observations are normally distributed.

8 WITHIN THE ERROR MARGIN

Now and then there are reports on TV or in the print media on investigations of the preferences among political parties. And often one comes across a commentator talking about a change within the error margin. It was not so in olden times, when there would have been just a report of the numbers and no talk of errors at all. The figures would be presented as if they were perfect results without any randomness. I am happy that the world has become better, at least in this respect. There is now an awareness of random variations and the journalists like to talk about error margins. This present section will include some thoughts on the concept of error margins.

Error margins have a lot to do with confidence intervals, which we discussed in a simple form earlier. And we start this section too with a simple form of confidence intervals for an idealised situation.

This e-book
is made with
SetaPDF



PDF components for PHP developers

www.setasign.com



Let us suppose that we have independent observations which may be assumed to be normally distributed with unknown expectation μ , but with known standard deviation σ . This last condition is perhaps not very realistic. The natural situation is, of course, that the standard deviation is also unknown. But we make this assumption here in order to get a simple introductory example.

For a series $X_1, X_2, X_2, \dots, X_n$ of independent observations of this kind, their mean \bar{X} will be normally distributed with parameters μ and $\frac{\sigma}{\sqrt{n}}$. The mean thus inherits its normality from the single observations, but the standard deviation is much smaller, according to the ‘square root rule’. Now we try to construct a confidence interval for the parameter μ , by choosing a suitable value for the constant a in $\left[\bar{X} - a \frac{\sigma}{\sqrt{n}}; \bar{X} + a \frac{\sigma}{\sqrt{n}} \right]$. One can say here that we have put safety margins of the size $a \frac{\sigma}{\sqrt{n}}$ on each side of the estimate \bar{X} . The safety margin is a constant a times the standard error s.e.

What is the risk that such an interval misses the parameter on the right side? The probability we want is thus written as $P\left(\bar{X} - a \frac{\sigma}{\sqrt{n}} > \mu\right)$. Since the mean is normally distributed with parameters μ and $\frac{\sigma}{\sqrt{n}}$, we can get this probability from the normal distribution. Using a transformation to the standard normal distribution we have,

$$P\left(\bar{X} - a \frac{\sigma}{\sqrt{n}} > \mu\right) = P\left(\bar{X} - \mu - a \frac{\sigma}{\sqrt{n}} > 0\right) = P\left(\bar{X} - \mu > a \frac{\sigma}{\sqrt{n}}\right) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > a\right)$$

From a table or from a statistical program we can see that in this standard distribution, 97.5% of the probability mass is to the left of 1.96. The risk of missing, on the right side, is thus bounded to 2.5% by choosing $a = 1,96$. The risk of missing, on the left side, will also be 2.5% with this choice. The interval

$$\left[\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}}; \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}} \right]$$

constitutes a confidence interval for μ with a confidence degree 95%. This is now an exact 95% confidence interval in the idealised situation with normal observations and a known standard deviation σ .

You may have noticed that I have used uppercase letters in the notations, both for the observations and their mean in the discussion here. This is to indicate that there is some randomness involved and we can talk about the probabilities of hitting and missing. Once we have got the numerical values for the observations, there is no randomness left and we cannot talk of probabilities any more.

If the interval limits have the outcomes as 23.2 and 28.5, we can make the statement

$$23,2 \leq \mu \leq 28,5 \quad 95\%$$

which indicates that the interval is constructed by a method which has a 95% probability of hitting (success). But it is not really correct to say that the probability for this numerical statement is 95%.

One could well call $1,96 \frac{\sigma}{\sqrt{n}}$ error margins for the estimate. In this simple case it is an exact error margin. In other more realistic situations, where σ is not known, one can get approximate error margins by estimating the theoretical standard deviation σ with the empirical standard deviation S , calculated in the observation series, or some other empirical estimate of it. Quite often, when talking of approximate error margins the table value of 1.96 is rounded off to 2, that is, one uses 2 times the standard error s.e. as an error margin.

For large sample sizes and for most types of observations this usually works well and one gets approximately 95% confidence degree. For small sample sizes it may perhaps not work so well.

In a later section we will study confidence intervals for the expectation in a normal distribution with an exact confidence degree even for the case of unknown theoretical standard deviation. But for now we will first look at some examples with simple approximate error margins.

Example

An investigation on preferences for various political parties was based on interviews with 1142 persons. It turned out that 371 of them preferred a special party. Then this party is estimated to have $\frac{371}{1142} = 0,345 = 34,5\%$ of the whole population. The variance in this estimate is computed as $\frac{371}{1142} = 0,345 = 34,5\%$, according to formulas for binomial distributions. This gives an error margin of $2 \cdot \sqrt{0,0001979} = 0,028 = 2,8\%$. The sample size is rather big and the approximation can be assumed to be good.

Example

In a certain production cycle, there are sometimes interruptions. In order to get clear information on how often this happens, one has measured the time (in unit days) for ten time intervals between interruptions. The result was:

3.08 3.29 1.87 1.85 2.65 1.27 5.06 23.68 12.12 2.93

Here the mean is 5.78 and the empirical standard deviation in the series is 7.02. The approximate error margin in the estimate 5.78 would then be $2 \cdot \frac{7,02}{\sqrt{10}} = 4,44$. If we subtract and add this error margin to the estimate, we get the approximate 95% confidence interval limits equal to 1.34 and 11.46.



Sharp Minds - Bright Ideas!

Employees at FOSS Analytical A/S are living proof of the company value - First - using new inventions to make dedicated solutions for our customers. With sharp minds and cross functional teamwork, we constantly strive to develop new unique products - Would you like to join our team?

FOSS works diligently with innovation and development as basis for its growth. It is reflected in the fact that more than 200 of the 1200 employees in FOSS work with Research & Development in Scandinavia and USA. Engineers at FOSS work in production, development and marketing, within a wide range of different fields, i.e. Chemistry, Electronics, Mechanics, Software, Optics, Microbiology, Chemometrics.

We offer
A challenging job in an international and innovative company that is leading in its field. You will get the opportunity to work with the most advanced technology together with highly skilled colleagues.

Read more about FOSS at www.foss.dk - or go directly to our student site www.foss.dk/sharpminds where you can learn more about your possibilities of working together with us on projects, your thesis etc.

The Family owned FOSS group is the world leader as supplier of dedicated, high-tech analytical solutions which measure and control the quality and production of agricultural, food, pharmaceutical and chemical products. Main activities are initiated from Denmark, Sweden and USA with headquarters domiciled in Hillerød, DK. The products are marketed globally by 23 sales companies and an extensive net of distributors. In line with the corevalue to be 'First', the company intends to expand its market position.



Dedicated Analytical Solutions

FOSS
 Slangerupgade 69
 3400 Hillerød
 Tel. +45 70103370
www.foss.dk



The error margin in the small investigation was considered to be so large that it was decided that a bigger investigation was to be made. In a new series with 100 observations of time distances one got the mean 3.37 and the empirical standard deviation 2.94. This led to an approximate error margin of $2 \cdot \frac{2,94}{\sqrt{100}} = 0,59$ and the limits of the approximate 95% confidence interval were now 2.78 and 3.96. Much better!

There are two problems with using approximate error margins in observation series that are too small. First, the normal distribution may be a bad approximation of the real distribution. And second, the empirical standard deviation may be a bad approximation of the theoretical one.

In the last example above, the reasonably good estimate of the theoretical standard deviation in the large series was 2.94, while the not-so-good estimate in the small series was 7.02. Quite a difference! But of course in a very small series the estimate can occasionally deviate much, up or down. And this implies that in (very) small series one can occasionally get either rather worthless long intervals or short intervals which do not hit the parameter.

Coming back to the investigations on the political party preference, when the results are presented, the changes from earlier investigations are often discussed. In the example above, we had an estimate of 34.5% for a certain party. This was based on the observation of 371 preferences in the group of 1142 individuals. Now suppose the same party had got 335 preferences in a group of 1063 individuals in a previous investigation. This gives an estimate of 31.5%. And the political journalist, who has learned his lesson about error margins, declares that this is outside the error margin. Error margin in this region of sample size and probability is about 2.8% and the difference here is 3.0%. Unfortunately he has only attended the first lesson on error margins and not lesson number two. I will now explain what he could have learned in that next lesson.

It is correct that an error margin for one estimate in this region is approximately 2.8%. But both the compared estimates have random errors. If there are different samples at the two times of investigation, it is reasonable to assume the two estimates to be independent. There exists some general theory in mathematical statistics on properties of variances, which we do not discuss in detail here. Among the results is the property that the variance of the sum or difference of two independent random variables equals the sum of the individual variances for the two random variables. Here we have two estimates, each one with the variance (=square of standard deviation!) approximately equal to $0,028^2 = 0,000784$. Thus their difference will have an approximate variance of $2 \cdot 0,000784 = 0,001568$. This gives a standard deviation for the difference approximately equal to $\sqrt{0,001568} = 0,0396 = 3,96\%$. As you can see, the error margin for the difference is approximately 40% higher than the error margin for the individual estimates. That margin ought to be used when differences are discussed. Unfortunately, it is not always the case in all presentations of party preference investigations.

The error margin in the difference
between two independent estimates
is a bit higher than the error margin
in the individual measurements.

If, on the other hand, it is a comparison of one empirical estimate and a fixed limit, it is correct to use the error margin for a single estimate. There is no random variation in the fixed limit. A political party with a preference estimate in the region 6–7% in an investigation with a sample size around 1000 would have an error margin of about 1.6%. If 4% is a critical limit for getting representation in the parliament at all, a party with 7.0% preference can be declared safe for getting representation. An approximate confidence interval of 95% would be situated above the 4% limit.

9 PURE JURIDICAL MATTERS

Tests, statistical tests, hypothesis testing...a beloved child has many names. And they are used in many different types of applications. While the ideas behind the methods are perhaps not so difficult to understand, it is easy to misinterpret some concepts. So let me present the basic philosophy quite carefully, without delving too deeply into technical details.

Performing a statistical test in a practical situation means, to first formulate some kind of hypothesis to be tested. One then has to see if the result of the trial is in accordance with the stated hypothesis or if the result of shows that the hypothesis has to be rejected. The procedure will end with a statement and it is important to get it right with possible statements and risks of making wrong statements. Those are quite intricate details.



"I studied English for 16 years but...
...I finally learned to speak it in just six lessons"

Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download

While formulating a statement, you really want it to be correct, and besides that, you also want other people to understand and accept the statement as correct. Suppose for instance, that you believe that a certain drug has a positive medicinal effect. If you collect data in a medical trial and then make the statement that there is a positive effect, you must be able to defend that statement against any opponents who claim that there is no effect or perhaps even that there is a negative effect of the drug. The burden of proof is on your shoulders. You have to have enough power to eliminate the possibility of the opposite alternative to be reasonable.

How can this be done in a responsible way? We have to face the fact that there is a random variation in the observations, depending on which individuals are included in the investigation and all sorts of intermittent effects, which influence the result. And yet we must be confident with our statements. What we can do, is to use a method with a very small risk of making a wrong statement, whatever the true parameters may be.

The technical way to achieve this is to consider the hypothesis that there is no effect or negative effect, and then test this hypothesis. The result of the test may either be that the hypothesis is rejected or that it cannot be rejected. The fundamental security precaution is now to have a decision rule such that there is a very small probability of rejecting the hypothesis by pure chance **if it should be true**. This small probability is called the level of significance for the test, and it is chosen in advance as a small value. The value is also declared in connection with the rejection statement.

When we have been able to reject the hypothesis that there is no effect (or negative effect), and thus claimed that there is a positive effect, the statement is defensible in the following way:

Each person who doesn't want to accept your statement thus holds the opinion that there is no effect or perhaps even a negative effect. But if he (or she) would be right, there would be just this certain small little probability, the level of significance, that I by pure chance would be able to make my statement of positive effect. Purely juristical matters! But it is all about the burden of proof.

Best would of course be if one could make a statement with 100% safety, that is with a level of significance 0%, but in general this is not possible since there is always randomness involved.

Now observe carefully, that if you have not been able to reject the hypothesis, there is no safety protection for a statement that the hypothesis is true. No possibility to defend oneself against an opponent believing the opposite there. Say nothing at all or say that you have not been able to reject the hypothesis.

In order to 'prove' that there is a positive effect,
 you need to test the hypothesis
 that the effect is zero or negative.
 If you can reject that hypothesis
 with a small level of significance
 it is defensible to state
 that there is a positive effect.

Example

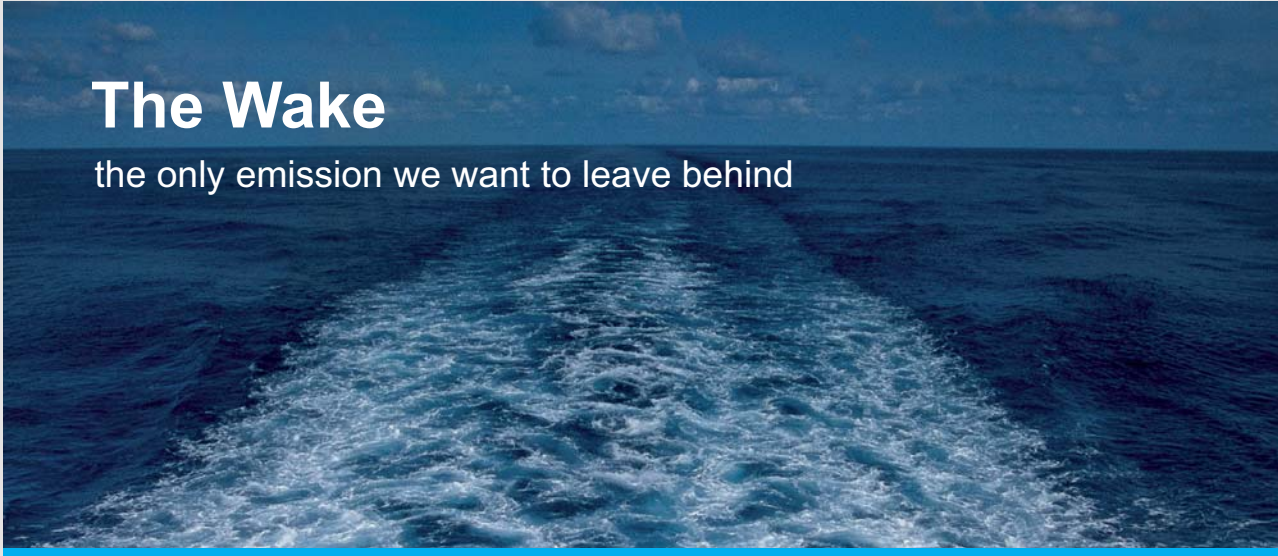
In an investigation on the effect of a certain food supplement on the content of a trace substance in the blood, the content of trace substance has been measured before the treatment period in 16 randomly chosen persons. After having eaten the food supplement during a given time, these persons have again had the trace substance measured in the blood. As data for further analysis, the ratio of the contents after and before, is first calculated. This may be a reasonable way to handle the problem since an individual variation in the content of trace material can be expected. The ratio results were:

0.89	1.31	1.08	1.03	1.00	1.13	1.30	1.46
1.46	1.77	0.79	1.38	1.75	0.89	1.30	0.85

The mean in the series is 1.212 and the standard deviation is 0.305. With the use of approximate normal distribution, a test is made with an approximate level of significance. The hypothesis here is that the expectation of the ratio is 1.00 or less. If the value 1.00 should be correct, the random variable $\frac{\bar{X} - 1,00}{S/\sqrt{16}}$ should be approximately normally distributed with parameters 0 and 1. On the approximate significance level 2.5% the hypothesis could be rejected if there was an outcome of this variable above 1.96, which is the point in the normal (0,1) distribution that has probability 2.5% to the right. If the expectation would be less than 1.00, the probability to reject is even smaller. The outcome for the test statistics was 2.78. Thus the hypothesis that the expectation is less than or equal to 1.00 may be rejected, and it can safely be stated that the expectation is more than 1.00. Thus it is defensible to state that the content of the trace substance shows a general increase when the food supplement is used on the chosen level of significance 2.5%.

There are different traditions within different application fields, and also to some extent different traditions within different countries when it comes to the handling of significance levels. In many cases one does not want to use a fixed level of significance, but works with what could be called an ‘obtained level of significance’. That is the level of significance which would be required in order to just be able to reject the hypothesis with the outcome obtained. This probability is often called just p value. If the hypothesis is true, the p value is random with a distribution which is uniform in the interval (0;1). Thus all points in this interval are equally probable as a random outcome of p if the hypothesis is true. If the hypothesis is not true, there is an increased probability for small outcomes of p.

If we want to use the p value method in the above example, we have to calculate the probability that the test statistics by pure chance would get an outcome of at least 2.78 if the hypothesis were true. According to a statistics program or table, this p value is 0.0027 or 0.27%.



The Wake


the only emission we want to leave behind

Low-speed Engines Medium-speed Engines Turbochargers Propellers Propulsion Packages PrimeServ

The design of eco-friendly marine power and propulsion solutions is crucial for MAN Diesel & Turbo. Power competencies are offered with the world's largest engine programme – having outputs spanning from 450 to 87,220 kW per engine. Get up front! Find out more at www.mandieselturbo.com

Engineering the Future – since 1758.

MAN Diesel & Turbo



Two comments are in place here. First, we have used an approximate method by using the empirical estimate S for the theoretical standard deviation σ . That can be overcome. In the next section you will see how one can make an exact test when σ is not known, when the observations have a normal distribution. Using this method would instead give the obtained significance level (p value) equal to $0.0070 = 0.70\%$. We will come back to this method in the next section. Second, it might perhaps be better to make the analysis for logarithms of the ratios instead of the ratios themselves. Then the natural hypothesis would be that the expectation in that case is zero or negative. Positive random variables, like the ratios here, often have distributions which are skewed to the right. Taking logarithms instead of using the values themselves may often make the distributions more symmetric and thus more 'normal-like'. This is often suggested for positive observations with a big relative variation.

We may also use the previous example for a little further discussion. You may have observed that 2 of the 16 persons in the trial got lower content after having used the food supplement, than they had before. This does not necessarily mean that the food supplement gives less content of the trace material in some people. In practice, almost all medical measurements (and for that matter also many other measures) have a natural random variation in time. So if the food supplement had no effect at all, all individuals would get a pure random increase or decrease between the measurements – about half of them up and about half of them down. We could make a test based only on the number of increases and decreases. If the number of increases is considerably more than half the total number of individuals, it is an indication of a positive effect. If there were no real effect, the number of increases would have a binomial distribution with parameters $n = 16$ and $p = 0,50$. The obtained level of significance, that is the p value, is the probability of getting at least 12 increases in that distribution. Using a statistical computer program or a table we find that this probability is equal to $0.0384 = 3.84\%$.

We got a larger p value here than in the other test. And small p values give good arguments in the discussion with a possible 'opponents' as I explained earlier. But remember that there is randomness in everything we do here. So it does not necessarily mean that this first test is better than the second one in general. This is just what happened in one single numerical example. What you definitively are not allowed to do, is to make several tests and choose the one that is most favorable for you. That is cheating! The technical calculated p values in the separate test are not valid if you commit this statistical crime. But before you see the data you may well, of course, consider what is a good test and then run just this test and nothing else.

The different sizes of p values for the two tests' statistics we discussed in the previous example lead us to the question of comparing the properties of different tests and to define quality measures for tests.

You may ask: What is the probability that a test gives rejection of a hypothesis in cases when it is false? That depends on 'how false the hypothesis is'. Consider now the second test above. The rejection probability is determined by the sample size and the real value of the probability for increased content of trace material in the blood. Suppose we make a test with a fixed level of significance by rejecting the hypothesis if there are 13 or more persons with increased value. The level of significance for this test is 1.1%, which can be calculated by a statistical program.

The probability of rejection is obtained if we calculate the probability $P(Z \geq 13)$ when Z has a binomial distribution with parameters $n = 16$ and p . It will be a function of p , which is called the power function of the test. It is given in the following figure for this example.

