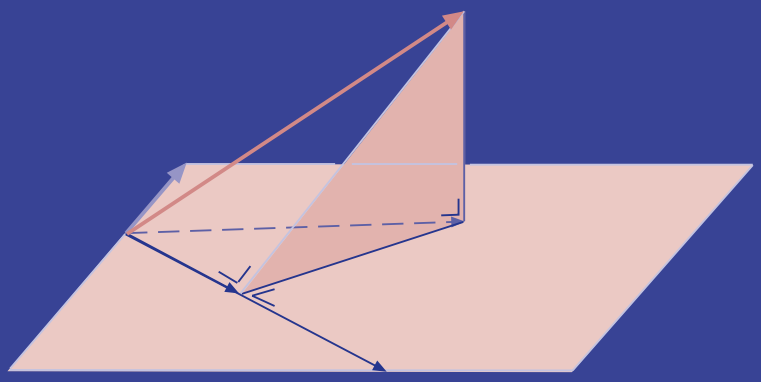
Testing theory

an introduction

P.J.G. Teunissen



Series on Mathematical Geodesy and Positioning

Testing theory
an introduction

Testing theory
an introduction

P.J.G. Teunissen

Delft University of Technology

Department of Mathematical Geodesy and Positioning

VSSD

Series on Mathematical Geodesy and Positioning
http://www.vssd.nl/hlf/a030.htm

Adjustment Theory
P.J.G. Teunissen
2003 / 201 p. / ISBN 90-407-1974-8

Dynamic Data Processing
P.J.G. Teunissen
2001 / 241 + x p. / ISBN 90-407-1976-4

Testing Theory
P.J.G. Teunissen
2000 / 147+viii p. / ISBN 90-407-1975-6

Hydrography
C.D. de Jong, G. Lachapelle, S. Skone, I.A. Elema
2003 / x+351 pp. / ISBN 90-407-2359-1 / hardback

First edition 2000-2006

Published by:
VSSD
Leeghwaterstraat 42, 2628 CA Delft, The Netherlands
tel. +31 15 278 2124, telefax +31 15 278 7585, e-mail: hlf@vssd.nl
internet: http://www.vssd.nl/hlf
URL about this book: **http://www.vssd.nl/hlf/a030.htm**

A collection of digital pictures and an elctronic version can be made available for lecturers who adopt this book. Please send a request by e-mail to hlf@vssd.nl

Printed version: ISBN 978-90-407-1975-2 Ebook: ISBN 978-90-6562-216-7
NUR 930

Keywords: testing theory, geodesy

Foreword

This book is based on the lecture notes of the course 'Testing theory' (Inleiding Toetsingstheorie) as it has been offered since 1989 by the Department of Mathematical Geodesy and Positioning (MGP) of the Delft University of Technology. This course is a standard requirement and is given in the second year. The prerequisites are a solid knowledge of adjustment theory together with linear algebra, statistics and calculus at the undergraduate level. The theory and application of least-squares adjustments are treated in the lecture notes *Adjustment theory* (Delft University Press, 2000). The material of the present course is a follow up on this course on adjustment theory. Its main goal is to convey the knowledge necessary to be able to judge and validate the outcome of an adjustment. As in other physical sciences, measurements and models are used in Geodesy to describe (parts of) physical reality. It may happen however, that some of the measurements or some parts of the model are biased or in error. The measurements, for instance, may be corrupted by blunders, or the chosen model may fail to give an adequate enough description of physical reality. These mistakes can and will occasionally happen, despite the fact that every geodesist will try his or her best to avoid making such mistakes. It is therefore of importance to have ways of detecting and identifying such mistakes. It is the material of the present lecture notes that provides the necessary statistical theory and testing procedures for resolving situations like these.

Following the *Introduction*, the basic concepts of statistical testing are presented in *Chapter 1*. In *Chapter 2* the necessary theory is developed for testing *simple* hypotheses. As opposed to its *composite* counterpart, a simple hypothesis is one which is completely specified, both in its functional form as well as in the values of its parameters. Although simple hypotheses rarely occur in geodetic practice, the material of this chapter serves as an introduction to the chapters following. In *Chapter 3*, the generalized likelihood ratio principle is used to develop the theory for testing composite hypotheses. This theory is then worked out in detail in *Chapter 4*, for the important case of linear(ized) models. Both the parametric form (observation equations) and the implicit form (condition equations) of linear models are treated. Five different expressions are given for the uniformly, most powerful, invariant teststatistic. As an additional aid in understanding the basic principles involved, a geometric interpretation is given throughout. This chapter also introduces the important concept of reliability. The internal and external reliability measures given, enable a user to determine in advance (i.e. at the designing stage, before the actual measurements are collected) the size of the minimal detectable biases and the size of their potential impact on the estimated parameters of interest.

Many colleagues of the Department of Mathematical Geodesy and Positioning whose assistance made the completion of this book possible are greatly acknowledged. C.C.J.M. Tiberius took care of the editing, while the typing was done by Mrs. J. van der Bijl and Mrs. M.P.M. Scholtes. The drawings were made by Mr. A.B Smits and the statistical tables were generated by Mrs. M. Roselaar. Various lecturers have taught the book's material over the past years. In particular the feedback and valuable recommendations of G.J. Husti, F. Kenselaar and N.F. Jonkman are acknowledged.

P.J.G. Teunissen
June, 2000

Contents

Introduction

The present lecture notes are a follow up on the book *Adjustment theory* (Delft University Press, 2000). Adjustment theory deals with the optimal combination of redundant measurements together with the estimation of unknown parameters. There are two main reasons for performing redundant measurements. First, the wish to increase the accuracy of the results computed. Second, the requirement to be able to check for mistakes or errors. The present book addresses this second topic.

In order to be able to adjust redundant observations, one first needs to choose a mathematical model. This model consists of two parts, the functional model and the stochastic model. The functional model contains the set of functional relations the observables are assumed to obey. For instance, when the three angles of a triangle are observed and when it is assumed that the laws of planar Euclidean geometry apply, the three angles should add up to π. However, since measurements are intrinsically uncertain (perfect measurements do not exist), one should also take the unavoidable variability of the measurements into account. This is done by means of a stochastic model in which the measurement uncertainty is captured through the use of stochastic (or random) variables. In most geodetic applications it is assumed that the results of measurement, the observations, are independent samples drawn from a normal (or Gaussian) distribution.

Once the mathematical model is specified, one can proceed with the adjustment. Although different methods of adjustment exist, one of the leading principles is the principle of least-squares (for a brief account on the early history of adjustment, see Appendix D). Apart from the fact that (linear) least-squares estimators are relatively easy to compute, they also possess two important properties, namely the property of unbiasedness and the property of minimum variance. In layman terms one could say that least-squares solutions coincide with their target value on the average (property of unbiasedness), while the sum of squares of their unavoidable, individual variations about this target value will be the smallest possible on the average (property of minimum variance). These two properties only hold true, however, under the assumption that the mathematical model is correct. They fail to hold in case the mathematical model is misspecified. Errors or misspecifications in the functional model generally result in least-squares estimators that are biased (off target). Similarly, misspecifications in the stochastic model will generally result in least-squares estimators that are less precise (larger variations).

Although one always will try one's best to avoid making mistakes, they can and will occasionally happen. It is therefore of importance to have ways of *detecting* and *identifying* such mistakes. In this book we will restrict ourselves and concentrate only on developing methods for detecting and identifying errors in the functional model. Hence, throughout this book the stochastic model is assumed to be specified correctly. This restriction is a legitimate one for many geodetic applications. From past experience we know that if modelling errors occur, they usually occur in the functional model and not so much in the stochastic model. Putting the exceptions aside, one is usually quite capable of making a justifiable choice for the stochastic model. Moreover, mistakes made in the functional model usually have more serious consequences for the results computed than errors made in the stochastic modelling.

Mistakes or errors in the functional model can come in many different guises. At this point it is of importance to realize, since every model is a caricature of reality, that every model has its shortcomings. Hence, strictly speaking, every model is already in error to begin with. This shows that the notion of a modelling error or a model misspecification has to be considered with some care. In order to understand this notion, it helps if one accepts that the presence of modelling errors can only be felt in the confrontation between data and model. We therefore speak of a modelling error when the discrepancies between the observations and the model are such that they can not be explained by, or attributed to, the unavoidable measurement uncertainty. Such discrepancies can have many different causes. They could be caused by mistakes made by the observer, or by the fact that defective instruments are used, or by wrong assumptions about the functional relations between the observables. For instance, in case of levelling, it could happen that the observer made a mistake when reading off the leveling rod, or in case of direction measurements, it could happen that the observer accidentally aimed the theodolite at the wrong point. These types of mistakes affect individual observations and are usually referred to as blunders or gross errors. Instead of a few individual observations, whole sets of observations may become affected by errors as well. This happens in case defective instruments are used, or when mistakes are made in formulating the functional relations between the observables. Errors with a common cause that affect whole sets of observations are sometimes referred to as systematic errors.

The goal of this book is to convey the necessary knowledge for judging the validity of the model used. Typical questions that will be addressed are: 'How to check the validity of a model? How to search for certain mistakes or errors? How well can errors be traced? How do undetected errors affect the final results?' As to the detection and identification of errors, the general steps involved are as follows:

(i) One starts with a model which is believed to give an adequate enough description of reality. It is usually the simplest model possible which on the basis of past experience has proven itself in similar situations. Since one will ordinarily assume that the measurements and the modelling are done with the utmost care, one is generally not willing, at this stage, to already make allowances for possible mistakes or errors. This is of course an assumption or an hypothesis. This first model is therefore referred to as the *null hypothesis*.

(ii) Since one can never be sure about the absence of mistakes or errors, it is always wise to check the validity of the null hypothesis once it has been selected. Hence, one would like to be able to *detect* an untrustworthy null hypothesis. This is possible in principle, when redundant measurements are available. From the adjustment of the redundant measurements, (least-squares) residuals can be computed. These residuals are a measure of how well the measurements fit the model of the null hypothesis. Large residuals are often indicative for a poor fit, while smaller residuals tend to correspond with a better fit. These residuals are therefore used as input for deciding whether or not one is willing to accept the null hypothesis.

(iii) Would one decide to reject the null hypothesis, one implicitly states that the measurements do not seem to support the assumption that the model under the null hypothesis gives an adequate enough description of reality. One will therefore have to look for an alternative model or an *alternative hypothesis*. It very seldom happens

however, that one knows beforehand which alternative to consider. After all, many different errors could have led to the rejection of the null hypothesis. This implies that in practice, instead of considering a single alternative, usually various alternatives will have to be considered. And since different types of errors may occur in different situations, the choice of these alternatives very much depends on the particular situation at hand.

(iv) Once it has been decided which alternatives to consider, one can commence with the process of *identifying* the most likely alternative. This in fact boils down to a search of the alternative hypothesis which best fits the measurements. Since each alternative hypothesis describes a particular mistake or modelling error, the most likely mistake corresponds with the most likely hypothesis. Once one is confident that the modelling errors have been identified, the last step consists of an *adaptation* of the data and/or model. This implies either a re-measurement of the erroneous data or the inclusion of additional parameters in the model such that the modelling errors are accounted for.