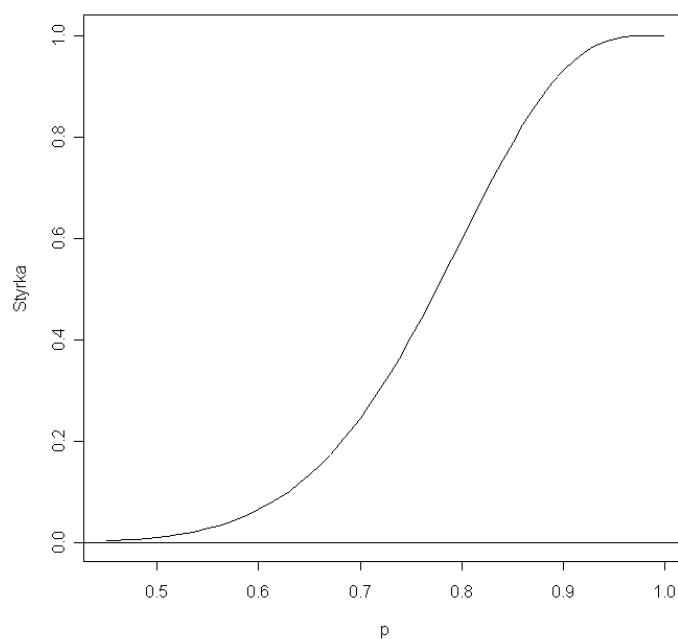
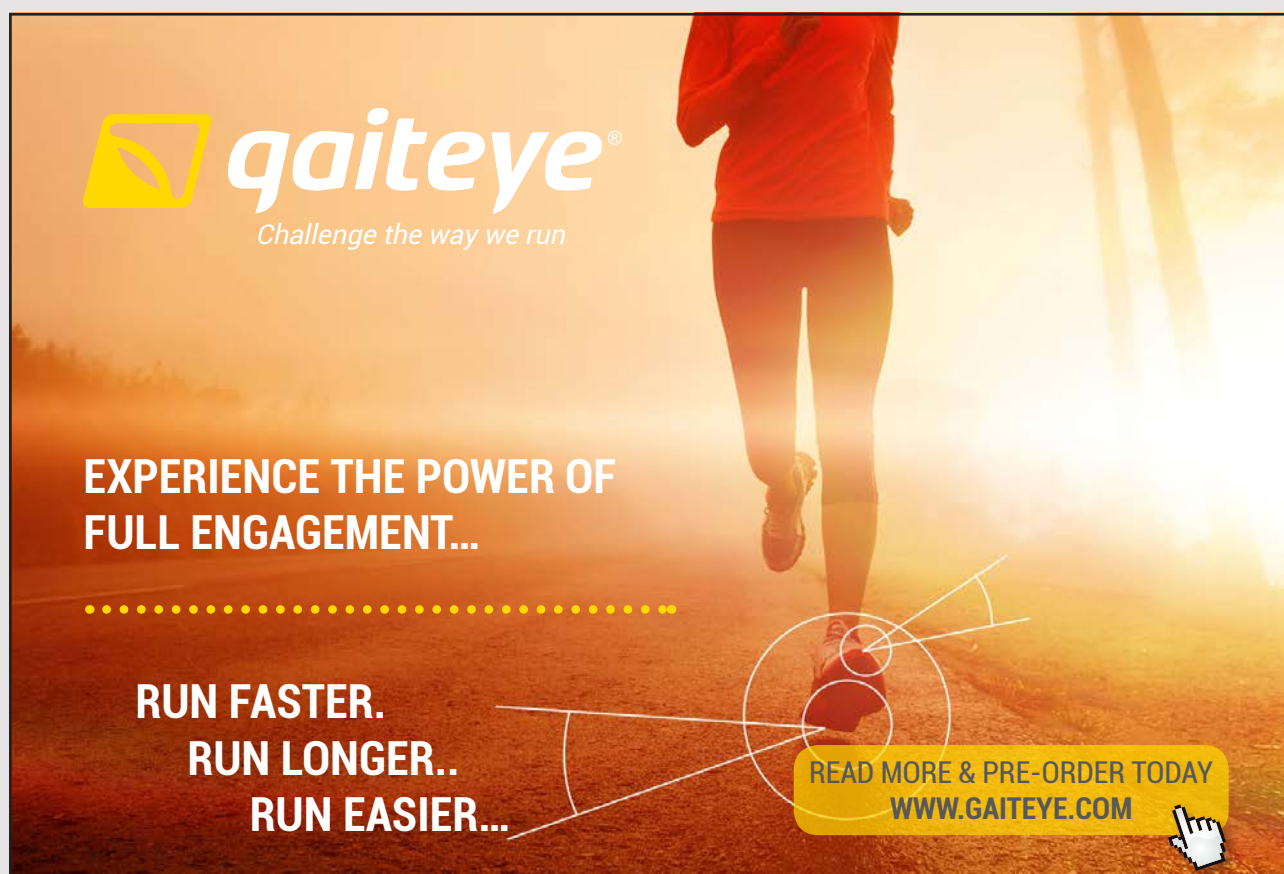


Figure 9.1. Power function for a simple test. Probability of rejection on y-axis and probability of failure on x-axis.

The value of the power function in a point is called the power for the alternative given by that point. A natural desire is of course, to have a test with a large power for different alternatives. But very close to the hypothesis it is not possible to get a very high power without using an extremely large sample size. Just at the boundary of the hypothesis the power function approaches the level of significance, and within the hypothesis it is always below or at the level of significance. Far away from the hypothesis it will approach 1.00. You can see in the above figure that if the probability of increased blood content of the substance for a single (randomly chosen) individual is around 0.8 we get approximately 50% power. The probability of discovering that the hypothesis of constant or decreased content is wrong, increases from rather moderate values at $p=0.7$ up to almost 100% at $p=0.9$. As you can see the power function reveals quite well what the test can do. If we increase the sample size, the power function will be steeper. As you can see, power calculations are good means for planning of experiments.

By way of the examples in this section, we have seen one approximate method and one simple method based on the number of positive differences only. This will suffice as illustration for the moment. In the next section you will see exact methods for normally distributed observations with unknown theoretical standard deviation.



gaiteye[®]
Challenge the way we run

**EXPERIENCE THE POWER OF
FULL ENGAGEMENT...**

.....

**RUN FASTER.
RUN LONGER..
RUN EASIER...**

**READ MORE & PRE-ORDER TODAY
WWW.GAITEYE.COM**

10 SOME OLD CLASSICS

Let us go back about 100 years in time. The statistical theory is still in its first phase of development. And since one has realised that the normal distribution for observations is reasonable in many cases, it is natural that there is an interest in developing a theory for statistical methods in that situation. Now we will here have a broader look at the simplest of these methods, which has stood the test of time.

The easiest way to present the different methods is to start with the methods for confidence intervals and tests related to standard deviations for observations. Thus suppose we have n independent observations which have a normal distribution with some unknown parameters μ and σ , and let the empirical standard deviation in the observation series be S , defined by $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ as before.

By pure mathematical methods it can now be shown that the distribution of the random variable $(n-1) \frac{S^2}{\sigma^2}$, depends only on the number n of observations. That distribution is called the χ^2 distribution (chi-square distribution) with $r = n - 1$ degrees of freedom. For instance the density function for the χ^2 distribution with 10 degrees of freedom and the χ^2 distribution with 20 degrees of freedom, have the following shapes.

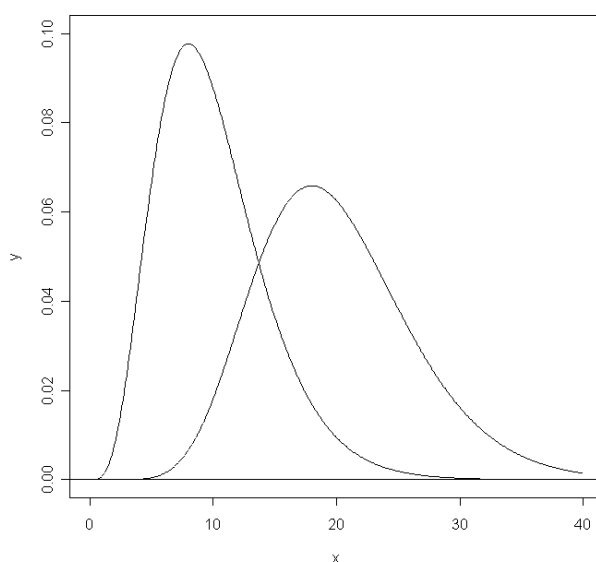


Figure 10.1. Density functions for χ^2 distributions with 10 degrees of freedom (left curve) and 20 degrees of freedom (right curve).

The degree of freedom being equal to $n - 1$ has to do with the theoretical deduction of the distribution. During the calculation it is discovered that the quantity $(n-1)S^2$ in fact consists of only $n - 1$ independent squares and not n as may be thought. These are intricate theoretical aspects which we do not have to worry about. Since we are most interested in application we just accept the fact that the degrees of freedom should be $n - 1$.

If you look up a table or use a statistical program, you will find that for 10 degrees of freedom, 2.5% of the probability mass of the χ^2 distribution lies to the left of 3.25 and 2.5% of the probability mass lies to the right of 20.48. If you look at the above figure you will find this reasonable.

Now suppose you intend to make a series of 11 observations. Then you could use these special points in the χ^2 distribution in order to get $P\left((11-1)\frac{S^2}{\sigma^2} < 3,25\right) = 0.025$, which after rearrangement will be equivalent to $P\left(\sqrt{(11-1)}\frac{S}{\sqrt{3,25}} < \sigma\right) = 0.025$ or $P(1,754 \cdot S < \sigma) = 0.025$. By using the other bound, you get $P\left(\sqrt{(11-1)}\frac{S}{\sqrt{20,48}} > \sigma\right) = 0.025$ or $P(0,699 \cdot S > \sigma) = 0.025$.

This means that the interval $[0.699 \cdot S; 1.754 \cdot S]$ is a random interval, which has 2.5% probability to miss the theoretical standard deviation σ to the left and the same probability to miss it to the right. The hitting probability is 95%. And when we put in the numerical outcomes for S , after having observed the result in the trial, we get a 95% confidence interval for the theoretical standard deviation σ with confidence degree 95%. For instance an outcome of $S = 14.6$ gives the interval $[10,2; 24,9]$, which may also be written as

$$10.2 \leq \sigma \leq 24.9 \quad 95\%.$$

You may observe that this interval is rather long. The upper bound is more than two times the lower one. And it is not caused by us being clumsy, by making incorrect calculations or having used an inferior method. To estimate a theoretical standard deviation well is a demanding task, in the sense that it always requires a lot of observations in order to give a short interval. Unfortunately, this is an intrinsic property of the problem which cannot be overcome in any way.

By tradition the χ^2 distribution is what has been tabulated for use in this problem. That is also what is obtained by means of modern statistical programs. It would have been simpler if there were tables and direct calculations of the multiplicative constants (in the example 0.699 and 1.754) which when multiplied by the empirical standard deviation would give the confidence interval with a suitable confidence degree e.g., 95%. Since such constants are very illustrative when it concerns length of the intervals, I have here made a small table of these constants for some sample sizes and confidence degree 95%.

Number of observations	10	20	30	40	60	80	100
Upper limit	0.688	0.760	0.796	0.819	0.848	0.865	0.878
Lower limit	1.826	1.461	1.344	1.284	1.220	1.184	1.162

Table. Constants to be multiplied by the empirical standard deviation in order to get a 95% confidence interval for the theoretical standard deviation

In the table you can see how the lower boundary approaches 1.00 from below and the upper boundary approaches 1.00 from above. But even for as much as 100 observations there is quite some distance between the constants.

**Technical training on
WHAT you need, WHEN you need it**

At IDC Technologies we can tailor our technical and engineering training workshops to suit your needs. We have extensive experience in training technical and engineering staff and have trained people in organisations such as General Motors, Shell, Siemens, BHP and Honeywell to name a few.

Our onsite training is cost effective, convenient and completely customisable to the technical and engineering areas you want covered. Our workshops are all comprehensive hands-on learning experiences with ample time given to practical sessions and demonstrations. We communicate well to ensure that workshop content and timing match the knowledge, skills, and abilities of the participants.

We run onsite training all year round and hold the workshops on your premises or a venue of your choice for your convenience.

For a no obligation proposal, contact us today at training@idc-online.com or visit our website for more information: www.idc-online.com/onsite/

- OIL & GAS ENGINEERING**
- ELECTRONICS**
- AUTOMATION & PROCESS CONTROL**
- MECHANICAL ENGINEERING**
- INDUSTRIAL DATA COMMS**
- ELECTRICAL POWER**

Phone: +61 8 9321 1702
 Email: training@idc-online.com
 Website: www.idc-online.com



You may also observe that the earlier presented approximate confidence intervals have been of the type with an estimate plus or minus a safety margin. It is not so in these exact intervals even if they hold approximately for very big sample sizes. The upper limit always deviates more from the estimate than the lower limit does.

For a series of normally distributed observations
the χ^2 distribution can be used to get the confidence
intervals for the theoretical standard deviation

Now we shift to the problem of calculating confidence intervals for the expectation parameter μ in a normal distribution. If the theoretical standard deviation σ were known, we could get such a confidence interval by starting the calculation from the normalised random variable

$$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

which is normally distributed with parameters 0 and 1. This is just another way to describe the calculation we have made in an earlier section. The most natural situation, however, is that we do not know the theoretical parameter σ . Then it is natural to study the corresponding expression with the theoretical parameter substituted by the corresponding estimate, that is

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

Theoretical probabilistic calculations can show that the distribution of this random variable depends only on the sample size n . We call its distribution the t distribution with degrees of freedom $r = n - 1$.

Since there is randomness also in the denominator, this distribution is a little broader than the normal distribution with parameters 0 and 1, which is the distribution for the corresponding expression with a fixed value σ in the denominator. For example the point 2.26 in the t distribution with 9 degrees of freedom (corresponds to 10 observations) has the probability mass 97.5% to the left. For the normal distribution with parameters 0 and 1, the corresponding point is 1.96. The following graph shows the density function for the t distribution with 9 degrees of freedom.

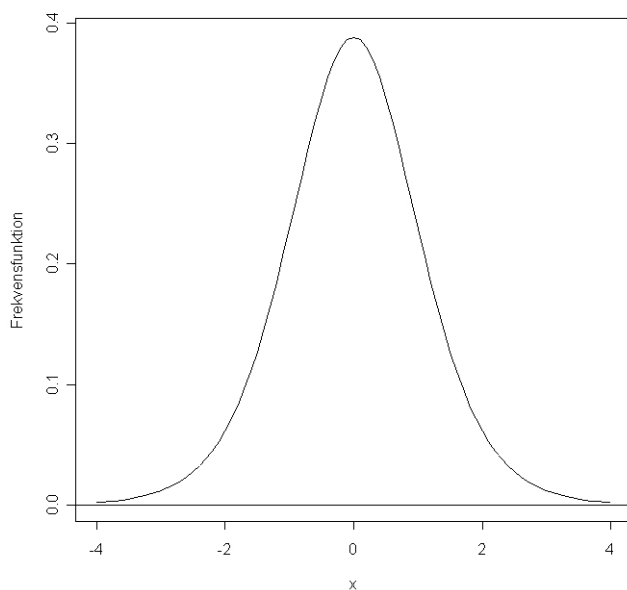


Figure 10.2. Density function for the t distribution with 9 degrees of freedom.

The graph in the figure is very much a reminder of the normal distribution with parameters 0 and 1. The similarities and differences are best shown in a figure with the cumulative distribution functions for both distributions.

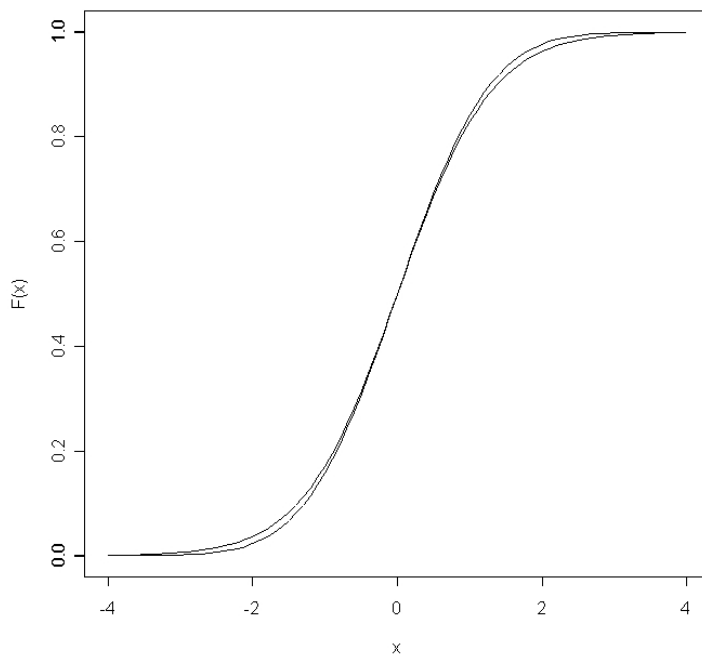


Figure 10.3. Cumulative distribution functions for the t distribution function with 9 degrees of freedom and the normal distribution with parameters 0 and 1. The upper curve on the left side and the lower curve on the right side belong to the t distribution.

We can conclude that taking theoretical results into consideration, we can make confidence intervals for the expectation when the theoretical standard deviation is unknown, in almost the same way as it is done for known theoretical standard deviation. It is just that we replace the unknown parameter σ with the known empirical estimate and replace the normal $(0; 1)$ distribution with the t distribution with $r = n - 1$ degrees of freedom.

And then there will be no approximation any more, but an exact confidence interval. The confidence interval based on the t distribution will always be a little longer than the approximated one based on the normal distribution. But on the other hand thus, the confidence degree will be exact.

The confidence intervals we have discussed here may be used when we have observations from a direct measurement, where the expectation is the interesting parameter. But the method can also be used in a situation when measurements are made before and after some treatment in the same units, in order to estimate a treatment effect parameter. Usually it is a good experimental design practice to make such types of measurements, when we want to study the effect of a treatment. Principally we could have double the number of individuals in the investigation group, measure half of them before treatment and half of them after treatment. But that would be less efficient as you can probably understand intuitively.

I joined MITAS because
I wanted **real responsibility**

The Graduate Programme
for Engineers and Geoscientists
www.discovermitas.com



Month 16
I was a construction supervisor in the North Sea advising and helping foremen solve problems

Real work
International opportunities
Three work placements



MAERSK

Example.

Suppose you have 9 units in an investigation, and that you have the following observations.

Unit	1	2	3	4	5	6	7	8	9
Before	17.7	17.3	15.8	15.4	19.5	22.5	17.9	19.0	18.3
After	21.3	19.6	16.4	18.0	19.9	23.1	18.6	21.2	22.7
Difference	3.6	2.3	0.6	2.6	0.4	0.6	0.7	2.2	4.4

You can see that there are large variations among the observations before treatment as well as after treatment. Values after treatment for some units are smaller than values before treatment for some other units. However the differences between values after treatment and before treatment for the same units, have a more clear structure. There is some variation, but all differences happen to be positive. In the following analysis we now ‘forget how we got the values’ and make an analysis for only the 9 difference values. These 9 values have an empirical mean 1.97 and the empirical standard deviation 1.45. From a table or a statistical program we see that in the t distribution with 8 degrees of freedom we have 97.5% probability to the left of the point 2.31. A 95% confidence interval for the expectation of the increase would be

$$\left[1.97 - 2.31 \cdot \frac{1.45}{\sqrt{9}}; 1.97 + 2.31 \cdot \frac{1.45}{\sqrt{9}} \right] = [1.97 - 1.12; 1.97 + 1.12] = [0.85; 3.09]$$

If we want to talk about error margin, we could say that the estimate 1.97 has an error margin of 1.12. The confidence interval is rather large, but it is fully on the positive side and far away from 0. We can be confident to say that there is a positive effect.

We could also make a formal statistical test. The hypothesis would then be that the effect is zero or negative. With the level of significance 2.5% we could reject the null hypothesis and state that there is a positive effect.

Another possibility is to state that we have estimated the effect to 1.97, and that we have got a p value of $0.0018 = 0.18\%$. This comes out simply from a statistical program if we take one minus the cumulative distribution function in the point $\frac{1.97}{1.45/\sqrt{9}} = 4.076$ for the t distribution with 8 degrees of freedom. It is essentially a matter of choice, what path we will take – a confidence interval, a test with a fixed level of significance or an estimate with a p value. If you have a parameter, which has some physical meaning, the confidence interval gives the most informative report of the result.

11 COMPARING TWO CASES

Suppose you want to compare two cases where you have completely independent observation series. Thus it is not a situation with measurements before and after treatment, of the type we had in the previous section. It may be a comparison between two methods or two groups, where each unit in the trial can be due to only one of the methods or belong to only one of the groups. If for instance, we want to study two different arrangements of mathematics learning, it is clear that each individual can do his or her mathematics studies in only one of the arrangements.

Such comparisons can feature in many applications with different types of measurement variables. There exist special statistical methods for situations where it is reasonable to suppose that the observations have some special type of distribution, for example normal distribution. But there also exist very general techniques, which work well for all types of distributions. We start with a method of the latter type.