It will be intuitively clear that not all errors can be traced equally well. Some errors are better traceable than others. Apart from being able of executing the above steps for the detection and identification of modelling errors, one would therefore also like to know how well these errors can be traced. This depends on the following factors. It depends on the model used (the null hypothesis), on the type and size of the error (the alternative hypothesis), and on the decision procedure used for accepting or rejecting the null hypothesis. Since these decisions are based on uncertain measurements, their outcomes will be to some degree uncertain as well. As a consequence, two kinds of wrong decisions can be made. One can decide to reject the null hypothesis, while in fact it is true (wrong decision of the 1st kind), or one can decide to accept the null hypothesis, although it is false (wrong decision of the 2nd kind). In the first case, one wrongly believes that a mistake or modelling error has been made. This might then lead to an unnecessary re-measurement of the data. In the second case, one wrongly believes that mistakes or modelling errors are absent. As a consequence, one would then obtain biased adjustment results. These issues and how to cope with them, will also be discussed in this book. Once mastered, they will enable one to formulate guidelines for the *reliable* design of measurement set-ups.

1 Basic concepts of hypothesis testing

1.1 Statistical hypotheses

Many social, technical and scientific problems result in the question whether a particular theory or hypothesis is true or false. In order to answer this question one can try to design an experiment such that its outcome can also be predicted by the postulated theory. After performing the experiment one can then confront the experimental outcome with the theoretically predicted value and on the basis of this comparison try to conclude whether the postulated theory or hypothesis should be rejected. That is, if the outcome of the experiment disagrees with the theoretically predicted value, one could conclude that the postulated theory or hypothesis should be rejected. On the other hand, if the experimental outcome is in agreement with the theoretically predicted value, one could conclude that as yet no evidence is available to reject the postulated theory or hypothesis.

Example 1

According to the postulated theory or hypothesis the three points 1, 2 and 3 of Figure 1.1 lie on one straight line. In order to test or verify this hypothesis we need to design an experiment such that its outcome can be compared with the theoretically predicted value.

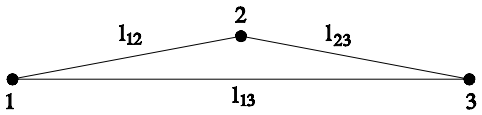


Figure 1.1: Three points on a straight line.

If the postulated hypothesis is correct, the three distances l_{12}, l_{23} and l_{13} should satisfy the relation:

$$l_{13} = l_{12} + l_{23}.$$

Thus, under the assumption that the hypothesis is correct we have:

(1) $$H: \quad l_{12} + l_{23} - l_{13} = 0.$$

To denote a hypothesis, we will use a capital H followed by a colon that in turn is followed by the assertion that specifies the hypothesis. As an experiment we can now measure the three distances l_{12}, l_{23} and l_{13}, compute $l_{12} + l_{23} - l_{13}$ and verify whether this computed value agrees or disagrees with the theoretically predicted value of H. If it agrees, we are inclined to accept the hypothesis that the three points lie on one straight line. In case of disagreement we are inclined to reject hypothesis H.

It will be clear that in practice the testing of hypotheses is complicated by the fact that experiments (in particular experiments where measurements are involved) in general do not give outcomes that are exact. That is, experimental outcomes are usually affected by an amount of uncertainty, due for instance to measurement errors. In order to take care of this uncertainty, we will, in analogy with our derivation of estimation theory in "Adjustment theory", model the uncertainty by making use of the results from the theory of random variables. The verification or testing of postulated hypotheses will therefore be based on the testing of hypotheses of random variables of which the probability distribution depends on the theory or hypothesis postulated. From now on we will therefore consider statistical hypotheses.

A *statistical hypothesis* is an assertion or conjecture about the probability distribution of one or more random variables, for which it is assumed that a random sample (mostly through measurements) is available.

The structure of a statistical hypothesis H is in general the following:

(2) $$H: \quad \underline{y} \sim p_{\underline{y}}(y|x) \text{ with } x \text{ fully or partially specified.}$$

This statistical hypothesis should be read as follows: According to H the scalar or vector observable random variable $\underline{y}$ has a probability density function given by $p_{\underline{y}}(y|x)$. The scalar, vector or matrix parameter x used in the notation of $p_{\underline{y}}(y|x)$ indicates that the probability density function of $\underline{y}$ is known except for the unknown parameter x. Thus, by specifying (either fully or partially) the parameter x, an assertion or conjecture about the density function of $\underline{y}$ is made. In order to see how a statistical hypothesis for a particular problem can be formulated, let us continue with our Example 1.

Example 1 (continued)

We know from experience that in many cases the uncertainty in geodetic measurements can be adequately modelled by the normal distribution. We therefore model the three distances between the three points 1, 2 and 3 as normally distributed random variables [1]. If we also assume that the three distances are uncorrelated and all have the same known variance $\frac{1}{3}\sigma^2$, the simultaneous probability density function of the three distance observables becomes:

[1] Note that strictly speaking distances can never be normally distributed. A distance is always nonnegative, whereas the normal distribution, due to its infinite tails, admits negative sample values.

$$(3) \qquad \begin{pmatrix} \underline{l}_{13} \\ \underline{l}_{12} \\ \underline{l}_{23} \end{pmatrix} \sim N\left(\begin{pmatrix} E\{\underline{l}_{13}\} \\ E\{\underline{l}_{12}\} \\ E\{\underline{l}_{23}\} \end{pmatrix}, Q \right) \text{ with } \quad Q = \tfrac{1}{3}\sigma^2 I_3 .$$

Statement (3) could already be considered a statistical hypothesis, since it has the same structure as (2). Statement (3) asserts that the three distance observables are indeed normally distributed with unknown mean, but with known variancematrix Q. Statement (3) is however not yet the statistical hypothesis we are looking for. What we are looking for is a statistical hypothesis of which the probability density function depends on the theory or hypothesis postulated. For our case this means that we have to incorporate in some way the hypothesis that the three points lie on one straight line. We know mathematically that this assertion implies that:

$$(4) \qquad l_{12} + l_{23} - l_{13} = 0 .$$

However, we cannot make this relation hold for the random variables $\underline{l}_{12}$, $\underline{l}_{23}$ and $\underline{l}_{13}$. This is simply because of the fact that random variables cannot be equal to a constant. Thus, a statement like: $\underline{l}_{12} + \underline{l}_{23} - \underline{l}_{13} = 0$ is nonsensical. What we can do is assume that relation (4) holds for the expected values of the random variables $\underline{l}_{12}$, $\underline{l}_{23}$ and $\underline{l}_{13}$:

$$(5) \qquad E\{\underline{l}_{12}\} + E\{\underline{l}_{23}\} - E\{\underline{l}_{13}\} = 0 .$$

For the hypothesis considered this relation makes sense. It can namely be interpreted as stating that if the measurement experiment were to be repeated a great number of times, then on the average the measurements will satisfy (5). With (3) and (5) we can now state our statistical hypothesis as:

$$(6) \qquad H : \begin{pmatrix} \underline{l}_{13} \\ \underline{l}_{12} \\ \underline{l}_{23} \end{pmatrix} \sim N\left(\begin{pmatrix} E\{\underline{l}_{13}\} \\ E\{\underline{l}_{12}\} \\ E\{\underline{l}_{23}\} \end{pmatrix}, \tfrac{1}{3}\sigma^2 I_3 \right) \text{ with } \quad E\{\underline{l}_{12}\} + E\{\underline{l}_{23}\} - E\{\underline{l}_{13}\} = 0.$$

This hypothesis has the same structure of (2) with the three means playing the role of the parameter x.

In many hypothesis-testing problems two hypotheses are discussed: The first, the hypothesis being tested, is called the *null hypothesis* and is denoted by H_0. The second is called the *alternative hypothesis* and is denoted by H_A. The thinking is that if the null hypothesis H_0 is false, then the alternative hypothesis H_A is true, and vice versa. We often say that H_0 is tested against, or versus, H_A. In studying hypotheses it is also convenient to classify them into one of two types by means of the following definition: if a hypothesis completely specifies the distribution, that is, if it specifies its functional form as well as the values of its parameters, it

is called a *simple hypothesis* (enkelvoudige hypothese); otherwise it is called a *composite hypothesis* (samengestelde hypothese).

Example 1 (continued)

In our example (6) is the hypothesis to be tested. Thus, the null hypothesis reads in our case:

$$(7) \qquad H_0 : \begin{pmatrix} \underline{l}_{13} \\ \underline{l}_{12} \\ \underline{l}_{23} \end{pmatrix} \sim N\left(\begin{pmatrix} E\{\underline{l}_{13}\} \\ E\{\underline{l}_{12}\} \\ E\{\underline{l}_{23}\} \end{pmatrix}, \tfrac{1}{3}\sigma^2 I_3 \right) \text{ with } \quad E\{\underline{l}_{12}\} + E\{\underline{l}_{23}\} - E\{\underline{l}_{13}\} = 0.$$

Since we want to find out whether $E\{\underline{l}_{12}\} + E\{\underline{l}_{23}\} - E\{\underline{l}_{13}\} = 0$ or not, we could take as alternative the inequality $E\{\underline{l}_{12}\} + E\{\underline{l}_{23}\} - E\{\underline{l}_{13}\} \neq 0$. However, we know from the geometry of our problem that the left hand side of the inequality can never be negative. The alternative should therefore read: $E\{\underline{l}_{12}\} + E\{\underline{l}_{23}\} - E\{\underline{l}_{13}\} > 0$. Our alternative hypothesis takes therefore the form:

$$(8) \qquad H_A : \begin{pmatrix} \underline{l}_{13} \\ \underline{l}_{12} \\ \underline{l}_{23} \end{pmatrix} \sim N\left(\begin{pmatrix} E\{\underline{l}_{13}\} \\ E\{\underline{l}_{12}\} \\ E\{\underline{l}_{23}\} \end{pmatrix}, \tfrac{1}{3}\sigma^2 I_3 \right) \text{ with } \quad E\{\underline{l}_{12}\} + E\{\underline{l}_{23}\} - E\{\underline{l}_{13}\} > 0.$$

When comparing (7) and (8) we see that the type of the distribution of the observables and their variance matrix are not in question. They are assumed to be known and identical under both H_0 and H_A. Both of the above hypotheses, H_0 and H_A, are examples of composite hypotheses. The above null hypothesis H_0 would become a simple hypothesis if the individual expectations of the observables were assumed known.

1.2 Test of statistical hypotheses

After the statistical hypotheses H_0 and H_A have been formulated, one would like to test them in order to find out whether H_0 should be rejected or not.

A *test* of a statistical hypothesis:

$$H_0 : \underline{y} \sim p_{\underline{y}}(y|x) \text{ with } x \text{ fully or partially specified}$$

is a rule or procedure, in which a random sample of $\underline{y}$ is used for deciding whether to reject or not reject H_0. A test of a statistical hypothesis is completely specified by the so-called *critical region* (kritiek gebied), which will be denoted by K.

The *critical region K* of a test is the set of sample values of $\underline{y}$ for which H_0 is to be rejected. Thus, H_0 is rejected if $y \in K$.

It will be obvious that we would like to choose a critical region so as to obtain a test with desirable properties, that is, a test that is "best" in a certain sense. Criteria for comparing tests and the theory for obtaining "best" tests will be developed in the next and following sections. But let us first have a look at a simple testing problem for which, on more or less intuitive grounds, an acceptable critical region can be found.