Frank Wilcoxon was the name of a chemist who suggested such a general test at the end of 1940. His idea was to skip the very values of the observations and use only the ranks of those in the amalgamation of the two series. The test was called the Wilcoxon two-sample test. His test was modified a little a few years later by Mann and Whitney. It was then called the Mann-Whitney test. This test is equivalent to the Wilcoxon test such that for any chosen level of significance the two tests will reject for exactly the same cases. And if one works with p values, they will be exactly equal too. The basic formulations differ however, and I think that the Mann-Whitney formulation gives a better intuitive understanding of the problem. So I will use that in the following.

Many authors called these rank methods ‘quick and dirty’ and assumed that they would be inefficient in their simplicity. But it was the other way around. In the 1960s there appeared theoretical papers which showed that those methods were not at all inefficient, but instead were surprisingly efficient.

Example

Suppose we have two independent series of observations, one with 8 observations and the other with 9 observations. The outcomes in the first series, which we call x series, were 22.8; 34.4; 30.4; 27.3; 27.9; 17.1; 38.3; 21.5, and the observations in the other series, the y series, were 43.6; 43.1; 39.8; 28.5; 25.9; 41.8; 35.9; 28.9; 34.0.

The calculations needed to get the result of the Mann-Whitney test will be simple if we arrange the observations in increasing order, in both series.

x	17.1	21.5	22.8	27.3	27.9	30.4	34.4	38.3	
y	25.9	28.5	28,9	34.0	35.9	39.8	41.8	43.1	43.6

A y observation, which is greater than an x observation will give a so called inversion. For example the y observation 28.9 will give an inversion with the x observation 27.3. Such an event is an indication, however small, that the y variables have a distribution with a tendency to the right in comparison to the distribution of the x variables. If we consider just the y observation 28.9, it gives rise to 5 inversions in all, with x observations. Now we add the inversions for all the y observations. In our example there are in all $3+5+5+6+6+8+8+8+8=57$ inversions.



www.job.oticon.dk

oticon
PEOPLE FIRST



If the two cases were to have the same continuous distribution, the order between the x and y observations would be completely random. And then each of the 9 y observations would have in mean $8/2 = 4$ inversions with the x observations, so the total number of inversions would be in mean $9 \cdot 4 = 36$ inversions. The random number of inversions could of course deviate from this mean number, but the question is, whether 57 is a big deviance or not. A p value for the test is calculated as the probability of a pure random deviation to give 57 or more inversions.

To calculate this probability is a rather tricky problem. There are tables prepared for such probabilities in the Mann-Whitney test for certain sample numbers. But when the number of observations is large, one can use an approximation with a normal distribution. Observe that this is a normal distribution approximation of the probability that the random number of inversions would be at least equal to some number, and it has nothing to do with the distribution of the observations themselves.

It can be shown mathematically that if the numbers m and n of observations in the two series are big, then the number of inversions is approximately normally distributed. The parameters in the approximate normal distribution are mean $\frac{m \cdot n}{2}$ and the standard deviation

$$\sqrt{\frac{m \cdot n (m + n + 1)}{12}}. \text{ Thus in our example the parameters are mean } 36 \text{ and the standard deviation } \sqrt{\frac{8 \cdot 9 (8 + 9 + 1)}{12}} = 10,4.$$

Since we are talking here of approximating a discrete distribution (with whole numbers as possible outcomes) with a continuous distribution, it may be reasonable to consider the discrete distribution to be the distribution of a rounding of a continuous result. To the probability for an outcome of a certain number in the discrete distribution, we attach the probability in the continuous distribution for an interval with end points as that number minus and plus half a unit. To the outcome 57 in the discrete distribution we then attach the probability of the interval 56.5 to 57.5 in the continuous normal distribution. And the probability of the outcome 57 or more in the discrete distribution is approximated by the probability for the interval from 56.5 to infinity in the continuous one. Expressed in the cumulative distribution function for the normal (0; 1) distribution we thus get $1 - \Omega\left(\frac{56,5 - 36}{10,4}\right) = 0,024$. The p value for a one sided test thus is approximately 2.4%.

In order to understand the idea of the test better, we may introduce the notation

$$u_{ij} = \begin{cases} 1 & \text{for } y_j > x_i \\ 0 & \text{for } y_j \leq x_i \end{cases}.$$

This notation gives a 1 for inversion and 0 for no inversion of y_j with x_i . The sum of all inversions thus is equal to the double sum

$$\sum_{i=1}^m \sum_{j=1}^n u_{ij}.$$

And then the ratio

$$\frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n u_{ij}$$

equals the relative frequency for the event that a y observation is greater than an x observation. This is the very soul of the Mann-Whitney test. The inversion indicators u_{ij} are estimates of the probability $\theta = P(Y_j > X_i)$, and in the test we use their empirical mean. The hypothesis to be tested is $\theta = 0.5$, and when we reject the hypothesis for a big number of inversions, we work in order to reveal if $\theta > 0.5$.

The Mann-Whitney test of the hypothesis that two observation series have the same distributions is simple and can be used for all types of distributions for the observations.

The most relevant alternatives to the hypothesis is the case that y observations have a tendency to give higher values than x observations, in the sense that the cumulative distribution function of the y observations is to the right of the cumulative distribution function of the x observations at all levels. In mathematical statistics this is expressed by saying that y variables are stochastically greater than x variables. This is illustrated by the following figure.

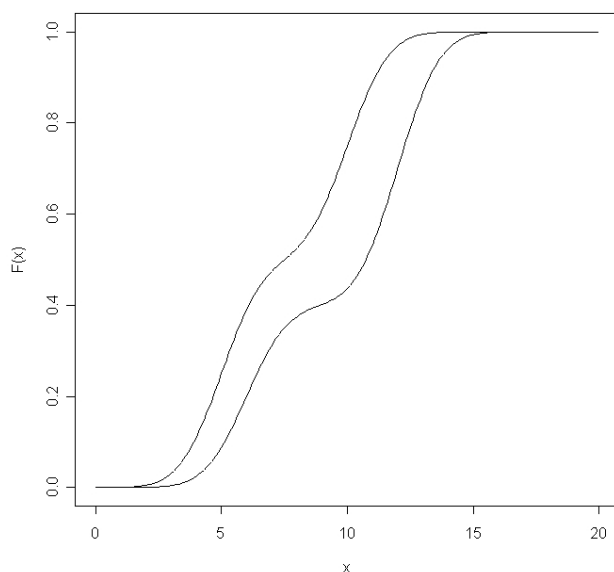


Figure 11.1. Example of cumulative distribution functions for two stochastically ordered random variables.

In the past four years we have drilled

81,000 km

That's more than **twice** around the world.

Who are we?
We are the world's leading oilfield services company. Working globally—often in remote and challenging locations—we invent, design, engineer, manufacture, apply, and maintain technology to help customers find and produce oil and gas safely.

Who are we looking for?
We offer countless opportunities in the following domains:

- **Engineering, Research, and Operations**
- **Geoscience and Petrotechnical**
- **Commercial and Business**

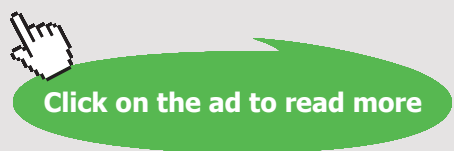
If you are a self-motivated graduate looking for a dynamic career, apply to join our team.

careers.slb.com



What will you be?

Schlumberger



In a figure with cumulative distribution functions you can very easily see if they are stochastically ordered as in the above case. In the following figure we have an example of two cumulative distribution functions for two random variables, which are not stochastically ordered. In fact it is a figure of cumulative distribution functions for two random variables, which differ only by having different variations.

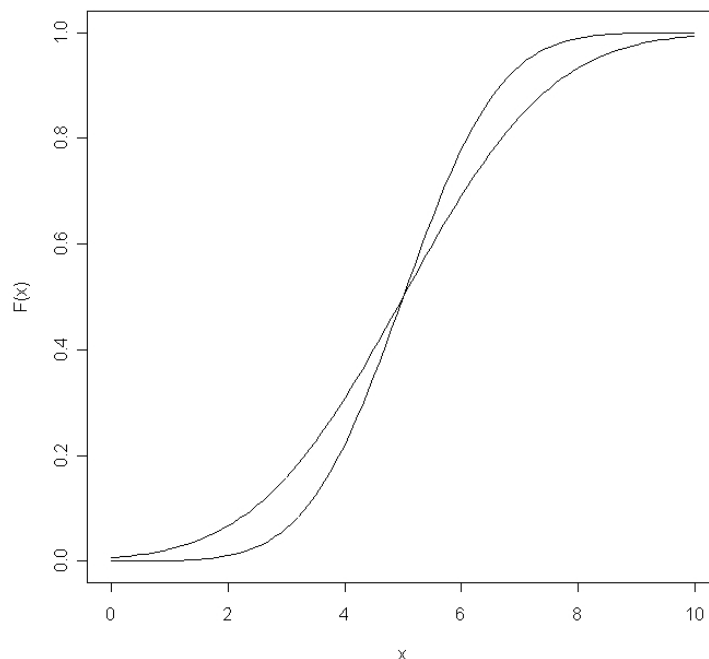


Figure 11.2. Cumulative distribution functions for two cases with same position (symmetry point 5.0) and different variations only.

The Mann-Whitney test has no power against an alternative where the two distributions differ only in deviation. The probability to reject the hypothesis of distributional equality is in fact even smaller than the chosen level of significance for such an alternative. We will not dig further into this, but please remember that the Mann-Whitney test is a test created in order to find positional differences.

A very good property of the Mann-Whitney test (and for that matter also the Wilcoxon test) works well also for observations with discrete distributions. But then we need to have reasonably similar sample sizes in the two cases and use a small correction of the standard deviation parameter in the normal approximation. When we have coinciding x_i and y_j values it is natural to let the inversion indicator u_{ij} equal to 0.5. If we do so, the corrected variance in the approximate normal distribution equals $\frac{mn(m+n+1)}{12}(1-\Delta)$, where the correction term Δ depends on how many coinciding values of different types there are. Suppose there are K points with any coinciding outcomes and let $\tau_k, k=1,2,3,\dots,K$ be the number of outcomes in these points. Then the correction term Δ equals

$$\Delta = \frac{1}{(m+n)(m+n-1)(m+n+1)} \sum_{k=1}^K \tau_k (\tau_k^2 - 1).$$

You can see that points with only one outcome do not contribute to the sum since $(\tau_k^2 - 1)$ equals 0 in such cases.

Example

Let us suppose that in a pedagogical trial, one wants to compare two groups. The participants have to face 7 questions and the number of correct answers is considered to be the result for the participant. Thus for each one, the outcome is one of the numbers 0, 1, 2, 3, 4, 5, 6 or 7. There are 100 participants in each group. The results are tabulated as follows..

Result	0	1	2	3	4	5	6	7
Group 1	1	0	4	14	26	28	19	8
Group 2	4	6	11	17	31	17	10	4
Total	5	6	15	31	57	45	29	12

In the correction factor we have the sum

$$\sum_{k=1}^K \tau_k (\tau_k^2 - 1) = 5 \cdot 24 + 6 \cdot 35 + 15 \cdot 224 + \dots + 12 \cdot 143 = 335742$$

and the whole correction factor equals $\Delta = \frac{335742}{199 \cdot 200 \cdot 201} = 0.042$. We see that there are large numbers in the calculation, but rather small numbers in the final result. The variance we should have if the data were continuous, thus should be decreased by 4.2%. Our estimate of the standard deviation in the normal approximation now equals

$$\sqrt{\frac{m \cdot n(m + m + 1)}{12}} \sqrt{1 - \Delta} = \sqrt{\frac{100 \cdot 100 \cdot 201}{12}} \sqrt{1 - 0,042} = 400,6.$$

It is a rather large standard deviation, but since we have many observations, the inversion sum is also big. In case of a true hypothesis of equal distributions, the inversion sum will be a random variable in the neighbourhood of its mean $100 \cdot 50 = 5000$.

You may have noticed that the values in group 1 seem to be bigger than those in group 2. We may perhaps prefer to let the values in group 1 be used as y values and the values in group 2 be used as x values. But of course the choice of name does not influence the result as long as we know what we do with the plus and minus.



Sweden
Sverige

Linköping University –
innovative, highly ranked,
European

Interested in Engineering and its various branches? Kick-start your career with an English-taught master's degree.

→ Click here!

li.u LINKÖPING
UNIVERSITY

On the calculation of the test statistics we do not need to consider all $100 \cdot 100 = 10000$ individual pairs of observations. It is enough to consider only pairs of outcome points in the two series to see how many cases there are for the different combinations. Observe that we may have both ‘full inversions’ and ‘half inversions’. In our example we get the inversion sum

$$1 \cdot \frac{5}{2} + 0 + 4 \cdot \left(10 + \frac{11}{2}\right) + 14 \cdot \left(21 + \frac{17}{2}\right) + \dots + 8 \cdot \left(96 + \frac{4}{2}\right) = 6551.$$

And then, in order to get the p value in the test we should calculate the normalised variable (difference between the observations and the estimated mean divided by the estimated standard deviation)

$$\frac{6551 - 5000}{400,6} = 3,87.$$

This gives a p value $1 - \Phi(3,87) = 5,4 \cdot 10^{-5}$. The result is ‘extremely significant’. It is ‘clearly proved’ that group 1 has a tendency of higher outcomes than group 2. In order to be formal we could say that the investigation has shown that $P(X > Y) > 0.5$ if we denote outcomes in group 1 by X and outcomes in group 2 by Y .

If we feel confident about the idea that the observations have some special type of distribution, we can make the assumption that the observations have this distribution, and then which requires no assumption on particular distributions. It is evident, that a test which is designed for a particular situation is better than a general test which should be able to adapt to a wider class of situations. The question is, however, how much better the specially designed test is when its distributional assumptions are satisfied; and how it works in comparison to the general test, when these assumptions are not satisfied. We will come back to this question later. But first I will present the common special method designed for the case when the observations are normally distributed.

Often it is assumed that the theoretical standard deviations for the two compared cases are the same, in particular if they are obtained with the same measurement method. In this case our model would be that we have two series of observations X_1, X_2, \dots, X_m and Y_1, Y_2, \dots, Y_m , where all observations are independent and normally distributed, in the x series with expectation μ_1 and standard deviation σ and the y series with expectation μ_2 and standard deviation σ . I have used uppercase letters in the notation for the observation to indicate that they are random. In principle the density functions for the two types of observations may appear as in the following figure.

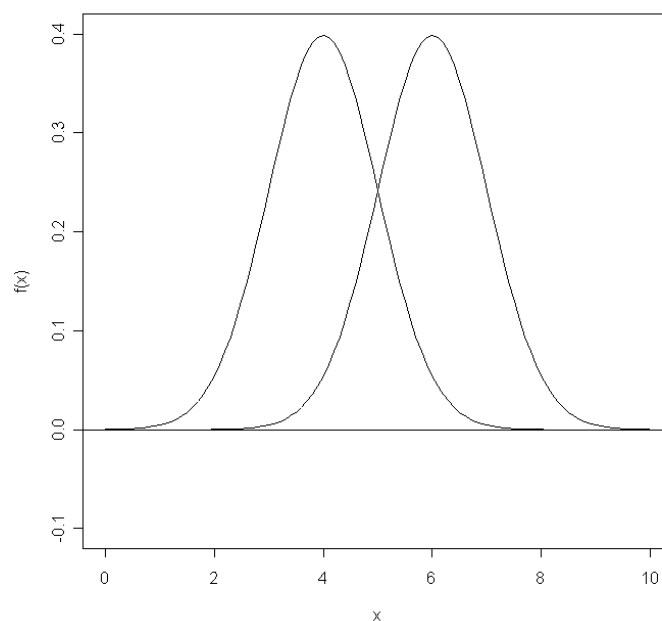


Figure 11.3. An example of density functions for two normal distributions with the same standard deviation but different expectations. Only a translation is the difference between the two cases.

The essential information in the two series of observations are their empirical means \bar{X} and \bar{Y} and their empirical standard deviations S_1 and S_2 . Since the two series are supposed to have the same theoretical standard deviations, there is information on this common parameter in both the empirical standard deviations S_1 and S_2 . In order to get the best information on the common theoretical standard deviation σ , we have to weight together the information from S_1 and S_2 in a suitable way. A larger series of observations, of course, provides better information than a smaller series does. A very good way to put the information together is to work with variances (squares of standard deviations) and to weight the cases with their respective degrees of freedom (sample sizes minus 1). Then the so called weighted empirical variance S^2 equals

$$S^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{(m-1) + (n-1)}.$$

This is a suitable estimate of the common theoretical variance σ^2 .

The theoretical parameter difference $\mu_2 - \mu_1$, which is a measure of the theoretical positional difference between the two cases, is suitably estimated by the empirical mean difference $\bar{Y} - \bar{X}$. Theoretical work has shown that the random variable

$$T = \frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{S \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

has a t distribution with $r = (m - 1) + (n - 1)$ degrees of freedom. With the help of this, we can easily make both confidence intervals for the difference $\mu_2 - \mu_1$ and test the hypothesis $\mu_2 - \mu_1$.

If for example, t denotes the point in the t distribution with the correct number of degrees of freedom, which has the value $0,975 = 97,5\%$ of the cumulative distribution function in the t distribution, then the random interval

$$\left[\bar{X}_2 - \bar{X}_1 - t \cdot S \sqrt{\frac{1}{m} + \frac{1}{n}}; \bar{X}_2 - \bar{X}_1 + t \cdot S \sqrt{\frac{1}{m} + \frac{1}{n}} \right]$$

**STUDY FOR YOUR MASTER'S DEGREE
IN THE CRADLE OF SWEDISH ENGINEERING**

Chalmers University of Technology conducts research and education in engineering and natural sciences, architecture, technology-related mathematical sciences and nautical sciences. Behind all that Chalmers accomplishes, the aim persists for contributing to a sustainable future – both nationally and globally.

Visit us on **Chalmers.se** or **Next Stop Chalmers** on facebook.

CHALMERS
UNIVERSITY OF TECHNOLOGY



has the probability 95% to hit the theoretical parameter difference $\mu_2 - \mu_1$. The outcome of such a random interval then gives a 95% confidence interval. In the same easy way we may handle the test of the hypothesis $\mu_2 - \mu_1$, if we so wish. These are classical methods from the early 20th century, which are among the most used statistical methods ever.

When we calculate the weighted standard deviation, we should not be surprised if the empirical standard deviations in the two series differ quite a bit. It is in the very nature of empirical standard deviations to have big relative variations even for rather big sample sizes. You don't have to worry about this if you use the same measurement method for both cases. If you have other cases, where you may suspect that the theoretical variances differ, you may use some good approximate methods which can also handle different variances. It will take us too deep into the subject to go into these problems here, but it may be good to know that there are possibilities. Looking for Welch-Aspin methods would lead you right up, if you want to study it more.

It is very important to distinguish between the situation we have here, two series with all observations independent, and a situation with repeated measurements under two different situations, on the same unit, which we discussed in an earlier section. If you do not distinguish between these cases, there will be a statistical catastrophe! In one way or the other.

I had promised to come back to the question of efficiency. Theoretical calculations have shown, for instance, that if we have normal observations with the same theoretical variance, then the Mann-Whitney method is just 4% less efficient than the t method, in the sense that we get approximately the same power if a Mann-Whitney test has 4% more observations than a t test. Not much of an efficiency loss! However, for other types of distributions for the observations the Mann-Whitneys test may be much more efficient than the t test, in particular if the observations distribution has slowly declining tails.

12 ONE SIDED OR TWO SIDED

Perhaps you have noted that I have purposely avoided a certain problem. It is the question regarding one-sided or two-sided statistical methods. All confidence intervals we have discussed have been two-sided and symmetric, in the sense that they have the same risk of missing the true parameter on both sides. We have had to make probabilistic risk bounds on both sides. When it comes to tests, all examples have been one-sided in the sense that we have rejected the null hypothesis only at extremely large deviations in one direction. We have calculated the level of significance as the probability of an excessive deviance in one direction only. In a way the ideas are limping here. All confidence intervals are two-sided and all tests one-sided.

There is a certain correspondence between tests and confidence intervals. For instance we can test the hypothesis $\mu_1 - \mu_2$ by rejecting it as soon as the point 0 is not included in the confidence interval for $\mu_2 - \mu_1$. If the confidence interval has a confidence degree 95%, the test will then have the level of significance at 5%. But here we reject the hypothesis for big deviations in both directions. We have got a two-sided test. But the confidence interval in fact gives more precise information than rejection does, in a test. When we get a confidence interval which is fully on the positive side, we can safely say that $\mu_2 - \mu_1 > 0$. And if the confidence interval is fully on the negative side we can safely say that $\mu_2 - \mu_1 < 0$. And all this means that within the same total risk, one minus the level of significance, we can not only reject the hypothesis but also state in which direction we reject it. We have the possibility to make two different statements within the same total risk.