Example 2

Let us assume that a geodesist measures a scalar variable, and that this measurement can be modelled as a random variable $\underline{y}$ with density function:

$$\underline{y} \sim \frac{1}{\sqrt{2\pi}} \exp[-\tfrac{1}{2}(y - E\{\underline{y}\})^2]. \tag{9}$$

Thus, it is assumed that $\underline{y}$ has a normal distribution with unit variance. Although this assumption constitutes a statistical hypothesis, it will not be tested here because the geodesist is quite certain of the validity of this assumption. The geodesist is however not certain about the value of the expectation of $\underline{y}$. His assumption is that the value of $E\{\underline{y}\}$ is x_0. This assumption is the statistical hypothesis to be tested. Denote this hypothesis by H_0. Then:

$$H_0 : \quad E\{\underline{y}\} = x_0. \tag{10}$$

Let H_A denote the alternative hypothesis that $E\{\underline{y}\} \neq x_0$. Then:

$$H_A : \quad E\{\underline{y}\} \neq x_0. \tag{11}$$

Thus the problem is one of testing the simple hypothesis H_0 against the composite hypothesis H_A. To test H_0, a single observation on the random variable $\underline{y}$ is made. In real-life problems one usually takes several observations, but to avoid complicating the discussion at this stage only one observation is taken here. On the basis of the value of $\underline{y}$ obtained, denoted by y, a decision will be made either to accept H_0 or reject it. The latter decision, of course, is equivalent to accepting H_A. The problem then is to determine what values of $\underline{y}$ should be selected for accepting H_0 and what values for rejecting H_0. If a choice has been made of the values of $\underline{y}$ that will correspond to rejection, then the remaining values of $\underline{y}$ will necessarily correspond to acceptance. As defined above, the rejection values of $\underline{y}$ constitute the critical region K of the test. Figure 1.2 shows the distribution of $\underline{y}$ under H_0 and under two possible alternatives H_{A_1} and H_{A_2}.

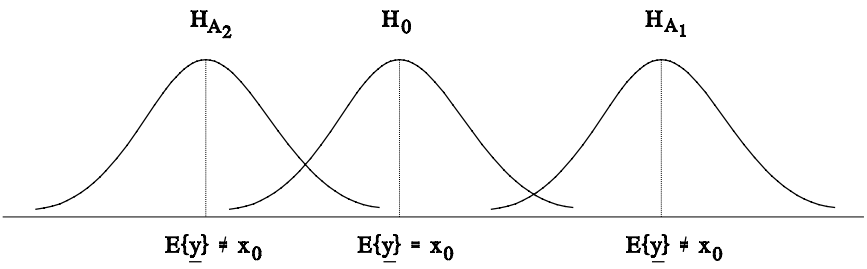


Figure 1.2: $H_0 : E\{\underline{y}\} = x_0$ versus $H_A : E\{\underline{y}\} \neq x_0$.

Looking at this figure, it seems reasonable to reject H_0 if the observation $\underline{y}$ is remote enough from $E\{\underline{y}\} = x_0$. If H_0 is true, the probability of a sample of $\underline{y}$ falling in a region remote from $E\{\underline{y}\} = x_0$ is namely small. And if H_A is true, this probability may be large. Thus the critical region K should contain those sample values of $\underline{y}$ that are remote enough from $E\{\underline{y}\} = x_0$. Also, since the alternative hypothesis can be located on either side of $E\{\underline{y}\} = x_0$, it seems obvious to have one portion of K located in the left tail of H_0 and one portion of K located in the right tail of H_0. Finally, one can argue that since the distribution is symmetric about its mean value, also the critical region K should be symmetric about $E\{\underline{y}\} = x_0$. This as a result gives the form of the critical region K as shown in Figure 1.3. Although this critical region has been found on more or less intuitive grounds, it can be shown that it possesses some desirable properties. We will return to this matter in a later section.

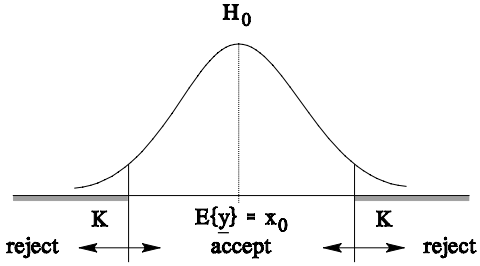


Figure 1.3: Critical region K for testing $H_0 : E\{\underline{y}\} = x_0$ versus $H_A : E\{\underline{y}\} \neq x_0$.

1.3 Two types of errors

We have seen that a test of a statistical hypothesis is completely specified once the critical region K of the test is given. The null hypothesis H_0 is rejected if the sample value or observation of $\underline{y}$ falls in the critical region, i.e. if $y \in K$. Otherwise the null hypothesis H_0 is accepted, i.e. if $y \notin K$. With this kind of thinking two types of errors can be made:

Type I error: Rejection of H_0 when in fact H_0 is true.

Type II error: Acceptance of H_0 when in fact H_0 is false.

Table 1.1 shows the decision table with the type I and II errors.

	H_0 true	H_0 false
Reject H_0 $y \in K$	Wrong Type I error	Correct
Accept H_0 $y \notin K$	Correct	Wrong Type II error

Table 1.1: Decision table with type I and type II error.