I think that we should treat the testing problem in the same way, if we want to keep it open to either a positive or a negative effect. What we then do is to make a multiple test of two formal hypotheses with a common multiple level of significance, which is defined as the risk to falsely reject any true hypothesis. In practice, this will be the same as making a two-sided test and then completing it with a statement about which side we have rejected it on, if we have. When we make a symmetric two-sided test, we divide the total level of significance into two equal parts. And if we divide the total multiple level of significance into two equal parts for the separate individual one-sided tests, it will, in practice, be the same.

This discussion may seem a little superfluous. But everything works more logically if we consider two separate one-sided tests with a total risk of making a wrong statement in any of these tests. If we think strictly on the logic of the statistical test theory, one should only be allowed to make the (rather uninformative) statement $\mu_1 \neq \mu_2$, if we make a two-sided test and reject the hypothesis $\mu_1 = \mu_2$. And I guess that you certainly want to tell, also, which direction there is deviation in.

You probably have noticed that I used the one-sided level of significance 2.5% in the earlier examples. There were two reasons for this. It is easier to understand the idea of a test if you have only a one-sided alternative. And I wanted the numerical value to coincide after the present discussion, with what I had used, if we wanted the total multiple level of significance at the conventional level 5%.

Finally I want to point out that it is sometimes motivating to make only a one-sided test after all. For example, if we want to discover if there is increased radio activity level in some geographical environment. Then there is reason to reject the hypothesis of non-increased level only if we get a measurement high enough over the standard level. It is probably reasonable to suppose that there exists a basic natural level everywhere, which can be raised somewhere, but not lowered.

MÄLARDALEN UNIVERSITY
SWEDEN

WELCOME TO OUR WORLD OF TEACHING!
INNOVATION, FLAT HIERARCHIES AND OPEN-MINDED PROFESSORS

STUDY IN SWEDEN - CLOSE COLLABORATION WITH FUTURE EMPLOYERS
MÄLARDALEN UNIVERSITY COLLABORATES WITH MANY EMPLOYERS SUCH AS ABB, VOLVO AND ERICSSON

TAKE THE RIGHT TRACK
GIVE YOUR CAREER A HEADSTART AT MÄLARDALEN UNIVERSITY
www.mdh.se

DEBAJYOTI NAG
SWEDEN, AND PARTICULARLY MDH, HAS A VERY IMPRESSIVE REPUTATION IN THE FIELD OF EMBEDDED SYSTEMS RESEARCH, AND THE COURSE DESIGN IS VERY CLOSE TO THE INDUSTRY REQUIREMENTS.
HE'LL TELL YOU ALL ABOUT IT AND ANSWER YOUR QUESTIONS AT MDUSTUDENT.COM

It is standard practice to make two-sided tests in applications. In many cases it is even argued that it is incorrect to make one-sided tests with the level of significance equal to 5%. That means then that one should always allow for deviation in both directions. If one first looks at the direction in which the results are going, and then makes a one-sided test on 5% level, the level in the test is not 5% any more. A 'post hoc' constructed test does not have the technical level attached afterwards. But as I said earlier, there exist some situations where one needs to take only one direction of deviation into consideration. And there exist situations where we need to consider the possibility of deviations in both directions.

Some advice for when you read something, which includes tests. Do always check whether the author uses one-sided or two-sided tests. And if you write something yourself with tests in it, be careful to indicate the same clearly. And of course talk about the level of significance you use. Ideally of course, use two multiple one-sided tests with declared multiple levels of significance.

When it comes to confidence intervals, I think that in almost all cases the use of two-sided confidence intervals is suggested, which give bounds in both directions. An exception for instance, may be if you want to show that a standard deviation of some kind has an upper bound. Since a standard deviation can never be negative, the confidence interval then has the type $[0; z]$, for some suitable bound z determined by the observations and a given low risk of missing the true parameter to the left.

Always check if presented p values are one-sided
or two-sided. Do always two, simultaneous one-sided tests
with the individual level equal to half the presented
multiple level, if both positive and negative
alternatives are reasonable.

13 IT DEPENDS

We have spoken a great deal about random results of different types. In some problems we had two different series of observations which came from two different cases of some kind. Parameters for the random observations could be different for the two cases, and we were interested in the possible differences. Now we will study statistical methods for situations where the distributions of the random observations depend on a continuous ‘background’ variable. If we denote a random observations with Y and denote the background variable with x , then in statistics we say that Y has regression on x , meaning that the distribution of Y is determined by x and perhaps some parameters. The original meaning of the word regression is ‘return’ or ‘go back’, which fits well here too. The random observation Y goes back to x , since its distribution is determined by x and some theoretical parameters.

Often it is meant to suppose that the observations are normally distributed. In the simplest regression model, the observation Y has a normal distribution with a constant standard deviation and an expectation μ , which is a linear function of x . In school mathematics, the form of this line could be $\mu = kx + l$. For several reasons it is here more convenient to let the variable x have a starting point in the midpoint of the observational material. Thus, with other, but equivalent, parameters we may write the equation of the (unknown) line as

$$\mu = \alpha + \beta(x - \bar{x}),$$

where \bar{x} is the mean of the x coordinates for all observation points. The parameter α can be interpreted as the height position of the line in the mid point \bar{x} in the material and β is the elevation parameter (gradient of the line).

This simple linear regression problem has three parameters σ , α and β . The following figure illustrates an example, where I have generated artificial normally distributed observations for $x = 1, 2, 3, \dots, 10$, with constant standard deviation $\sigma = 0,5$ and the true regression line $\mu = 3,65 + 0,3(x - 5,5)$. You can see how the generated observation points are spread randomly around the true regression line.

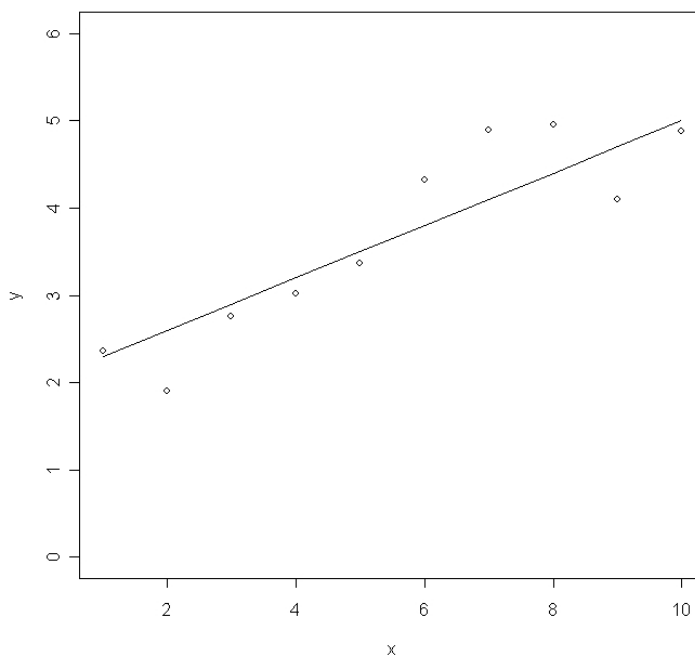



Figure 13.1. A theoretical regression line and 10 artificially generated observations.

LIFE SCIENCE IN UMEÅ, SWEDEN - YOUR CHOICE!

- 32 000 students • world class research • top class teachers
- modern campus • ranked nr 1 in Sweden by international students
- study in English

- Bachelor's programme in Life Science
- Master's programme in Chemistry
- Master's programme in Molecular Biology

Download brochure here!



UMEÅ UNIVERSITY
FACULTY OF SCIENCE & TECHNOLOGY



In a real situation, you have a cloud of observation points of corresponding x and Y values, and the statistical problem then is to use this data in order to estimate the three parameters σ , α and β , and then to also create confidence intervals or make tests of different hypotheses concerning these parameters.

The estimation of the line parameters α and β can be made by use of the so called least squares method. That method means, to find the straight line which fits the observation points best by minimising the sum of squares of all deviations from observations to the line in the y direction. In practice, you do not need to sit and minimise. Formulas have been developed which give the solution directly. We have the line equation in the form $\mu = \alpha + \beta(x - \bar{x})$.

Then the estimates of α and β , denoted $\hat{\alpha}$ and $\hat{\beta}$, are equal to

$$\hat{\alpha} = \bar{y}$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^n (x - \bar{x})(y - \bar{y})}{\sum_{i=1}^n (x - \bar{x})^2}.$$

With mathematical methods it can be shown that these two parameter estimates are independent, which makes life easier when one wants to make statistical calculations in this model. The suitable estimate of the standard deviation σ is based on the distance between observation points and the estimated line in y direction. It can be written as

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - \left(\hat{\alpha} + \hat{\beta}(x_i - \bar{x}) \right) \right)^2.$$

You need not worry about the formulas, since statistical programs calculate this variance estimate directly, if you just type the data into the computer. But the character of the formula also gives a feeling of how variance estimates work. Hence I included the formula here.

Perhaps now you begin to understand the funny numerators in the earlier variance estimates too. In a simple measurement series with n observations and only one position parameter, the expectation, we divided by $n-1$, and we got a χ^2 distribution with $n-1$ degrees of freedom. In a comparison of two cases with $m+n$ observations and two expectation parameters, we divided by $m+n-2$, and we got $m+n-2$ degrees of freedom. Now we have n observations and two expectation parameters, and we divide by $n-2$. A deeper mathematical analysis reveals that $(n-2)\frac{\hat{\sigma}^2}{\sigma^2}$ has a χ^2 distribution with $n-2$ degrees of freedom.

All this is based on the same principle, which can most easily be described in the following way. Among all dimensions corresponding to observations, a certain number of dimensions must be used for expectation parameter estimates. The rest of the dimensions can be used for estimating the variations, that is, the variance parameter. The corresponding degrees of freedom are the number of dimensions used for the variance estimate. The same number appears in the numerator when we calculate a mean of the variation. Now let us consider an example.

Example

In a trial, one wanted to see how the temperature influenced a chemical process. A reaction experiment with regulated temperatures was performed, and the obtained mass of a certain substance was measured for temperatures 20, 25, 30, 35, 40, 45, 50 degrees Celsius. The results were as follows.

Temperature	20	25	30	35	40	45	50
Obtained mass	78.8	80.6	85.1	84.4	86.3	84.9	85.3

The mean of the x values (temperature) is 35. A statistical program will deliver the estimates $\hat{\alpha} = 83.6$, $\hat{\beta} = 0.209$ and $\hat{\sigma} = 1.80$. First we will see what it looks like when we put both observations and estimated regression line $y = 83,6 + 0,209(x - 35)$ in a figure.

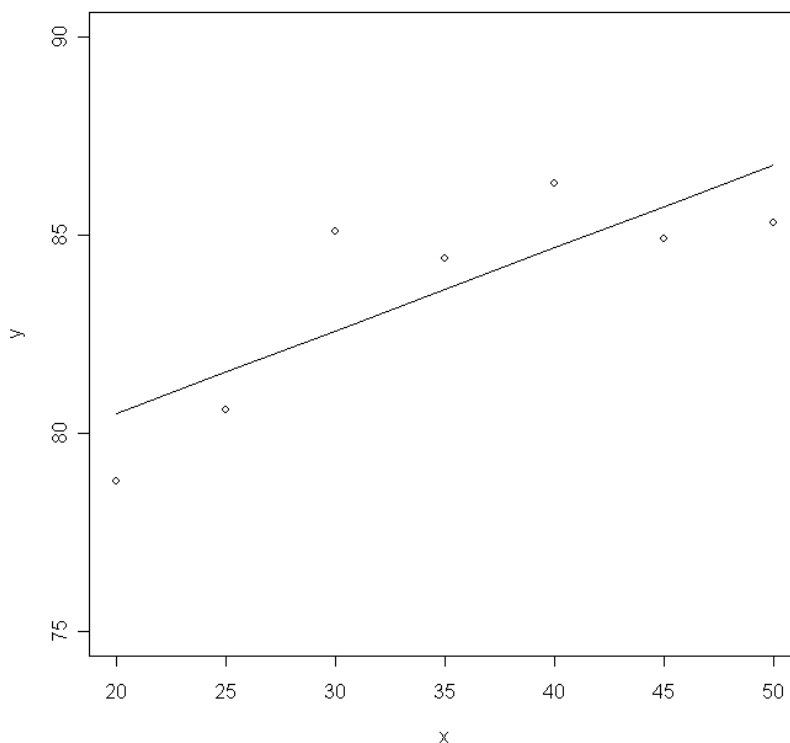


Figure 13.2. Obtained mass in a chemical reaction. Observations and estimated regression line.



Scholarships



Open your mind to new opportunities

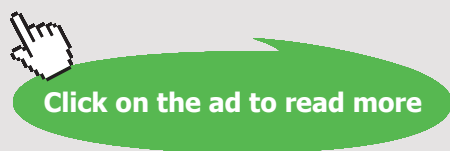
With 31,000 students, Linnaeus University is one of the larger universities in Sweden. We are a modern university, known for our strong international profile. Every year more than 1,600 international students from all over the world choose to enjoy the friendly atmosphere and active student life at Linnaeus University. Welcome to join us!

Linnaeus University
Sweden

Bachelor programmes in
Business & Economics | Computer Science/IT | Design | Mathematics

Master programmes in
Business & Economics | Behavioural Sciences | Computer Science/IT | Cultural Studies & Social Sciences | Design | Mathematics | Natural Sciences | Technology & Engineering

Summer Academy courses



Can we safely state that there is a temperature dependence in the studied temperature interval? A multiple test consisting of a one-sided test of the hypothesis $\beta = 0$ against the alternative $\beta > 0$ and a one-sided test of the hypothesis $\beta = 0$ against the alternative $\beta < 0$ can give the answer to that question. Most statistics program give the value of a test statistics based on the estimates $\hat{\beta}$ and $\hat{\sigma}$ in form of a t statistic

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

which is based on the estimates $\hat{\beta}$ and $\hat{\sigma}$. Under the hypotheses this test data has a t distribution with $n-2$ degrees of freedom. Here it is thus 5 degrees of freedom. Often, the statistical programs also give p values. In our material, the outcome of T is 3.07 and the p value is $0.0139 = 1.39\%$. The experiment showed that the obtained mass depends significantly on the temperature. The figure also clearly indicates that there ought to be some dependence. Since T is positive, it is in the test of the hypothesis $\beta = 0$ against the $\beta > 0$ where we can expect a rejection and make the statement $\beta > 0$.

When you fit a straight line to observation material of a number of pairs of x and y values, we talk of a simple linear regression. Such a statistical model can be used when we have a good motivation for the regressions function to be linear. If we have a small interval of observed x values, we can often use the simple linear regression as an approximation, as also if we do not have a particular motivation for linearity.

In simple linear regression we fit
an ordinary straight line to the data material.
Using statistical tests we can also
find whether a dependence can be safely stated.

The term regression function has a much wider meaning than just a straight line. It may for instance be a linear function of several background variables. Or it may be a polynomial of some degree, for instance a second degree polynomial in one variable. In that case the regressions function is determined by three parameters, a constant term and coefficients for polynomial terms of first and second degree. A question which could be posed in that case is, whether there is a significant curvature of the function, i.e., if the second order term is needed in the function. In our example of the chemical reaction we might have posed that question. In the next section we will come back to similar questions.

Then there is parameter linear regression, which means that the parameters are involved in the regressions functions as multiplicative coefficients. For example a second order polynomial regression is a parameter linear problem. The regression function is non-linear in the x values, but it is a linear function in the parameters, which are coefficients of x polynomials of the order zero, one and two.

It may of course also be natural in some situations to use genuinely non-linear regression functions. An example is an exponential regression function of the type $\mu = \alpha \cdot e^{\beta x}$ which is genuinely non-linear in parameters.

A parameter linear regression problem is not so difficult to handle, since the estimates are obtained directly by formulas. Genuinely non-linear regressions functions need suitable numerical methods to find the estimates, and also from a theoretical point of view they are much more complicated. We will not discuss that in any detail here.

There also exists so called generalised linear regression, where one has observations with distributions other than the normal one, and in a suitable way the parameters in the distribution are transformed to a linear form of background variables via a so-called link function. We will come back to generalised linear regression in a later section.

14 ANALYSIS OF VARIANCE

Everyone who has read scientific papers or investigations with statistical content has probably stumbled upon the concept of analysis of variance sometime. Thus it is extremely relevant to give the technique of analysis of variance in this book. In an earlier section I have discussed the problem of comparison of two cases. In a sense, this is the simplest form of analysis of variance. What we usually call simple analysis of variance is a method to compare a number of similar cases. And two is, of course, a special case of several. But generally, in simple analysis of variance we compare more than two cases. The application surroundings can vary. It may be comparison of a number of communities, a number of professions, a number of medical treatments and so on.

**YOUR WORK AT TOMTOM WILL
BE TOUCHED BY MILLIONS.
AROUND THE WORLD. EVERYDAY.**

Join us now on www.TomTom.jobs

follow us on **LinkedIn**



#ACHIEVEMORE

TomTom 

Model

In the simplest model we have for instance n cases with m observations for each case. All observations are supposed to be independent and normally distributed with the same (unknown) standard deviation σ . The different cases may have different expectations, $\mu_i, i = 1, 2, \dots, n$. Yes, of course they may also occasionally be equal, which is in fact the hypothesis we are going to test in our analysis of variance. If we make the test on a small level of significance and reject that hypothesis, we may safely state that there are (some) differences between the expectations.

Now denote the observations by $X_{ij}, i = 1, 2, \dots, n, j = 1, 2, \dots, m$. The observations for the different cases could be characterised by their means $\bar{X}_i = \frac{1}{m} \sum_{j=1}^m X_{ij}$ and their variances $S_i^2 = \frac{1}{m-1} \sum_{j=1}^m (X_{ij} - \bar{X}_i)^2$, for $i = 1, 2, 3, \dots, n$. All information on the common variance parameter σ^2 is collected in the weighted variance $S^2 = \frac{1}{n} \sum_{i=1}^n S_i^2$. Here it can now be mathematically deduced that the distribution of the random variable $W = n(m-1) \frac{S^2}{\sigma^2}$ is a χ^2 distribution with $n(m-1)$ degrees of freedom.

In analysis of variance, one estimates
the common variance for the different cases
by putting the different pieces of information
together in the weighted variance.

Now when we are trying to get a measure of how much the means differ, we can suitably calculate a variance expression for the means, that is

$$V = \frac{1}{n-1} \sum_{i=1}^n (\bar{X}_i - \bar{X}_{TOT})^2,$$

where \bar{X}_{TOT} is the total mean over all cases. This can be seen as a calculation of an empirical variance, where the means act as observations. Since the variance for an individual mean is $\frac{\sigma^2}{m}$, the variable V would estimate this variance if there were no true differences between the expectations. If the outcome of the variable is divided by the corresponding theoretical variance, we get $\frac{(n-1)V}{\sigma^2/m} = \frac{m(n-1)V}{\sigma^2}$, which would have a χ^2 distribution with $n-1$ degrees of freedom.

On the other hand, if there are some differences between the expectations, there is a tendency for V to have increased values. This can be used in the construction of the analysis of variance test statistics, which is the ratio of the two random variables $\frac{mV}{\sigma^2}$ and $\frac{S^2}{\sigma^2}$. Both include the unknown variance χ^2 , but that is not the case for their ratio

$$\frac{\frac{mV}{\sigma^2}}{\frac{S^2}{\sigma^2}} = \frac{mV}{S^2}.$$

If there are no differences between the expectations, this ratio would have an outcome in the vicinity of 1. In this case it will also have an old classical distribution, which is called the F distribution with $n - 1$ degrees of freedom in the numerator and $n(m - 1)$ degrees of freedom in the denominator. If the hypothesis of equal expectations is not true, there is a tendency of increasing value in this ratio. Thus it may favourably be used as test statistics to test the hypothesis that all the expectations are equal.

The idea of analysis of variance
is to compare variations
between series and
variations within series.

This F distribution began to be studied already in the early days of statistical theory, almost a hundred years ago. With much effort, tables were made for certain percentage points in the distribution for different numbers of degree of freedom. Nowadays the statistical programs on computers can quickly deliver both test bounds and obtained p values. Let us look at an example of an F test in a simple analysis of variance.

Example

In a trial four groups of patients are compared with respect to pH value in the urine. There are three individuals in each group, and the measurement values are the following

Group	1			2			3			4		
Value	5.2	6.0	6.1	7.1	6.6	7.3	4.8	5.5	5.8	5.9	6.4	6.6

The following figure depicts the data. Both the table values and the figure indicate a difference between the groups. Is there a significant difference at all? Can group 2 perhaps be shown to have a greater expectation than the others?