The *size* of a type I error is defined as the probability that a sample value of $\underline{y}$ falls in the critical region when in fact H_0 is true. This probability is denoted by α and is called the size of the test or the *level of significance* of the test (onbetrouwbaarheid van de test). Thus:

$$\alpha = P(\text{type I error}) = P(\text{rejection of } H_0 \text{ when } H_0 \text{ true})$$

or

(12) $$\boxed{\alpha = P(\underline{y} \in K | H_0) = \int_K p_{\underline{y}}(y | H_0) \mathrm{d}y} \; .$$

The size of the test, α, can be computed once the critical region K and the probability density function of $\underline{y}$ is known under H_0. The size of a type II error is defined as the probability that a sample value of $\underline{y}$ falls outside the critical region when in fact H_0 is false. This probability is denoted by β. Thus:

$$\beta = P(\text{type II error}) = P(\text{acceptance of } H_0 \text{ when } H_0 \text{ is false})$$

or

(13) $$\boxed{\beta = P(\underline{y} \notin K | H_A) = 1 - \int_K p_{\underline{y}}(y | H_A) \mathrm{d}y} \; .$$

The size of a type II error, β, can be computed once the critical region K and the probability density function of $\underline{y}$ is known under H_A.

Example 3

Assume that $\underline{y}$ is distributed as:

(14) $$\underline{y} \sim N(E\{\underline{y}\}, \sigma^2)$$

with known variance σ^2.

The following two simple hypotheses are considered:

(15) $$H_0 : \; E\{\underline{y}\} = x_0$$

and

(16) $$H_A : \; E\{\underline{y}\} = x_A > x_0.$$

The situation is sketched in Figure 1.4.

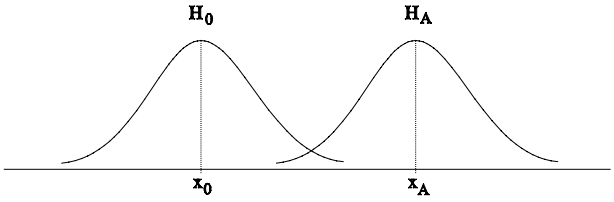


Figure 1.4: The two simple hypotheses: $H_0 : E\{\underline{y}\} = x_0$ and $H_A : E\{\underline{y}\} = x_A > x_0$.

Since the alternative hypothesis H_A is located on the right of the null hypothesis H_0, it seems intuitively appealing to choose the critical region K right-sided. Figure 1.5a and 1.5b show two possible right-sided critical regions K.

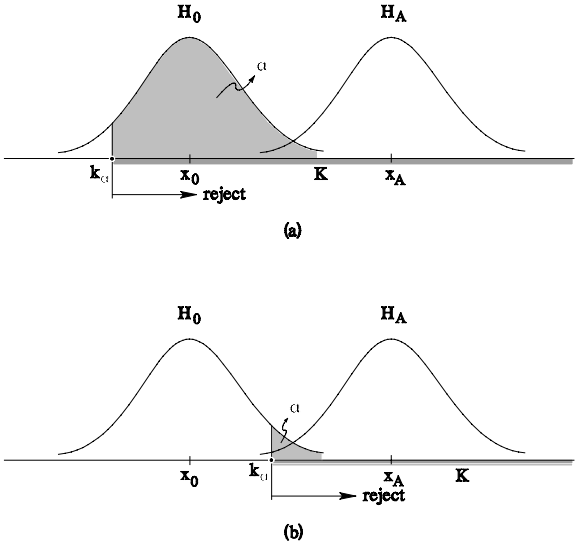


Figure 1.5: Critical region K and size of test, α.

They also show the size of the test, α, which corresponds to the area under the graph of the distribution of $\underline{y}$ under H_0 for the interval of the critical region K.

The size of the test, α, can be computed once the probability density function of $\underline{y}$ under H_0 is known and the form and location of the critical region K is known. In the present example the form of the critical region has been chosen right-sided. Its location is determined by the value of k_α, the so-called *critical value* (kritieke waarde) of the test. Thus, for the present example the size of the test can be computed as:

$$\alpha = \int_{k_\alpha}^{\infty} p_{\underline{y}}(y|x_0)\mathrm{d}y$$

or, since:

$$p_{\underline{y}}(y|x_0) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2}\frac{1}{\sigma^2}(y-x_0)^2]$$

as:

$$\alpha = \int_{k_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2}\frac{1}{\sigma^2}(y-x_0)^2]\mathrm{d}y. \tag{17}$$

When one is dealing with one-dimensional normally distributed random variables, one can usually compute the size of the test, α, from tables given for the standard normal distribution (see appendix B). In order to compute (17) with the help of such a table, we first have to apply a transformation of variables. Since $\underline{y}$ is normally distributed under H_0 with mean x_0 and variance σ^2, it follows that the random variable $\underline{z}$, defined as:

$$\underline{z} = \frac{\underline{y}-x_0}{\sigma} \tag{18}$$

is standard normally distributed under H_0. And since:

$$\alpha = P(\underline{y} > k_\alpha|H_0) = P(\underline{z} > \frac{k_\alpha - x_0}{\sigma}|H_0) \tag{19}$$

we can use the last expression of (19) for computing α. Application of the change of variables (18) to (17) gives:

$$\alpha = \int_{\frac{k_\alpha - x_0}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[-\tfrac{1}{2}z^2]\mathrm{d}z \ . \tag{20}$$

We can now make use of the table of the standard normal distribution. Table 1.2 shows some typical values of the α and k_α for the case that $x_0 = 1$ and $\sigma = 2$.

α	$\frac{k_\alpha - x_0}{\sigma}$	k_α
0.1	1.28	3.56
0.05	1.65	4.29
0.01	2.33	5.65
0.001	3.09	7.18

Table 1.2: Test size α, critical value k_α for $x_0 = 1$ and $\sigma = 2$.

As we have seen the location of the critical region K is determined by the value chosen for k_α, the critical value of the test. But what value should we choose for k_α? Here the geodesist should base his judgement on his experience. Usually one first makes a choice for the size of the test, α, and then by using (20) or Table 1.2 determines the corresponding critical value k_α. For instance, if one fixes α at $\alpha = 0.01$, the corresponding critical value k_α (for the present example with $x_0 = 1$ and $\sigma = 2$) reads $k_\alpha = 5.65$. The choice of α is based on the probability of a type I error one is willing to accept. For instance, if one chooses α as $\alpha = 0.01$, one is willing to accept that 1 out of a 100 experiments leads to rejection of H_0 when in fact H_0 is true.

Let us now consider the size of a type II error, β. Figure 1.6 shows for the present example the size of a type II error, β. It corresponds to the area under the graph of the distribution of $\underline{y}$ under H_A for the interval complementary to the critical region K.

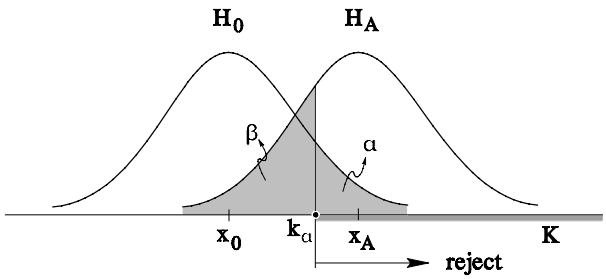


Figure 1.6: The sizes of type I and type II error, α and β, for testing H_0: $E\{\underline{y}\} = x_0$ versus H_A: $E\{\underline{y}\} = x_A > x_0$.

The size of a type II error, β, can be computed once the probability density function of $\underline{y}$ under H_A is known and the critical region K is known. Thus, for the present example the size of the type II error can be computed as:

$$\beta = \int_{-\infty}^{k_\alpha} p_{\underline{y}}(y|x_A)\mathrm{d}y$$

or since:

$$p_{\underline{y}}(y|x_A) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2}\frac{1}{\sigma^2}(y-x_A)^2]$$

as:

(21) $$\beta = \int_{-\infty}^{k_\alpha} \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2}\frac{1}{\sigma^2}(y-x_A)^2]\,dy.$$

Also this value can be computed with the help of the table of the standard normal distribution. But first some transformations are needed. It will be clear that the probability that a sample or observation of $\underline{y}$ falls in the critical region K when H_A is true, is identical to 1 minus the probability that the sample does not fall in the critical region when H_A is true. Thus:

(22) $$\beta = P(\underline{y} \notin K|H_A) = 1 - P(\underline{y} \in K|H_A).$$

Since for the present example:

$$P(\underline{y} \in K|H_A) = \int_{k_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2}\frac{1}{\sigma^2}(y-x_A)^2]\,dy$$

substitution into (22) gives:

(23) $$1 - \beta = \int_{k_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2}\frac{1}{\sigma^2}(y-x_A)^2]\,dy.$$

This formula has the same structure as (17). The value $1-\beta$ can therefore be computed in exactly the same manner as the size of the test, α, was computed. And from $1-\beta$ it is trivial to compute β, the size of the type II error.

Figure 1.7 gives the probability $1-\beta$ of rejecting H_0, when indeed H_A is true, as function of the unknown mean x_A under H_A. When this probability is requested to be at least $1-\beta = 0.80$, the unknown mean under H_A has to be at least $x_A = 7.34$. We return to the probability $\gamma = 1-\beta$, the power, in Section 4.5 on reliability. The size of the test was fixed to $\alpha = 0.01$.

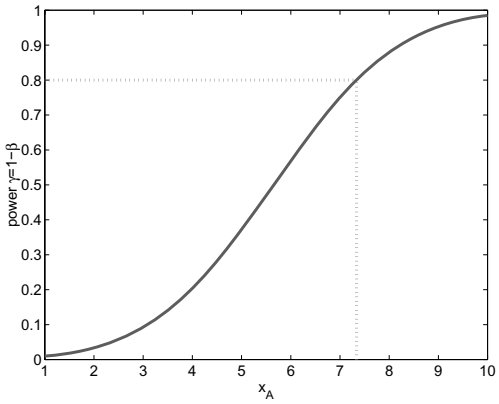


Fig. 1.7: Probability $\gamma = 1-\beta$ as function of x_A, for testing H_0: $E\{\underline{y}\} = x_0$ versus H_A: $E\{\underline{y}\} = x_A > x_0$, with $x_0 = 1$ and $\sigma = 2$.

1.4 A testing principle

We have seen that two types of errors are involved when testing a null hypothesis H_0 against an alternative hypothesis H_A: (1) The rejection of H_0 when in fact H_0 is true (type I error); (2) the acceptance of H_0 when in fact H_0 is false (type II error). One might reasonably use the sizes of the two types of errors, α and β, to set up criteria for defining a best test. If this is possible, it would automatically give us a method of choosing a critical region K. A good test should be a test for which α is small (ideally 0) and β is small (ideally 0). It would therefore be nice if we could define a test, i.e. define a critical region K, that simultaneously minimizes both α and β. Unfortunately this is not possible. As we decrease α, we tend to increase β, and vice versa. The *Neyman-Pearson* principle provides a workable solution to this situation. This principle says that we should fix the size of the type I error, α, and minimize the size of the type II error, β. Thus:

A testing principle (Neyman et al., 1933): Among all tests or critical regions possessing the same size type I error, α, choose one for which the size of the type II error, β, is as small as possible.

The justification for fixing the size of the type I error to be α, (usually small and often taken as 0.05 or 0.01) seems to arise from those testing situations where the two hypotheses, H_0 and H_A, are formulated in such a way that one type of error is more serious than the other. The hypotheses are stated so that the type I error is the more serious, and hence one wants to be certain that it is small. Testing principles other than the above given one can of course easily be suggested: for example, minimizing the sum of sizes of the two types of error, $\alpha + \beta$. However, the Neyman-Pearson principle has proved to be very useful in practice. In this book we will therefore base our method of finding tests on this principle. Now let us consider a testing problem from the point of view of the Neyman-Pearson principle.