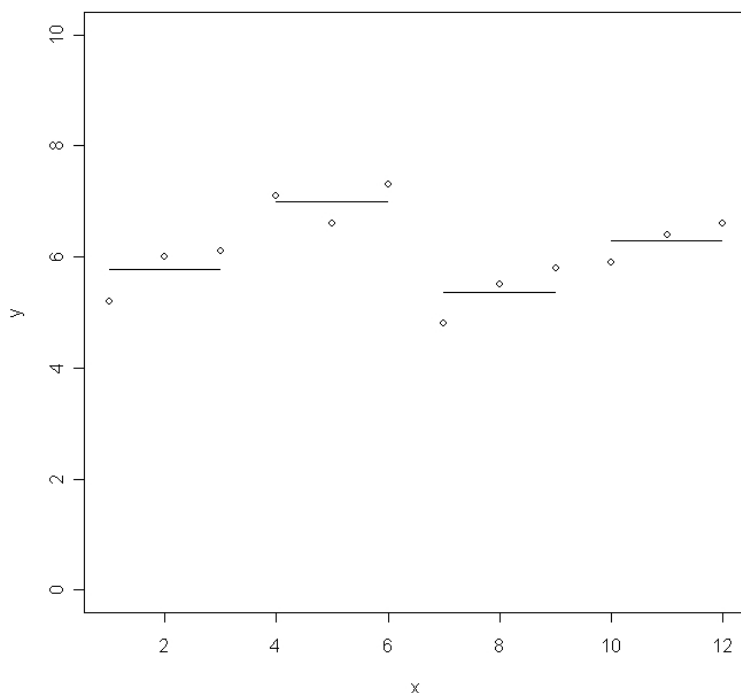


Figure 14.1. An example of analysis of variance. The observations are marked as y values. In each group the mean is illustrated by a horizontal line.

.....Alcatel-Lucent 

www.alcatel-lucent.com/careers

What if you could build your future and create the future?

One generation's transformation is the next's status quo. In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".



Our weighted variance S^2 has to be composed of the four individual variances within the four groups. Its value becomes 0.192. For testing, we need the variance expression V calculated using the four means 5.77; 7.00; 5.37; 6.30. The value is 0.497. Putting in the correct factors including m and n gives the outcome $3\frac{V}{S^2} = 7.77$. If this had been thirty years ago, we would have had to consult a table of the F distribution with 3 degrees of freedom in the numerator and 8 degrees of freedom in the denominator. The closest table value would have been the 99 percentile, which is 7.59% and the 99.9 percentile, which is 9.60. We would have been able to reject the hypothesis of equal expectations on the 1% level but not on the 0.1% level. But now in the age of computers, we can also get the exact one-sided p value from a statistics program. This value appears to be 0, 0093 = 0, 93%.

Note that it is correct to use a one-sided p value here. The only possible deviation from the F distribution of the null hypothesis, is an increase. Sometimes an extremely small value may occur, but that can only happen due to pure chance. No alternative whatsoever can give a tendency of a smaller value compared to the null hypothesis.

If we make a test, and reject the null hypothesis of equal expectations, we can safely state: 'There are **some** differences between the expectations'. That statement is within the risk of the chosen level of significance. To safely give an explanation for what the differences consists of, is a much more complicated problem, which requires more theory of multiple statistical tests.

This basic sketch of analysis of variance included only the very simplest type of analysis of variance. There are several more structured problems, each of them requiring its own type of analysis of variance. There is no chance that we could cover much more in this small book. But I will give one more scenario, which, in all its simplicity, gives some more insight into the basic thoughts of analysis of variance.

Let us go back to the regression problems in section 13, and in particular the example with the measurement of mass obtained in a chemical reaction. The figure with data and the estimated straight line in it perhaps gave the impression that the straight line model was not enough for the experiment. Do we need a model with a second order polynomial there, in order to involve curvature of the regression function?

Now we ought to think of the problem in steps. First, we have a constant term. It is estimated by the mean of all observations, which is in fact the least sum of squares estimate of that constant. After this, there still remains some deviation from the model.

Next we add a linear term and make an estimate of the model with both constant term and linear term, i.e., a simple linear regression model. This is a better fit than the first attempt, but deviances still remain. There is however, always some decrease in the sum of squares. If that decrease is very small, we would not need the linear term in the model, that is, we would not be able to establish dependence.

After this we add a second order term (quadratic term), which means bending of the regression function. Fitting a model with all three types of terms decreases the sum of the squares of deviations even more. If this decrease is large, we certainly need the second order term in the model.

In the figure below, you can see the observations together with the fitted second order polynomial. I think that you would agree that there is now a good fit. For examining the fit we need some calculations, which are based on the numbers in the following table including the sums of squares in the different steps.

	Remaining sum of square	Decrease	Dimension	Decrease divided by dimension
Constant	46.79	-	1	-
First order term	16.13	30.66	1	30.66
Second order term	4.47	11.66	1	11.66
Pure error	0	4.47	4	1.12

If you like geometry and abstract thinking, the following discussion may give you more insight. Our 7 observations correspond to 7 dimensions. When we fit a constant to the observations, it requires one dimension, and after that fit only 6 free dimensions remain. The remaining sum of squares is determined in these 6 dimensions. When we also fit a linear term, we use a further dimension, and the remaining sum of squares is determined in the remaining 5 dimensions. Finally when we also fit a quadratic term, we use a third 'estimation dimension' and there remain only 4 free dimensions. In this space we can make the estimate of the variance.

But now we come to the very idea of the analysis of variance. If there are no systematic effects, but just pure random variations, then the outcomes in the last (rightmost) column estimate the same theoretical variance. And in the very last row of that column is a pure error estimate, which the others can be compared to. This means of course, that we have reason to suppose that the model need not have more terms (like third degree terms etc.). The tests for different degree terms will be F tests with the degrees of freedom determined by the dimension numbers.



Nido

Luxurious accommodation

Central zone 1 & 2 locations

Meet hundreds of international students

BOOK NOW and get a £100 voucher from voucherexpress

Nido Student Living - London

Visit www.NidoStudentLiving.com/Bookboon for more info.

+44 (0)20 3102 1060

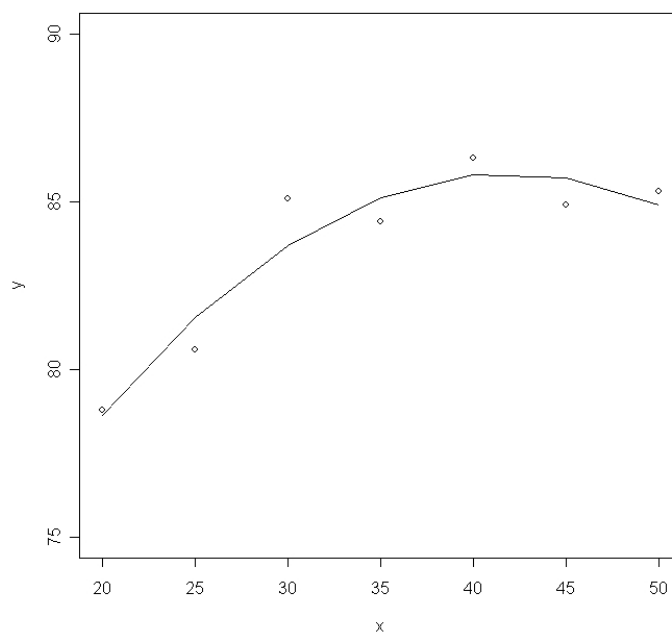


Figure 14.2. Observations in a regressions problem and a fitted second order polynomial.

In order to test if a second order term is needed in the model, we should use the test statistics $\frac{11.66}{1.12} = 10.41$. If the second order term were not to be manifested in the theoretical model, we would get a random variable here, which would have an F distribution with 1 degree of freedom in the denominator and 4 degrees of freedom in the numerator. A statistical program gives us the p value $0.032 = 3.2\%$. If we use a conventional level of significance 5%, we could say that the second order term is significant, that is, we have a curvature in the regressions function.

Another question is of course, if a second order polynomial is a suitable model for the dependence. But in a rather small investigation interval (for the x variable, in this case temperature), it may always be a reasonable model to use a second order polynomial as an approximation if we think that there may be curvature.

The column that I have labelled 'Decrease' in the table above, is often called 'Sum of squares' or shortened to 'S S' in statistical software programs. In the same way, the last column is often denoted as 'Mean square' and shortened to 'M S'.

In more advanced analysis of variance,
one often compares decreases in sum of squares
with an error variance obtained from a later sum of squares.

15 AND ANALYSIS OF VARIANCES

Many situations involve random variations on several levels, in the sense that some sources of variation may influence several observations simultaneously, while some other sources of variation may influence individual observations only. For instance, a patient who is involved in an investigation where we study a medical measure, often has a variation of that measure, which is characteristic for him as a patient. And to this variation is added a more spontaneous extra variation from day to day. Some patients may have a high personal level and a random variation around this, while others may have a low personal level and a random variation around this. In a model we can think of these variations as two components of variation. And in the analysis we must take both types of variation into consideration.

It is not easy to measure the two types of variations in such a model. When analysing it we must clearly work both with variation within patients and variation between patients. Variations within patients, measured as variations between time points for the same individual, can give estimates of the internal variation component. In order to estimate the other type of variation, we need to use the variation between means for patients having the same background characteristic. The structure will best be understood with a simple example.


Example

Suppose the variation between individuals in a group can be described by considering random individual characteristics which have a fixed unknown expectation and an unknown variance denoted by τ^2 , and added to this individual characteristic is an extra independent internal variation within the individual with a variance denoted by σ^2 .


Suppose further that we make m observations for each one of n individuals. Denote the empirical mean and variance in terms of individual number i by \bar{X}_i and S_i^2 . Then the variance within individuals is estimated without systematic errors by the weighted variance $S^2 = \frac{1}{n} \sum_{i=1}^n S_i^2$. If we consider the model we have used, we see that the theoretical variance for individual means is equal to $\tau^2 + \frac{\sigma^2}{m}$. This is explained by the fact that the individuals' characteristic value with variance τ^2 is included in all observations, and this characteristic value shows a variance of the mean of the extra random variation from observation to observation, which is $\frac{\sigma^2}{m}$.

The formal empirical variance $T^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{X}_i - \bar{X}_{TOT})^2$ of the empirical means for individuals gives an estimate of the theoretical quantity $\tau^2 + \frac{\sigma^2}{m}$. If m is very large, this is of course approximately equal to τ^2 but in general we need to make a correction for $\frac{\sigma^2}{m}$. Thus we can generally estimate τ^2 by the difference $T^2 - \frac{S^2}{m}$.

A concern here is that, in particular if τ^2 is small, occasionally the estimate may be negative. And since the theoretical value of the variance is non-negative, this would estimate an impossible value. This problem is however easily solved by substituting the estimated value by 0. And after all this is an improvement, since it just forces the estimate a bit towards the true value, which is non-negative.

SIMPLY CLEVER


WE WILL TURN YOUR CV INTO AN OPPORTUNITY OF A LIFETIME



Do you like cars? Would you like to be a part of a successful brand? As a constructor at ŠKODA AUTO you will put great things in motion. Things that will ease everyday lives of people all around Send us your CV. We will give it an entirely new new dimension.

Send us your CV on
www.employerforlife.com

Whether you makes statistical investigations yourself or just read reports from others, it is important to understand the different types of variations in data and to distinguish between variations between individuals and variations within individuals. This difference may be used in the planning of experiments. Since variations within individuals are often smaller than variations between individuals, and it will be helpful if an experiment can be designed to make comparisons between cases within individuals. Examples of this can be found in efficiency measurements as differences in individuals between after treatment and before treatment. Many times, however, this is not possible. If the aim is to compare two types of treatments, different individuals must be targeted for the two treatments, since each individual can only be given one of the (curative) treatments. But for continuous medical treatments one can make a plan where individuals shift between treatments. If you are interested in this matter, you should read about cross-over trials. Unfortunately there is not space enough in this small book to discuss these interesting and important things in greater detail.

Random variations in data often has some sources which influence several observations. Then these observations will be dependent and the analysis of data must take this structure into consideration.

It is wrong to make the analysis as if all observations were independent.

16 NOW AND THEN OR HERE AND THERE

In a television program on crimes, a listener asked an expert in criminology about his opinion on an increased number of murders from one year to the next. The expert had probably studied some statistics, because he had a good understanding of the quality of statistical data and he commented: ‘When it comes to small data sets, one must understand that big relative variations can appear just randomly. It may be just natural random variation, when there are no murders one year and perhaps four murders the next year.’

I must say that I was quite happy to find such an intuitive insight by a person in a public broadcast medium. So often I have come across the opposite, for example long and heated discussions on some small numbers with no real content. Now we will discuss the type of data where events of some type occur randomly in time or in space. And in that connection we will also consider the case when there are few events occurring in the studied time interval or in the studied space.

In order to make a simulation on the computer, of such a ‘now-and-then-phenomena’, we could for instance divide the year in 100 equal parts and in each of them make a random generation with a probability of 0.02 for occurrence of the event. This would give in the mean ‘2 events per year’ in an ‘almost continuous manner’, which reminds one of the truly continuous manner in real life. I ran 1000 years and I got the following empirical distribution.

Result	0	1	2	3	4	5	6
Proportion	0.134	0.278	0.280	0.169	0.092	0.037	0.006

Table of a simulation result.

Here is a graphical picture of the simulation result.

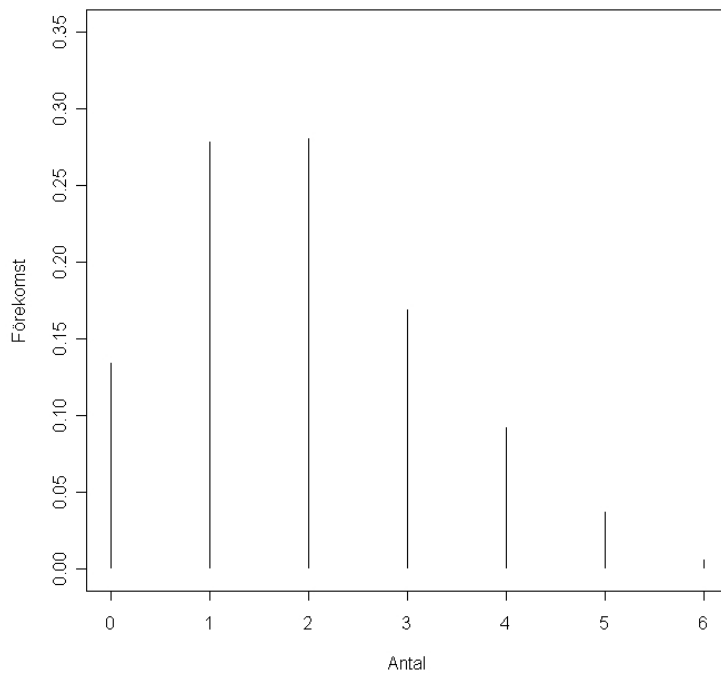
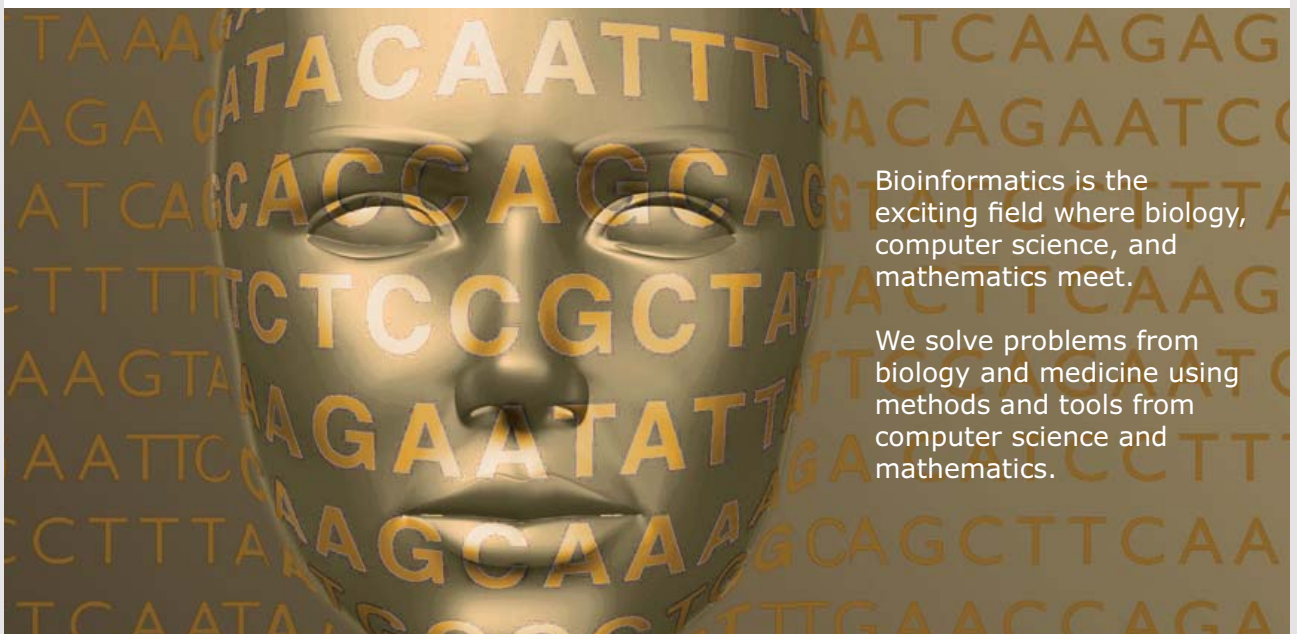


Figure 16.2. Result of a realistic simulation of yearly numbers with theoretical mean 2.



UPPSALA
UNIVERSITET

Develop the tools we need for Life Science Masters Degree in Bioinformatics



Bioinformatics is the exciting field where biology, computer science, and mathematics meet.

We solve problems from biology and medicine using methods and tools from computer science and mathematics.

Read more about this and our other international masters degree programmes at www.uu.se/master



The trick to come close to reality was to cut time into small intervals with a small probability of occurrences in each of them. In theoretical work in mathematical statistics, one goes ‘all the way to infinitely small intervals’ in order to get a general model. Then the result for the theoretical limit for the probability distribution is obtained. In our example, with a mean of 2 events per time unit, the limiting result is the one given in the following table.

Result	0	1	2	3	4	5	6
Probability	0.135	0.271	0.271	0.180	0.090	0.036	0.012

Table of theoretical distribution of the number of events per year.

Our small calculations show very clearly the statement given in the criminology television program that there are large relative variations in data consisting of small natural numbers. And it is good advice to not trust this data too much. One ought to be skeptical of making too much of such data. If you cannot explain the big relative variations to others yourself, you can always advise them to read this little elementary book, which they get without any cost.

Of course, data with large relative variations can sometimes be used for responsible conclusions, if there are many such data points available. For instance there may be a good estimate for a time period consisting of a number of shorter intervals even if there are large relative variations in the shorter intervals.

So far we have talked mostly about the relative variations in random data which are small natural numbers. Such a large relative variation is a characteristic of data of the type ‘now and then’ or ‘here and there’. But there are also other characteristics of such data.

‘One accident rarely comes alone’ is a saying I guess you have heard. Is it true that accidents have a tendency to accompany each other? Many people think so. On the other hand I guess you have heard people say ‘Such accidents occur in the mean every second year; it is now 20 months since the last one, so now it will happen within the next few months’. What do you think? Are adverse events occurring together or are they spread evenly?

I have to say, there is some truth in both and there is some falsehood in both statements. Let me try to explain. Accidents can certainly be assumed to occur randomly. There is hardly any ‘memory process’ built in; the potential accident places does not have any information on earlier events, if we disregard ice, snow, traffic flow and similar things, which disrupt the common randomness. But such times with increased risk can be thought of as a random process with higher accident intensity. But internally in time intervals I must say that I strongly believe in pure randomness.

What does it look like then? In order to illustrate this, I have made a little simulation. For 1000 small time intervals I have used the computer to generate ‘an accident’ with a probability of 0.02 for each interval. Then accidents happened to appear in the following time intervals:

12, 15, 28, 57, 76, 80, 109, 290, 315, 334, 456, 476, 481, 551, 569, 586, 616, 648, 682, 735, 788, 813, 872, 875, 905, 940, 971

In four cases there happened to be ‘one accident does not come alone’. There are less than 5 time units between accidents in the intervals 12–15, 76–80, 476–481 and 872–875. The theoretical mean distance should be 50 time units. In three cases, 682–735, 735–788 and 813–872 there happened to be approximately 50 time units in between. In one case interval 109–290, there is a distance of 181 time units between accidents! As you see, our simulated example gave a mixture of small, medium and long distances without any particular pattern. It is just random! Perhaps a figure depicting the time points obtained, can contribute further to your understanding.