Example 4

Assume that $\underline{y}$ has the following probability density function:

(24) $$p_{\underline{y}}(y|x) = xe^{-yx}\ ,\ x > 0\ ,\ y \geq 0\ .^{2}$$

The following two simple hypotheses are considered:

(25) $$H_0 : x = 2 \qquad \text{and} \qquad H_A : x = 1\ .$$

Figure 1.8 shows the density function of $\underline{y}$ under H_0 and H_A.

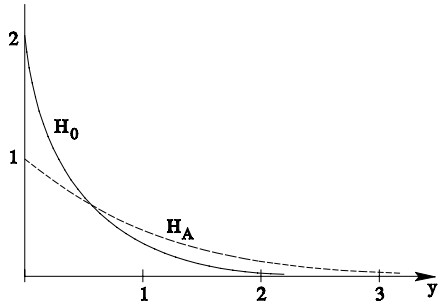


Figure 1.8: The function xe^{-yx}, $x>0$, $y\geq 0$ for $x = 2$ and $x = 1$.

Contrary to our Example 3, it is now not that obvious how to choose the form of the critical region K. Let us first consider the case of a right-sided critical region K. Thus:

(26) $$K = \{y \in \mathbb{R}_0^+ \mid y > k_\alpha\}\ .$$

In order to compute α and β we need to evaluate an integral of the type:

(27) $$\int_a^b xe^{-yx}dy = e^{-ax} - e^{-bx}.$$

For the right-sided critical region (26) this gives for the size of the type I error:

(28) $$\alpha = \int_{k_\alpha}^{\infty} 2e^{-y^2}dy = e^{-2k_\alpha}\ .$$

The corresponding size of the type II error is:

---

[2] Prove yourself that this function is indeed a probability density function.

(29) $$\beta = 1 - \int_{k_\alpha}^{\infty} 1e^{-y1}dy = 1 - e^{-k_\alpha}.$$

Now let us consider a left-sided critical region K^* as alternative. Thus:

(30) $$K^* = \{y \in \mathbb{R}_0^+ \mid 0 \leq y < k_\alpha^*\}.$$

For this critical region the size of the type I error becomes:

(31) $$\alpha^* = \int_{0}^{k_\alpha^*} 2e^{-y2}dy = 1 - e^{-2k_\alpha^*}.$$

And the corresponding size of the type II error is given by:

(32) $$\beta^* = 1 - \int_{0}^{k_\alpha^*} 1e^{-y1}dy = e^{-k_\alpha^*}.$$

Let us now compare the two tests, that is, the one with the right-sided critical region K with the one with the left-sided critical region K^*. We will base this comparison on the Neyman-Pearson principle. According to this principle, both tests have the same size of type I error. Thus:

(33) $$\alpha = \alpha^*.$$

With (28) and (31) this gives $e^{-2k_\alpha^*} = 1 - e^{-2k_\alpha}$ or:

(34) $$(e^{-k_\alpha^*})^2 = (1 - e^{-k_\alpha})(1 + e^{-k_\alpha}).$$

Using (29) and (32) this equation can be expressed in terms of β^* and β as:

$$(\beta^*)^2 = \beta(2 - \beta).$$

Hence:

(35) $$\beta^* = (2\beta - \beta^2)^{\frac{1}{2}}.$$

Figure 1.9 shows the graph of this function. It clearly shows that:

(36) $$\beta < \beta^*.$$

The conclusion reads therefore that of the two tests the one having the right-sided critical region K is the best in the sense of the Neyman-Pearson principle.

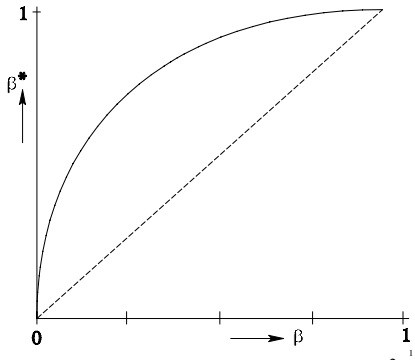


Figure 1.9: The function $\beta^* = (2\beta - \beta^2)^{\frac{1}{2}}$.

1.5 General steps in testing hypotheses

Thus far we have discussed the basic concepts underlying most of the hypothesis-testing problems. The same concept and guidelines will provide the basis for solving more complicated hypothesis-testing problems as treated in the next chapters. Here we summarize the main steps on testing hypotheses about a general probability model.

(a) From the nature of the experimental data and the consideration of the assertions that are to be examined, identify the appropriate null hypothesis and alternative hypothesis:

$$H_0\colon \underline{y} \sim p_{\underline{y}}(y|x_0) \quad \text{versus} \quad H_A\colon \underline{y} \sim p_{\underline{y}}(y|x_A)\,.$$

(b) Choose the form of the critical region K that is likely to give the best test. Use the Neyman-Pearson principle to make this choice.

(c) Specify the size of the type I error, α, that one wishes to assign to the testing process. Use tables to determine the location of the critical region K from:

$$\alpha = P(y \in K|H_0) = \int_K p_{\underline{y}}(y|x_0)\,\mathrm{d}y.$$

(d) Compute the size of the type II error:

$$\beta = P(y \notin K|H_A) = 1 - \int_K p_{\underline{y}}(y|x_A)\mathrm{d}y$$

to ensure that there exists a reasonable protection against type II errors.

(e) After the test has been explicitly formulated, determine whether the sample or observation y of $\underline{y}$ falls in the critical region K or not. Reject H_0 if $y \in K$, and accept H_0 if $y \notin K$. Never claim however that the hypotheses have been proved false or true by the testing.