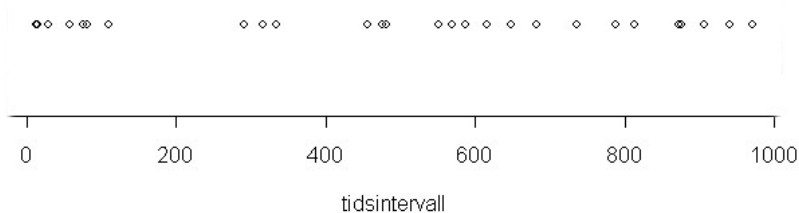


Figure 16.3. During 1000 time intervals the event occurred in 27 intervals

This is just simulation, but this behaviour is also what we could expect in real life. The computer simulation is a very good picture of reality in this case, which is seen in many applications. The assumption that the occurrence of accidents in different time intervals is independent is very realistic if we correct for varying environmental effects like weather and traffic flow. And mathematically, there is a common structure for such different things as radioactive emission, traffic accidents, and telephone calls to a service unit. In other applications there is a random occurrence not in time but in space. But it is the same mathematical structure.

In such situations, it is possible to make a mathematical deduction of a suitable type of distribution to apply. That distribution has been called the Poisson distribution, named after the French mathematician, Siméon Denis Poisson (1781–1840).

The Poisson distribution has only one parameter, which is usually denoted by λ . The probability that the number X of events will be equal to k is given by

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!},$$

UNIVERSITY OF COPENHAGEN



Copenhagen Master of Excellence

Copenhagen Master of Excellence are two-year master degrees taught in English at one of Europe's leading universities

Come to Copenhagen - *and aspire!*

Apply now at
www.come.ku.dk



cultural studies

religious studies

science

for all $k = 0, 1, 2, 3, \dots$. The numerator $k!$ (k -faculty) denotes the product of all natural numbers from 1 up to and including k . For k equal to 0 we have the special definition $0! = 1$.

A wonderful property of that distribution is that it fits many common application situations and that it has a single parameter λ . The expectation in the distribution is λ , and the standard deviation is $\sqrt{\lambda}$.

Events occurring randomly in time
can usually be supposed to have a distribution
where the standard deviation is the square root
of the expectation.

Example

Let us go back to the previous illustration example. When I generated the outcomes in the 1000 intervals I had put the probability of occurrence equal to 0.02 in each interval. It means that the event ought to occur about 20 times in the whole interval. That is an approximate Poisson distribution with parameter $\lambda = 20$. In our simulated interval it happened 27 times. Was that an unexpectedly large number, taking into account that the 'theoretical mean' was 20? Not at all. The standard deviation of the Poisson distribution is equal to the square root of the parameter λ , in our case $\sqrt{20} = 4.47$. The 'overshoot' 7 we got is just 1.56 standard deviations away from the theoretical mean. And since the standard deviation is a kind of mean deviation from the theoretical mean, it is natural that we sometimes get an observation a bit more than the standard deviation away from the theoretical mean. Anything less than 2 standard deviations away from the theoretical mean is quite natural. It might be interesting to see what a Poisson distribution with parameter 20 looks like. The graph is given in the following figure. As you can see, the observation 27 is not very far out in the tail.

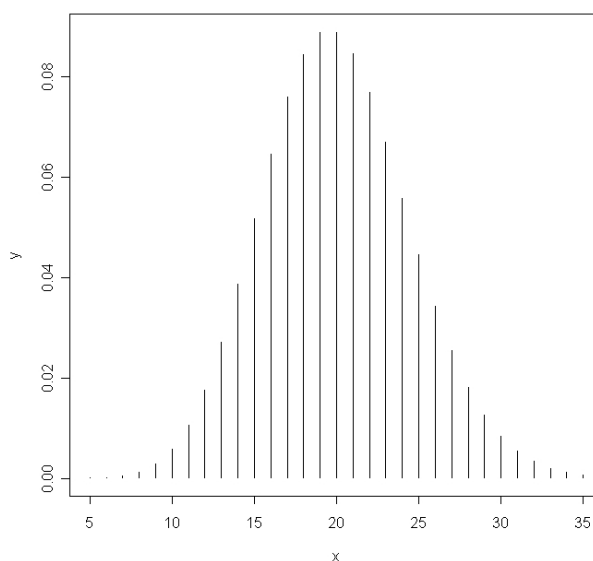


Figure 16.4. Poisson distribution with parameter 20.

Next time you see that some kind of event has occurred a certain number of times in a time interval, let your thoughts go to the Poisson distribution, and in particular to its standard deviation. If for instance something has occurred 9 times in a year, remember that if the expectation is around 9, the standard deviation is around $\sqrt{9} = 3$. Quite a large relative variation, at one third of the expectation. If, however, you have seen a case where an event has occurred 100 times during some time interval, the estimated standard deviation is approximately $\sqrt{100} = 10$, which is only one tenth of the expected value. This is a much lesser relative variation.

Can you always suppose that you have a Poisson distribution when you count how many times an event occurs during a time interval or in a space? No, not always. If you think of the occurrences to be completely random, it is OK fine. But if there is some kind of systematic structure involved, you have reason to be sceptical. For the position of trees along a roadside you may perhaps use the Poisson distribution if the road passes a forest. But in an open landscape with equidistant planted trees by the roadside it does not fit well at all. The number of bacteria in a part of a preparation may well be supposed to have a Poisson distribution in some cases, But if each ‘original’ bacteria may cause groups of ‘secondary’ bacteria around itself, that Poisson distribution model is perhaps not so good.

The reasoning behind the Poisson distribution showed that it has a lot to do with the binomial distribution. This also means that the Poisson distribution may serve as an approximation of the binomial distribution if the p parameter in the binomial distribution is small. In the approximation the Poisson parameter should then equal $\lambda = n \cdot p$. We may also observe now that for small outcomes in the binomial distribution we may use the rule of thumb that the standard deviation is approximately equal to the square root of the expectation, like in the Poisson distribution. This is good to know when judging the quality of numerical binomial data.

Statistical results in the form of natural numbers
can always be suspected to have a big relative variation.
The standard deviation is often approximately equal
to the square root of the observed numbers.



Brain power

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations. Therefore we need the best employees who can meet this challenge!

The Power of Knowledge Engineering

Plug into The Power of Knowledge Engineering.
Visit us at www.skf.com/knowledge

SKF

17 GENERALISED LINEAR MODELS

The simple linear regression is a statistical model which is very easy to understand. The analysis methods we have seen require that the observations are normally distributed with constant variance. Besides the problem of possible curvature which we have discussed earlier, the method can work inefficiently if the distributions are not normal and if the variance is not constant. However, there exists a similar technique, which adapts to distributions like binomial distributions and Poisson distributions. A linear model is linked to the distribution in a suitable way with respect to the properties of that distribution. That is what we call a generalised linear model. Let us look at an introductory example.

Assume that there has been a problem with some kind of production and that work is being done in order to fix the problem. During the last five weeks, the number of produced units and the number of complaints were as follows:

Produced units	204	187	227	173	193
Complaints	26	24	14	4	7

The engineer responsible calculated a regression line with the complaint numbers as data. With x denoting the week number, he derived the regression line estimate $y = 15 - 5.8(x - 3)$ and the following figure for the regression function.

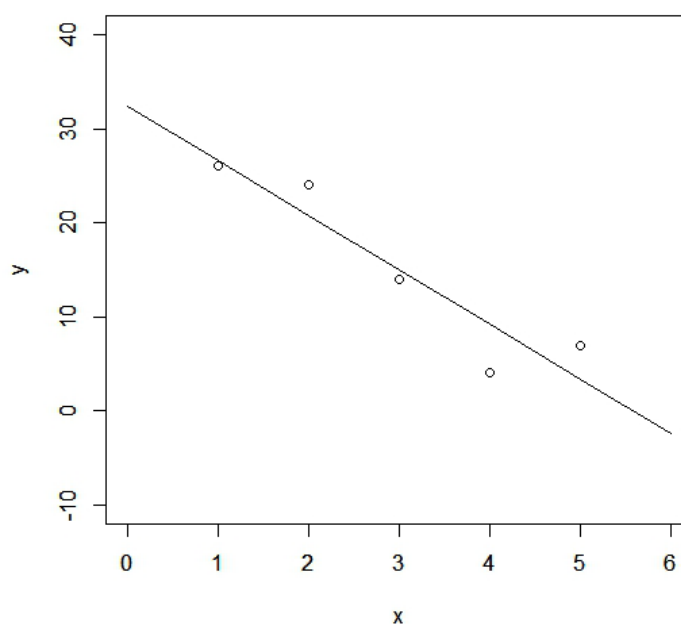


Figure 17.1. A doubtful regression analysis

Now, what is the problem with this analysis, besides the fact that just outside the observation interval the estimate presents a negative value, which is an impossible parameter value? The originally suggested simple linear regression analysis requires normally distributed observations with the same variance. What we have here may be supposed to have binomial distributions. The first observations 26 and 24 may possibly be accepted as approximately normal but the last ones 4 and 7 may certainly not. Further, the variances differ a lot. The first binomial observation 26 (in 204 'trials') may be estimated to have variance of the order of $\frac{1}{204} \frac{26}{204} \left(1 - \frac{26}{204}\right) = 0.00055$, while the fourth observation, 4 (in 173 'trials') has a variance which may be estimated to be $\frac{1}{173} \frac{4}{173} \left(1 - \frac{4}{173}\right) = 0.00013$. The latter is many times smaller than the first one.

How can these inconveniences be overcome and a more reliable regression type of analysis be made? A trick is to make a generalised linear regression. In a model for generalised linear regression, we make a coupling between the parameters of the assumed observation distributions and a linear model in another (abstract) space via a link function. In the example here, we have observations which certainly may be assumed to have binomial parameters. The size parameter n is already exactly known and to the other parameter p we now attach a new abstract parameter η via the so called link function, which for example could be

$$\eta = \log\left(\frac{p}{1-p}\right) \text{ with inverse } p = \frac{e^\eta}{1+e^\eta}.$$

Other link functions are also possible, but this is the most common one. You must observe that when the parameter p goes from 0 to 1, the more artificial parameter η goes all the way from $-\infty$ to ∞ . If we fit a linear function in x to the η values we have a generalised linear model

$$\eta = \alpha + \beta(x - \bar{x}) = \log\left(\frac{p}{1-p}\right).$$

There is no risk here of getting parameters estimated outside the allowed region. In order to estimate the parameters and also to get estimates of their variances, numerical methods are needed. In a suitable statistical program the necessary computational methods are built in. The principle is an iterative procedure to find the parameters, which give the highest probability for the observations we have got. And the program also uses this probability function to get approximate values of variances.

We do not have to worry about this. The program will give us the figures; we just have to make the interpretation. Calculations showed that in the η scale, the estimated regression line was estimated to be $-1.358 - 0.435x$. Further, the standard error of the steepness parameter estimate was 0.095 and an approximate 95% confidence interval $[-0.435 - 1.96 \cdot 0.095; -0.435 + 1.96 \cdot 0.095] = [-0.62; -0.25]$ for the steepness. We may now go back to the probability parameter p via the inverse of the link function. We then get

$$p = \frac{e^\eta}{1 + e^\eta} = \frac{e^{-1.358 - 0.435x}}{1 + e^{-1.358 - 0.435x}}.$$

In the following figure we have probability estimates (relative frequencies) for the five weeks together with the regression estimate in the p scale.

Trust and responsibility

NNE and Pharmaplan have joined forces to create NNE Pharmaplan, the world's leading engineering and consultancy company focused entirely on the pharma and biotech industries.

Inés Aréizaga Esteva (Spain), 25 years old
Education: Chemical Engineer

– You have to be proactive and open-minded as a newcomer and make it clear to your colleagues what you are able to cope. The pharmaceutical field is new to me. But busy as they are, most of my colleagues find the time to teach me, and they also trust me. Even though it was a bit hard at first, I can feel over time that I am beginning to be taken seriously and that my contribution is appreciated.



NNE Pharmaplan is the world's leading engineering and consultancy company focused entirely on the pharma and biotech industries. We employ more than 1500 people worldwide and offer global reach and local knowledge along with our all-encompassing list of services.
nnepharmaplan.com

nne pharmaplan®

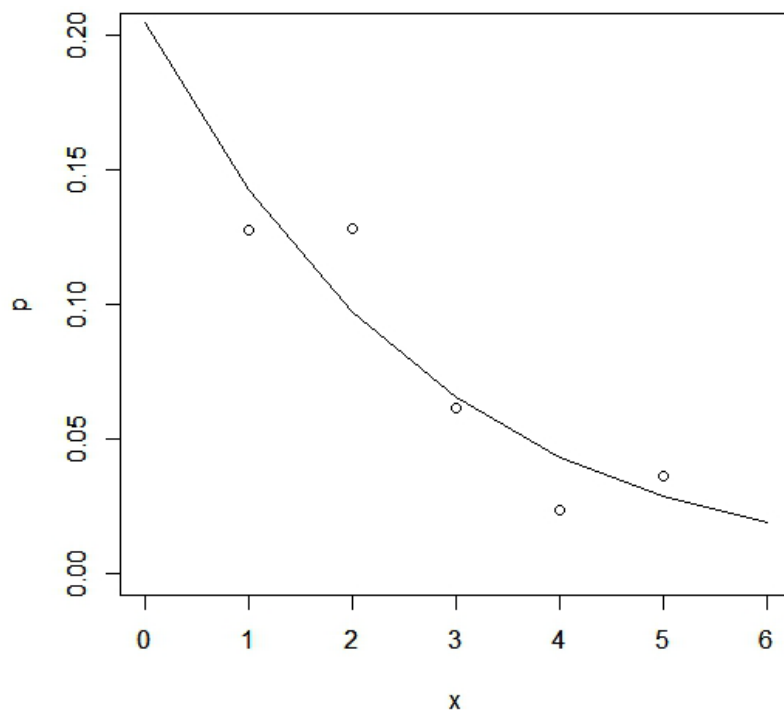


Figure 17.2 Observations of relative frequencies at five time points in an estimate of the regression function obtained with a generalized linear model.

Here we have got an estimate which is more truthful to the character of the data and the experimental situation. For example, the estimated function cannot be negative in any time interval. Different quality of different data is taken into consideration in the estimation procedure by the maximum likelihood method. We still have to think of the realism in, for example, the x -linearity in the generalised linear regression model though. There might be curvature tendencies in a generalised linear regression model just as can be so in an ordinary linear model. This problem can be overcome, for instance, by using a second order polynomial in the generalised linear model instead of the purely linear one.

Generalised linear model methods exist for several discrete and continuous distributions. One example is the discrete Poisson distribution and another one is the continuous exponential distribution which can be used for life spans without aging.

Related to this is the general Cox model for hazard rate dependence on background variables. In this small text we cannot dig deeper into these topics, but I hope that you have got a little understanding of what a generalised linear model is and how it can be used.

18 TO MEASURE THE ALMOST NON-MEASURABLE

The classical methods of physical measurements always relate to some generally accepted normal quantity. I guess you have heard of the archive metre in Paris which had very accurate copies spread over the world. And there were copies of copies all the way down to your own ruler or measuring-tape. Nowadays the definition of metre is instead coupled to the wave length of a certain orange-red light from a krypton-86-atom.

Other units of measurement can be defined in similar ways. A common feature of these definitions is that the measurement unit in natural sciences is very accurately determined.

But how do you define measurement units in the 'soft sciences'? How can you define a measurement unit for pain or happiness? In fact it is not at all natural to define a measurement unit in these cases. What you can best hope for is a good ordered scale which may give the possibility to judge which class in the scale the measure situation belongs to. But even that is not so easy. How can you, for example, make a good ordered scale to judge bleeding in the brain with the help of computerized pictures?

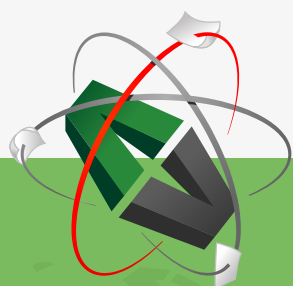
There are different possibilities for getting good quality. Careful verbal descriptions of scale steps are important. If the scale is to be used internationally the translation of the descriptions is a special issue here. One method is to use 'gold standards'. New and less experienced judges can learn by making judgments for some units which are compared to the results of more experienced judges' result for the same units.

Often there have been in use the VAS scales (visual analog scales), which usually consist of a 10 centimeter long line with extreme judgments at the ends. The judgment should be given by a mark on the line. This may seem to be a very accurate method, with this continuous set of alternatives. But that is not the case. The absolute determination of the position between the extremes is left to the judge. This inconvenience is most important in self-judgments, for instance in medical investigations, where the patient himself should make the judgment. It is of course slightly better if a treating doctor makes the judgment for several patients. That may be a little more comparable.

Is it then at all possible to make, for instance, a comparison between two treatments by use of judgment scales? Yes, by careful construction of ordered categorical scales it is possible, after all. In situations where it is possible to make good verbal descriptions of ordered categories it is certainly better to use such a scale than to use a simple continuous VAS scale. The verbal descriptions have a normalising effect.

How can results from ordered categorical scale judgments be analysed in a good way? We may make a good test for comparing two groups of units (e.g. individuals) by use of a technique we have talked about in a previous section. It is quite perfect to use a Mann-Whitney test in order to compare the positions of two 'distributions on ordered categories'. That test is based just on the order of observations. Earlier we discussed the test in connection with continuous data, but also discussed and gave a correction for coinciding values in rounded continuous data, which are in fact discrete. The present case with results in ordered categories is also of a discrete type. It can easily be handled using the correction of the variance in the approximating normal distribution, which we presented before. Here the hypothesis is that the two cases have the same distribution on the ordered categories and the alternative to equality is that there is a systematic tendency.

This e-book
is made with
SetaPDF



PDF components for PHP developers

www.setasign.com



Example

Two medical treatments are compared in an investigation that includes 200 patients. Medicine A is given to 100 randomly selected patients and the remaining patients are given medicine B. The effect is judged in a scale with 4 ordered categories, described in verbal terms. The results are shown in the following.

Category	1	2	3	4
A	11	37	31	21
B	3	23	35	39

In order to test the hypothesis of equal effects of medicines A and B, we just have to go through all categories of one of the medicines, for example B, and sum the number of inversions with medicine A. For equal category we count a half inversion for each pair. The result will be

$$3 \cdot 5,5 + 23 \cdot (11 + 18,5) + 35 \cdot (48 + 15,5) + 39 \cdot (79 + 10,5) = 6408.$$

The hypothesis in the test is that the theoretical distributions are the same for the two medicines. If that hypothesis is true, the sum of inversions ought to be of the order of $100 \cdot 50 = 5000$. It seems that medicine B is better than medicine A. But how big is the variance of the test statistic? We know from an earlier section that the variance, on comparing, is equal to

$$\frac{mn(m+n+1)}{12}(1-\Delta)$$

with its correction term

$$\Delta = \frac{1}{(m+n)(m+n-1)(m+n+1)} \sum_{k=1}^K \tau_k (\tau_k^2 - 1).$$

The numerical value of the correction term here is equal to 0.090 and with that included, we get the standard deviation 390.4 from the first formula. Observe that the formula gives the variance! The random variable which should have an approximate normal distribution with parameters 0 and 1, would get an outcome equal to $\frac{6408 - 5000}{390.4} = 3.607$. A one-sided p value is here $0.00016 = 0.016$ percent. Really small! We dare to state that of the two medicines, B has a better effect.

I hope you remember the ‘juridical’ character of test philosophy. If any ‘opponent’ tries to convince you that B is not favorable in comparison to A, you can defend your statement by telling him that if A and B had the same distribution, then the p value would have a uniform distribution in the interval (0;1). And yet you have got 0.016 percent far out in the lower end. The chances of getting such a small, or smaller value by pure chance is just 0.016 percent.

It is not relevant to work with means and standard deviations for the numbers that possibly mark the ordered categories. Remember that the results are only results on order between observations and that there does not exist any real ‘metric’ in this type of scale measurement. A simple and clear illustration of the two cases of ordered category data is to plot the empirical cumulative distribution functions for the two cases together in the same figure. The cumulative distribution function also works well for data which is not really metric but only has order information. In the following figure you can see the functions for our example.

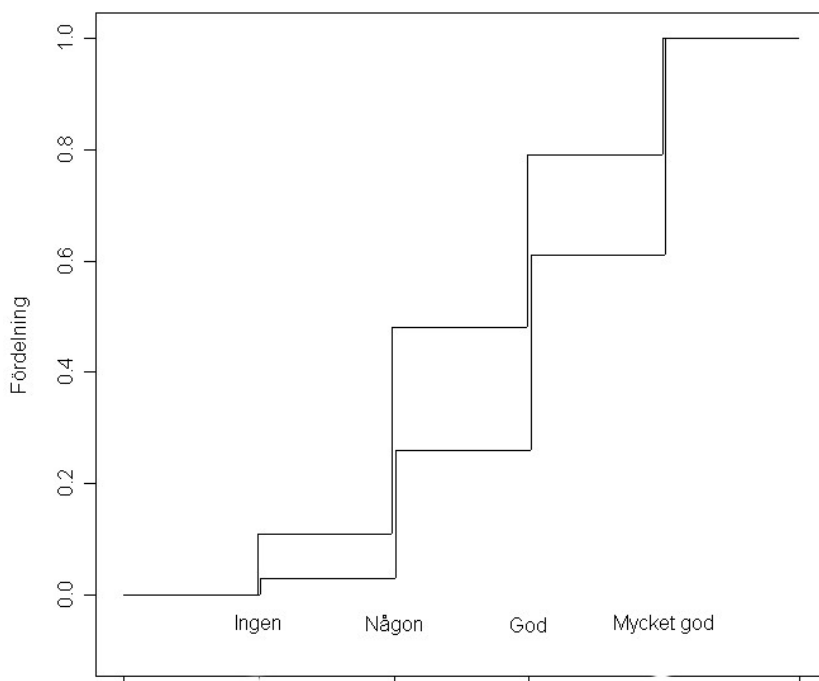


Figure 17.1. Cumulative distribution functions for results of scale judgments for two medicines. The medicine B is the one corresponding to the graph that is lower and more to the right.

It is rather common to come across investigations where several scale judgments are put together in some kind of index, which is a weighted or non-weighted mean of the category number in the basic judgment scale. In particular, if the scale judgments concern different types of questions, it will be completely impossible to really understand the index. What does it mean, for instance, if some category of elderly people has a quality of life index that has decreased from 7.3 to 6.9 in the last 5 years? Is it a question of physical or psychological well-being? Is it a question of problems with health care or is it a question of worsening economical conditions?

My clear opinion is that one should be able to understand the meaning of the measures used in an investigation. And in order for the investigation to have any impact on the real world, we should be able to find out in some detail, where the possible problem lies. By putting all kind of things together in an index, we have no chance of finding out what the problem really is. An ambitious investigation is worthwhile, to present a more detailed picture than just some curious overall index.



FOSS

Sharp Minds - Bright Ideas!

Employees at FOSS Analytical A/S are living proof of the company value - First - using new inventions to make dedicated solutions for our customers. With sharp minds and cross functional teamwork, we constantly strive to develop new unique products - Would you like to join our team?

FOSS works diligently with innovation and development as basis for its growth. It is reflected in the fact that more than 200 of the 1200 employees in FOSS work with Research & Development in Scandinavia and USA. Engineers at FOSS work in production, development and marketing, within a wide range of different fields, i.e. Chemistry, Electronics, Mechanics, Software, Optics, Microbiology, Chemometrics.

We offer
A challenging job in an international and innovative company that is leading in its field. You will get the opportunity to work with the most advanced technology together with highly skilled colleagues.

Read more about FOSS at www.foss.dk - or go directly to our student site www.foss.dk/sharpminds where you can learn more about your possibilities of working together with us on projects, your thesis etc.

Dedicated Analytical Solutions

FOSS
 Slangerupgade 69
 3400 Hillerød
 Tel. +45 70103370
www.foss.dk



19 LIKE A STARRY SKY

Before we had such fast computers and a wide range advanced statistics programs, one had to consult paper tables for the probability distributions needed in analysis. Naturally, the tables could not be too extensive. It was quite common to have, for instance, only 95%, 99% and 99.9% percentiles in the distributions. That meant that the possibilities were essentially to make tests for significance levels 5%, 1% and 0.1%, or to check if the p value was below these fixed levels. It became a convention to use one star (*), two stars (**) and three stars (***) to mark results with p levels below these limits.

Sometimes you could see papers so full of star-marks that the whole paper was reminiscent of the map of a starry sky. The reason for this often was that the authors squeezed their data material too hard, making tests in a lot of subsets and on various questions, hoping to get a good capture of stars. Data without a structured plan and with a common handling of all the error possibilities has often been seen. Perhaps beautiful with all these stars, but not good (useful) at all!

Today when computers and statistics programs are much more powerful, the stars are increasingly substituted by calculated p values. However the basic problem also remains when a lot of p values are calculated. And now I will try to explain what the problem is.

Suppose that we take observations from 10 different cases, for example, 10 groups of individuals. Perhaps when you look at the data you see that there are two cases, where the means differ much. And the p value of a test of equality in position between those two groups, gives a p value which is below the traditional limit of 5%. This ought not to be a surprise. There are in fact $\frac{10 \cdot 9}{2} = 45$ different pairs of cases, when there are 10 cases in all. If there are no real differences between the cases at all, the p values should be random numbers in the interval (0; 1). The probability of getting a particular p value below 5% by pure chance in some of 45 attempts ought not to be surprising. Also, 2 or 3 p values below 5% by pure chance in 45 tests of true hypotheses, is a very reasonable result.

If we think of one-sided tests in the pairs there would be 90 attempts to find a significant result. And if we still use a significant limit of 5% in the individual tests, a number like 5 random rejections is the common result, but for even 8 or 9 rejections is still a natural random outcome if there are no real differences between cases.

Sometimes this phenomenon is called the mass significance problem. To make an unrestrained lot of separate tests on different, more or less planned questions, on a conventional individual level of significance like 5% is an irresponsible method.

How should it be done in a responsible way then, if we really want to handle several detail questions in our investigation? I want to point out that the latter ambition is reasonable if planned in advance and on taking some basic multiple inference theory into consideration.

A simple way to make a judgment on a number of calculated individual p values, is to use an old inequality from the elementary probability theory. There it was called the Boole inequality and when used in statistical problems it is often called the Bonferroni method, after an Italian statistician who was active in the beginning of the 20th century. He advocated the use of this inequality on statistical problems, hence the method is so named.

What does the inequality look like then? Suppose we have n errors of some type and denote their probabilities with $p_1, p_2, p_3, \dots, p_n$. Then the probability that at least one of the errors occurs is, at the most equal to $p_1 + p_2 + p_3 + \dots + p_n$. If n tests have the individual level of significance α/n , the multiple (total) level of significance is at most equal to α . If n confidence intervals have individual confidence degrees $1 - \alpha/n$, the risk that any of the intervals will miss their parameter, at the most equals to $1 - \alpha$.

Example

If we make 5 confidence intervals, all of which have (individual) confidence degree 99%, then the risk that a given (fixed) one of the confidence intervals misses its parameter, is equal to 1%. But (according to Boole-Bonferroni), the risk that any (at least one) of the intervals misses their parameter, at the most equals to $5 \cdot 1\% = 5\%$. The probability that all the confidence intervals hit their parameter is at least equal to 95%. We have got an over-all (multiple) risk of at the most 5% and an over-all (multiple) level of confidence of at least 95%, for all the 5 confidence intervals together.

In the case of tests, I have myself made a refinement of the Bonferroni method which has been much used. Like the original Bonferroni method, it is completely general and can simply be used in any multiple test situation. The underlying principle for this stepwise test is the following. If we have used the Bonferroni method on a number of tests and made some rejections, we may use the same method again on the remaining tests, where there has not been rejection. Since we now work with a smaller number of tests, there is a new chance of rejecting some further hypotheses with a larger limit of the individual p values. And this may be repeated as long as we get new rejections. The method is always as good as the original Bonferroni method, and may give some further rejections, but yet the originally intended multiple level of significance is assured.

Since the method is usually called the Holm-Bonferroni method you may find more complete information on the method by doing a web search on this name. There do exist numerous multiple inference methods for more specific situations, but we will not delve deeper into that here.



"I studied English for 16 years but...
...I finally learned to speak it in just six lessons"

Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download

20 ALL THESE RANKS

I read in my local newspaper that the politicians in my town are very enthusiastic. A yearly ranking has just been presented, of the climate for enterprise in all 290 communes in the country. And our town has got a high rank with 4.1 points on the six-grade scale. That's an increase from the level of 4.0 last year. Happy politicians and town officers celebrate it with a good lunch – far away from tedious things like error margins and other disturbances to good appetite.

So typical, so typical! Not only when it comes to climate for entrepreneurship but also when it comes to service at health care institutions, chirurgical operation results at hospitals, results at schools, and much, much more. A lot of scale point data and ranks, are presented as though they were very precise but completely without analysis of the natural random variations.

Let us take a look at the figures. What does it signify that the mean of the scale numbers is 4.1? The indication 4 on the scale means 'good' and 5 means 'very good' in this case. Does 4.1 mean that most answers in the questionnaire have been good and very few answers have been very good? Or does it mean that there is quite a spread over all the alternatives, very bad, bad, good and very good, with slightly extra weight for good and very good. If the distribution of answers on alternatives were presented, one could gain more insight. For instance, it could reveal if there was in fact a substantial group of entrepreneurs who claimed that the service was very bad. I have looked for such information every time I have come across this type of data presentation. And never found it. I found just all these ranks and possibly some index results.

Means of scale numbers for categories defined by verbal descriptions are generally not good for conveying result information. In such a calculation there is an inherent structure meaning that there is 'the same distance' between all neighbouring categories. It may, however, be more important to distinguish between bad and good than to distinguish between good and very good. Beside this, the mean values are not always easy to interpret. A simple understandable overall measure is obtained by joining the negative answers in one main category and joining the positive answers in one main category. The main measure of the proportion of positive answers is always understandable. And of course this could be completed with two measures on further details, the proportion of very good among all good answers and the proportion of very bad among all bad answers. The information on the answers to a question would be much more meaningful if it were accompanied with an analysis of the quality of the estimates.

As I said, I have never seen results presented as numbers of answers in all categories, which would allow me to calculate some standard errors myself. But in the introductory example here I was surprised to find information on proportions of positive and negative answers in an internet complement not only for the present year but also for the last 11 years. The presented percentages (rounded to a single decimal point) are

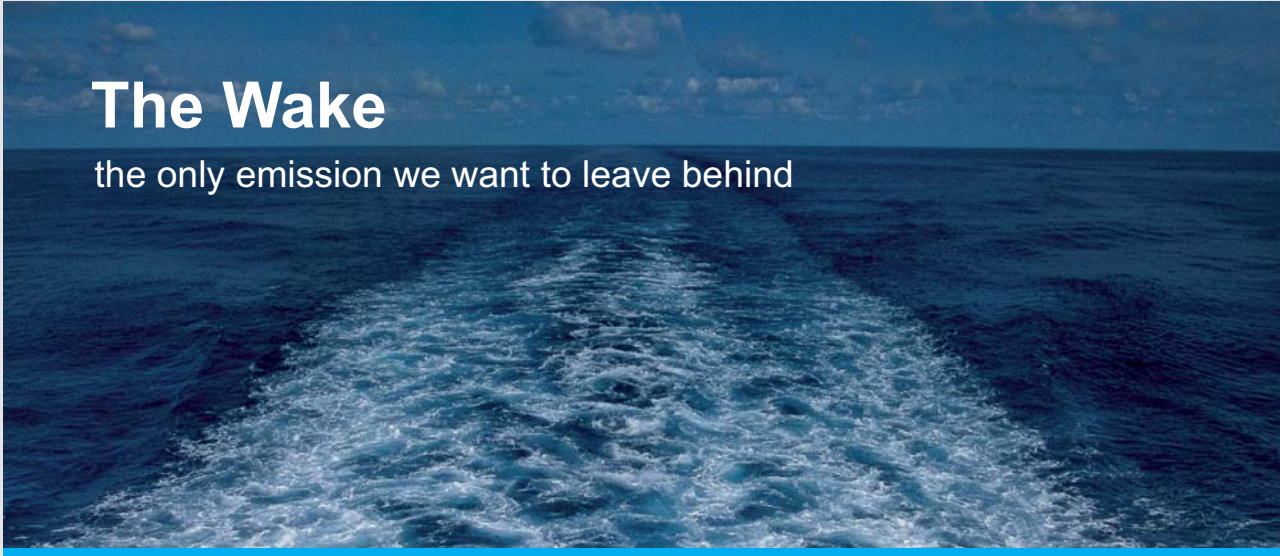
61.3, 76.9, 71.8, 68.8, 72.6, 73.7, 76.0, 70.5, 66.7, 69.0, 68.8.

But tell me about a state of happiness that lasts forever. The percentages are not accompanied with data on the number of answers, which would be needed in order to make calculations for standard errors. However, in a place in the text it is said that the last year there were ‘almost 100 answers’. In order to get some idea of the variation in the series, let us make some calculations for the sample size 100.

For last year we have the positive fraction $0.688 = 68.8\%$. With a sample size 100 this has a standard error $\sqrt{\frac{0.688 \cdot (1 - 0.688)}{100}} = 0.046 = 4.6\%$. A 95% confidence interval for this individual year would then be $[59,2;78,4]$. You may observe that all the other 10 yearly estimates are included in this interval. I dare say that all discussion on trends up and down in the series is just rubbish. But I am sure that if I go to the newspaper archives for the last 10 years I can find much discussion on the ups and downs during this period. The whole series is a good example of a natural, just random variation.

In a presented investigation on the results of a type of surgical operation for different hospitals in Sweden, it so happened that some small hospitals had the top results with the highest fraction of successful operations, while large famous hospitals were not so highly ranked. Are the big and famous Swedish hospitals bad? No, I think they are not. This is probably just a sample size effect. A small hospital has a much smaller number of operations than a big hospital. If both have the same theoretical probability of success for an operation, there is more random variation in the fraction of success for the small hospital. And furthermore, there is quite a number of small hospitals so there is a good chance that a few of them are in the top of the observed fractions. Looking at the lowest observed results, one also finds some small hospitals in the bottom region. As should be the case if there are no real differences but only randomness and different sample sizes.

In order to enlighten the above properly, I have made a small simulation. It consists of 20 small units, 10 medium size units and 5 big units. For the small units I have generated 100 operations' data, for the medium size units I have generated 200 operations' data and for the big units I have generated 300 operations' data. All operation results are supposed to be independent with a 10% risk of extra complications. The data I generated for 5 years gave the following risk estimates.



The Wake


the only emission we want to leave behind

Low-speed Engines Medium-speed Engines Turbochargers Propellers Propulsion Packages PrimeServ

The design of eco-friendly marine power and propulsion solutions is crucial for MAN Diesel & Turbo. Power competencies are offered with the world's largest engine programme – having outputs spanning from 450 to 87,220 kW per engine. Get up front! Find out more at www.mandieselturbo.com

Engineering the Future – since 1758.

MAN Diesel & Turbo



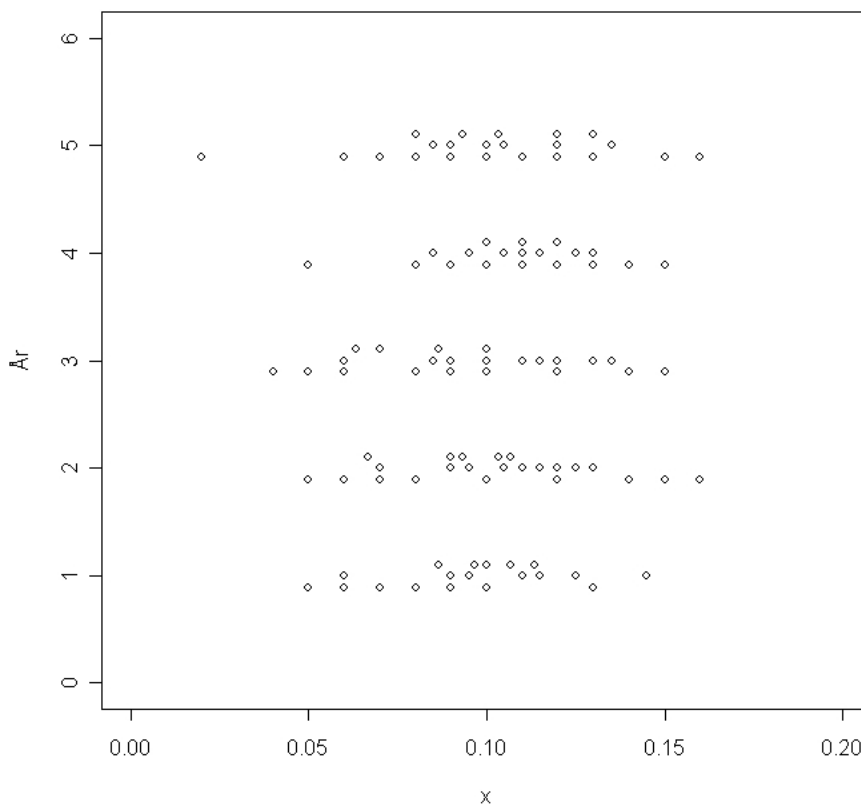


Figure 19.1. Simulated example for the risk of extra complications in 20 small units, 10 medium sized units and 5 big units, for 5 years. For each year the observations for the different types of units are arranged with big units in the upper layer, medium size units in the middle layer and small units in the lower layer. Points in the figure may denote one single observation or some coinciding observations.

In the years 2, 3, 4 and 5, small units have both top results and bottom results. In the first year, a small unit is best (smallest estimated risk) and a medium unit is worst (biggest estimated risk). The big units are quite well ‘centred’ in all five years. Even if these simulations have few units in order not to make the picture blurry, I think it will demonstrate the effect well. Now in this simulation there is no real difference between the units at all but if there are minor differences, it will essentially be the same behavior.

21 SIMULATION, IMPUTATION AND ELIMINATION

You have already been introduced to simulations in the previous sections. They have been used to illustrate how randomness works in different situations.

What is a simulation then? Modern computers can do calculations very fast and among the possible operations is the one for generating random numbers. Basically, random numbers with an equal distribution in the interval 0 to 1 can be generated, but built in mathematical operations make it possible to obtain random numbers with various distributions. By mathematical means, several random outcomes can also be combined in the way we want. And since computers are so fast we may easily run a lot of cases. By studying the results of those runs we may learn a lot about the properties of the random system we have imitated.

Now it is very important to observe here that a simulation in a computer is like a doll's-house or a model railway. It is not the reality; it is just our picture of reality. When we decide how to make a simulation, we also decide what picture of reality we have. And the usefulness of what we learn from the simulation depends on how well our picture depicts reality. Bad conformity between the simulation and reality will make the simulation a worthless result. Just like if you want to use a doll's-house to find a good way of changing the places of your home furniture without carrying around heavy sofas and tables too much, if the doll's-house itself and the furniture in it are not correct in form and scale, then the suggestions from the model study may be a complete disaster in reality.

One example of simulation in statistics is the so called bootstrap technique. The basic principle of bootstrap is simple. In statistical methods we want to calculate quality measures like confidence degrees, variances and significance levels or bias (=systematic error) correction. But in some cases it may be very hard or practically impossible to do that analytically. Then we may do the calculation by way of a careful simulation. And it is very reasonable to make the simulation calculation for the distribution we have got empirically in our observations. The essence of bootstrap technique is just to estimate statistical properties by making a simulation with the obtained empirical distribution as a base. Let us consider a very simple example in order to illustrate how the bootstrap technique works in principle.

Example

For a certain measurement method, there is some doubt that the observations do not have a normal distribution. For 10 measurements in a special situation we have got the following results:

46.2, 44.8, 51.3, 45.6, 52.1, 46.5, 50.1, 45.9, 50.6, 48.1

The mean of the observation series is 48.12 and the standard deviation is 2.68. If we could assume that the measurement errors are normally distributed we could make a confidence interval for the expectation in the common way by using the t distribution with 9 degrees of freedom.

Now we could, for instance, use the formula for the t statistic but make a special table for its distribution by the bootstrap technique. This means that we should study the distribution of the random variable

$$T = \frac{(\bar{X} - \mu)\sqrt{10}}{S}$$

qaiteye[®]
Challenge the way we run

**EXPERIENCE THE POWER OF
FULL ENGAGEMENT...**

.....

**RUN FASTER.
RUN LONGER..
RUN EASIER...**

**READ MORE & PRE-ORDER TODAY
WWW.GAITEYE.COM**

for observations whose distribution is equal to the obtained empirical distribution. Here (in the computer picture of reality) μ is now the mean in the observation series and ‘new bootstrap samples’ will be obtained by choosing 10 observations (with replacement!) from the original observation series. The trick to generate observations in this manner is that we make the ‘computer simulation calculation’ for the empirical distribution function we have obtained.

Now I have generated 10000 such T values. I sorted the results and got the following empirical cumulative distribution function for the T variable.

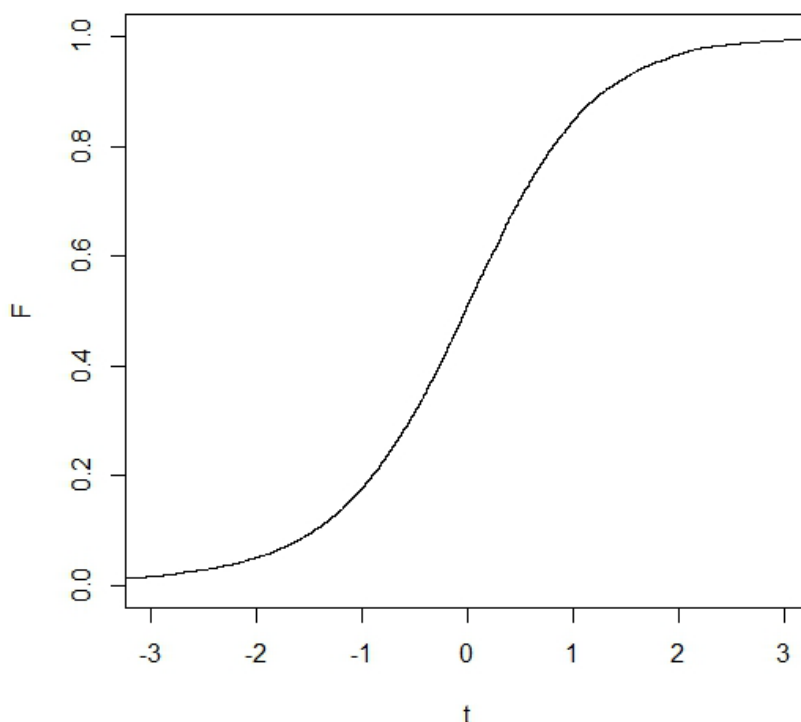


Figure 20.1. Cumulative distribution function for a T variable generated by the bootstrap procedure

At first sight the figure is a reminder of the cumulative distribution function for the T statistic with normally distributed observations. But if we look at the obtained function values, we find, for instance, that the cumulative distribution function equals 0.025 at -2.65 and 0.975 at 2.12. The same values in the T distribution with 9 degrees of freedom are obtained at -2.26 and 2.26. We find that there is some difference, in particular on the negative side. If we use these bootstrap generated table values, we would get the 95% (symmetric) confidence interval

$$\left[48.12 - 2.65 \frac{2.68}{\sqrt{10}}; 48.12 + 2.12 \frac{2.68}{\sqrt{10}} \right] = [45.87; 49.92]$$

for the expected value of the measured values. The interval is made in order to get it to be symmetric in the hitting sense, with same risk of missing below and above. But it appears to be skewed in relation to the point estimate 48.12 in the measurement scale. We can explain it since the observed skew character of the observed data via bootstrap technique has made a small adjustment of the confidence limits.

Even if the example is both small and simple, you can see some characteristic properties. Bootstrap is not really a statistical method but a method to make statistical calculations. In the simple example, the statistical method is to try to make a symmetric confidence interval based on the t type statistic. Bootstrap comes in just as a method to make a probabilistic calculation of 'a suitable table'. If we had been clever enough we could have done that calculation analytically. That, however, would have been extremely tedious and complex in this case, even though the example is small and simple. Bootstrap on the other hand gives us the result in a jiffy once we know a little programming and have written the few lines of code needed.

It is extremely important to be aware of the fact that a bootstrap calculation is of value only if it is well adapted to the situation in real life where we have got the original observations. If for instance, there are a number of observations made on each individual and a bootstrap sample is taken among all observations by random choice with replacement, there will be a disaster. That sort of choice of sample will mean the same as assuming that all observations were independent in an analytical calculation, which is also disastrous since observations on individuals should be assumed to be dependent.

Bootstrap is a technique to calculate
properties of statistical estimates
by simulation in a computer image of reality.
It is important that this picture resembles reality well enough.

There are many other situations where bootstrap calculations can fail because there is a bad adaption to the real life situation. The bootstrap technique uses the empirical distribution as 'a distribution assumption' in the calculation, but otherwise this technique does not in itself make the calculation more application oriented. But as I have said, the bootstrap technique reduces the burden of calculation, and that is very helpful.

Now we leave the bootstrap technique and consider some other simulation techniques. The problem of missing values has always been an irritating problem for people working with analysis of statistical data. A long time ago, when computers were slow and the programs in use were rather primitive, missing values were substituted by hand with reasonable values chosen with regard to known background variables. New values were imputed in the places of missing values. Simple statistical calculations could run smoothly and everything looked fine.

What is then to be said about the principles of imputation? It is very much a question of handling correlated data. To illustrate this, let us consider an example of an extremely simple situation. In a questionnaire investigation, there are a number of distinct groups of participants. Let us suppose that for a certain group there are some participants who have not answered a particular question. Then, perhaps each participant with a missing answer could get an answer imputed by getting a random yes or no, with the probabilities for yes and no given by the proportions of yes and no among those in the group that have answered. The rationality for this is that the imputation strengthens the influence of that particular group; either it has a low, medium or high fraction of yes answers, and that influence may have been weakened by some group members not having answered.



Technical training on *WHAT* you need, *WHEN* you need it

At IDC Technologies we can tailor our technical and engineering training workshops to suit your needs. We have extensive experience in training technical and engineering staff and have trained people in organisations such as General Motors, Shell, Siemens, BHP and Honeywell to name a few.

Our onsite training is cost effective, convenient and completely customisable to the technical and engineering areas you want covered. Our workshops are all comprehensive hands-on learning experiences with ample time given to practical sessions and demonstrations. We communicate well to ensure that workshop content and timing match the knowledge, skills, and abilities of the participants.

We run onsite training all year round and hold the workshops on your premises or a venue of your choice for your convenience.

For a no obligation proposal, contact us today at training@idc-online.com or visit our website for more information: www.idc-online.com/onsite/

- OIL & GAS ENGINEERING**
- ELECTRONICS**
- AUTOMATION & PROCESS CONTROL**
- MECHANICAL ENGINEERING**
- INDUSTRIAL DATA COMMS**
- ELECTRICAL POWER**

Phone: +61 8 9321 1702
 Email: training@idc-online.com
 Website: www.idc-online.com



The imputation makes an adjustment of the obtained data in a reasonable direction. However, there are some drawbacks here. These imputed values are not real observations so they do not have the same quality as those that are. In this simple example they only depict the proportions of yes and no among the group individuals who have given the answer. They cannot contribute to a better knowledge of the real theoretical proportions in the group. If the imputed values now are handled as real observations in a later analysis, the error estimates will be too optimistic. The standard errors will be estimated as smaller than they really are. A reliable estimate of standard error would have paid more attention to the differing qualities of the real observations and the artificial ones.

In investigations of knowledge levels in schools, quite often so called plausible values are used, which are a type of imputation. This technique is more complicated than simple imputation, but the basic structure is the same. Artificial computer generated observations are calculated from correlated information and used in later analysis as though they were real observations. In these investigations, often a great number of artificial data entries are generated and used in further analysis.

When information is missing it may
to some extent be obtained from
correlated information. However this does not
have the same quality as that of real observations.

Deviating observations should never be rejected just because they are deviating. Well if a careful investigation shows that something has happened, that really explains that the observation is not valid, then it may of course be skipped. There may, for instance, have been a break in the electricity supply during experimentation or there is proof that the wrong measurement equipment was being used and so on. But otherwise, never reject odd observations. And be alert to others who skip obtained observations

Odd observations are often called outliers. People collect statistical data and intend to make some simple statistical analysis like regression or analysis of variance. Then suddenly they get worried if they find some observation which does not fit the picture. They have got a problem, which seems to solve itself if that disturbing outlier is eliminated. Everything appears to be smooth and fine.

Quite often, however, some essential information is lost when the deviating observation is excluded. One such situation where it may happen is the following. The data consists of positive continuous numbers and a standard statistical method (originally constructed for normally distributed data) is used. Now it seems that the basic application's suitability for the normal distribution method for comparison is not satisfied. One of the observation series has an outlier, which contributes very much to the fact that variances in the two series differ quite a bit. It seems that the series with greater observation values also has considerably greater variance than the other one.

Let us think of the suitability of the use of the normal distribution, which is that the observations could be thought of as **sums of many additive** effects. For positive random variables, it is more natural to think of variables as **products of many multiplicative** effects. This rationale leads instead to the so called lognormal distribution which is nice and smooth but has a longer upper tail than the normal distribution. For systems with different types of lognormal observations, it is most natural to talk of multiplicative parameters instead of the additive parameters we use in standard normal distribution methods. The technical analysis of models for lognormal observations is done simply to make the analysis for the logarithms of the positive observations in the usual way. And there is a good chance that the analysis works smoothly without discarding any observations. The analysis fits well in situations with positive data where it is natural to think of relative variations.

Example

In an experiment, two methods are to be compared. There are 9 observations made for each case, and the results are

X: 4.49, 5.01, 5.99, 7.20, 6.23, 7.17, 5.55, 6.14, 4.14

Y: 8.78, 6.09, 9.53, 7.89, 11.10, 9.86, 7.96, 10.50, 9.53

The investigators intended to make a 95% confidence interval for the difference in expectations in the ordinary normal model. Since the means and standard deviations are

$$M_X = 5.77 \quad S_X = 1.08 \quad M_Y = 9.03 \quad S_Y = 1.53,$$

the confidence interval was obtained as

$$9.03 - 5.77 \pm 2.12 \sqrt{\frac{1}{9} + \frac{1}{9}} \sqrt{\frac{8 \cdot 1.08^2 + 8 \cdot 1.53^2}{16}} = 3.16 \pm 1.87 = [1.29; 5.03].$$

Everything seemed just fine, but then the investigators calculated the residuals in the series, that is, the difference between observations and the mean in the series. The result was

X series: -1.28, -0.76, 0.22, 1.43, 0.46, 1.40, -0.22, 0.37, -1.63

Y series: -0.25, -2.94, 0.50, -1.14, 2.07, 0.83, -1.07, 1.47, 0.50

The second observation in the Y series is extremely small compared to other numbers in that series and the investigators intended to remove that observation as an unwanted outlier.

Is it appropriate to do so?

No, I think it isn't. As I pointed out above, it is common for positive data to have the character that greater values also have greater variation. This is also indicated in the Y series having estimated standard deviation 1.53, while the X series has 1.08. A multiplicative model and assumption of lognormal distributions would lead to an analysis of the logarithms of the observations. The logarithms of the values in the series are

Log X series: 1.50, 1.61, 1.79, 1.97, 1.83, 1.97, 1.71, 1.81, 1.42

Log Y series: 2.17, 1.81, 2.25, 2.07, 2.41, 2.29, 2.03, 2.35, 2.25

I joined MITAS because
I wanted **real responsibility**

The Graduate Programme
for Engineers and Geoscientists
www.discovermitas.com



Month 16
I was a construction supervisor in the North Sea advising and helping foremen solve problems

Real work
International opportunities
Three work placements



MAERSK

Still of course the second observation is the smallest one, but now the difference from the mean for the series is not so drastic

Difference from mean for the log Y series: -0.01, -0.37, 0.07, -0.11, 0.23, 0.11, -0.15, 0.17, 0.07

In this analysis there is no reason to be surprised at the deviation. And the two log series have rather similar standard deviations. A common normal theoretical 95% confidence interval for the difference will be $0.45 \pm 0.19 = [0.26; 0.64]$. This is an interval for the logarithm of the multiplicative factor, and the confidence interval for the factor itself will be $[e^{0.26}; e^{0.64}] = [1.30; 1.90]$.

Another reason for not getting tempted to exclude deviating observations is that the model used in the analysis may not fit reality so well. The model may for instance need refinement. Excluding a deviation observation then may lead to a quite different estimate, which is still bad because the model is then not accurate. Let us look at a very simple example.

Example

In a regression situation, measurements of a variable y have been taken for the nine x values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. The results were

1.03, 0.75, 0.82, 0.78, 0.73, 0.70, 0.72, 0.88, 0.93.

A straight line was fitted to the observations. The gradient was estimated at -0.029. The deviations between the observations and the line were also calculated. The result was

0.2024, -0.0747, -0.0018, -0.0389, -0.0860, -0.1131, -0.0902, 0.0727, 0.1256.

The first deviance is substantially greater than “the normal ones” so it was considered to be an outlier and excluded. A new calculation with only eight pairs (x, y) gave a gradient of +0.161. Quite a difference between the first value and this one and even a changed sign! What happened?

In fact, linear regression is an unsuitable model here. We can see it in a pictorial representation of the observations pairs here, together with the simple linear regression estimate for all observations and also with the estimated regression line for the reduced set of observations, where the first observation has been removed.

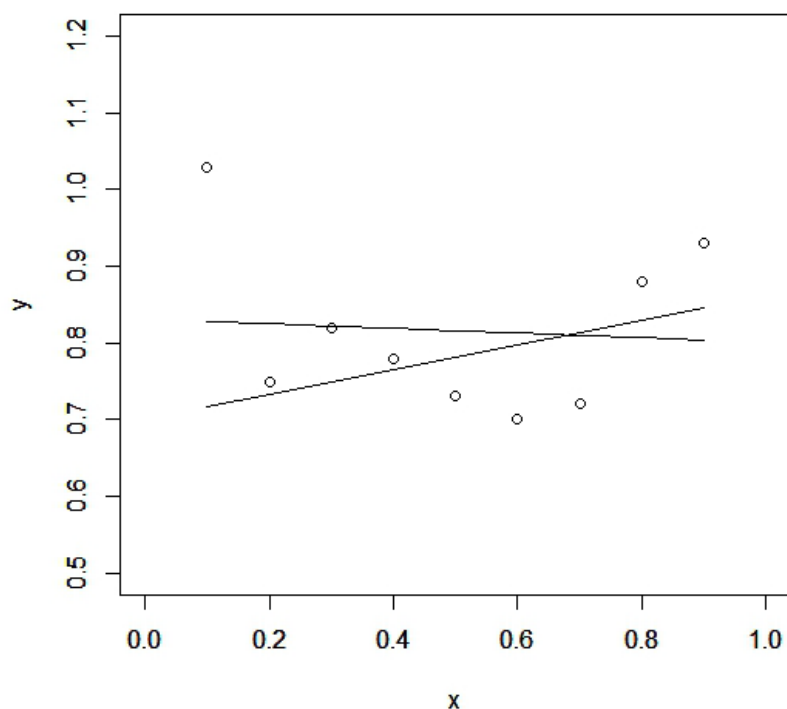


Figure 20.2. A data set with simple linear regression estimate for all data and with the first data point omitted.

The simple linear model (straight line) is not suitable here. The figure for the observations clearly indicates that the regression function may be convex. One ought to have a model which includes the possibility of curvature. The simplest would be the second order polynomial.

Deviating observations should not be removed
just because they deviate.

They may instead imply further investigation of
the used statistical model and statistical method.

WHERE IS THE CONCEPT DEFINED?

A

analysis of variance 7, 99, 100, 101, 102, 103, 105, 106, 139, 144

B

binomial coefficient 20

binomial distribution 19, 20, 22, 23, 28, 41, 67, 68, 117

Bonferroni 128, 129

Boole 128

bootstrap 7, 134, 135, 136, 137, 138

C

central limit theorem 56

chi-square distribution 70

classical probability model 11, 12, 15

combinatorial 11

confidence degree 27, 28, 29, 30, 58, 59, 71, 72, 75, 89, 128

confidence interval 9, 24, 26, 27, 28, 29, 30, 58, 60, 61, 62, 71, 72, 73, 75, 76, 88, 89, 91, 120, 131, 135, 136, 137, 140, 142

Cox model 121

D

degrees of freedom 70, 71, 73, 74, 75, 76, 86, 87, 95, 97, 100, 101, 103, 105, 106, 135, 136

density 24, 25, 70, 73, 85, 86

E

empirical 16, 23, 31, 33, 34, 35, 38, 39, 40, 45, 55, 59, 60, 61, 62, 67, 70, 72, 75, 76, 80, 86, 87, 88, 100, 107, 108, 110, 125, 134, 136, 137

error margin 8, 9, 57, 59, 60, 61, 62, 76

www.job.oticon.dk

oticon
PEOPLE FIRST

expectation 31, 32, 33, 34, 35, 36, 37, 40, 41, 42, 43, 46, 47, 50, 51, 53, 55, 58, 59, 65, 67, 73, 75, 76, 85, 92, 95, 101, 107, 115, 116, 117, 135

F

F distribution 101, 103, 106

frequency function 24, 25, 26, 30, 47, 50, 52, 53, 54, 55

G

Gauss 47

generalised linear regression 98, 119, 121

gold standards' 122

H

hitting probability 27, 29, 71

hypothesis 63, 64, 65, 66, 67, 68, 69, 76, 80, 82, 84, 87, 88, 89, 90, 97, 100, 101, 103, 123, 124

I

imputation 7, 134, 138, 139

independent 17, 18, 19, 23, 26, 34, 40, 41, 45, 47, 51, 52, 53, 56, 58, 62, 70, 71, 77, 85, 88, 94, 100, 107, 109, 114, 132, 137

inversion 78, 80, 83, 84, 85, 124

L

least squares method 94

level of significance 64, 65, 66, 67, 68, 69, 76, 77, 82, 89, 90, 91, 100, 103, 106, 128, 129

link function 98, 119, 120

location measure 31, 35

M

Mann-Whitney test 78, 79, 80, 82, 83, 88

mass significance 128

mean 9, 10, 16, 28, 31, 32, 33, 34, 35, 38, 39, 40, 44, 45, 46, 47, 51, 53, 54, 55, 58, 60, 61, 65, 67, 76, 79, 80, 84, 85, 87, 92, 95, 100, 102, 103, 107, 110, 111, 112, 113, 115, 126, 130, 135, 136, 137, 141, 142

median 25, 26, 27, 28, 29, 31

missing values 138

multiple inference 128, 129

N

normal distribution 47, 50, 51, 52, 53, 54, 55, 56, 58, 59, 61, 65, 67, 70, 73, 74, 75, 77, 79, 83, 92, 123, 124, 135, 140

O

obtained level of significance 66, 67

one-sided 89, 90, 91, 97, 103, 124, 127

one sided test 79

ordered scale 122

outliers 139

P

parameter 10, 20, 23, 30, 36, 43, 44, 51, 58, 61, 73, 75, 76, 83, 86, 87, 88, 89, 91, 92, 94, 95, 98, 100, 114, 115, 116, 117, 119, 120, 128

parameter linear regression 98

Poisson 114, 115, 116, 117, 118, 121

Poisson distribution 114, 115, 116, 117, 121

power 64, 68, 69, 82, 88

R

regression model 92, 104, 121

relative frequency 9, 16, 43, 44, 80

S

simple linear regression 92, 97, 104, 118, 119, 142, 143

simulation 28, 110, 111, 113, 114, 132, 133, 134, 136, 137, 138

standard deviation 36, 37, 38, 40, 41, 43, 44, 45, 46, 47, 50, 51, 53, 54, 58, 59, 60, 61, 62, 65, 67, 69, 70, 71, 72, 73, 75, 76, 79, 83, 84, 85, 86, 88, 91, 92, 94, 100, 115, 116, 117, 124, 135, 141

standard error 46, 58, 59, 120, 131, 139

stepwise test 129

systematic errors 40, 107

T

t distribution 73, 74, 75, 76, 87, 97, 135
test 16, 23, 63, 64, 65, 66, 67, 68, 69, 70, 76, 77,
78, 79, 80, 82, 83, 85, 87, 88, 89, 90, 91, 97,
100, 101, 103, 106, 123, 124, 125, 127, 129
theoretical 7, 16, 23, 26, 31, 33, 34, 35, 36, 38,
40, 41, 43, 59, 61, 67, 69, 71, 72, 73, 75, 77,
85, 86, 87, 88, 92, 93, 98, 100, 105, 106, 107,
108, 111, 112, 113, 115, 124, 131, 139, 142
two-sided 89, 90, 91

U

unbiased 40

V

variance 7, 36, 38, 40, 59, 62, 83, 84, 86, 88, 94,
95, 99, 100, 101, 102, 103, 104, 105, 106, 107,
108, 118, 119, 123, 124, 139, 140
variation between 107
variation within 9, 107
VAS scales 122

W

weighted empirical variance 86
Wilcoxon test 77, 83