Statistics in Language Research: Analysis of Variance

Statistics in Language Research: Analysis of Variance

by

Toni Rietveld and Roeland van Hout

Mouton de Gruyter
Berlin · New York

Mouton de Gruyter (formerly Mouton, The Hague)
is a Division of Walter de Gruyter GmbH & Co. KG, Berlin.

Library of Congress Cataloging-in-Publication Data

Rietveld, Toni, 1949–
Statistics in language research : analysis of variance / by Toni Rietveld and Roeland van Hout.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-3-11-018580-5 (cloth : alk. paper)
ISBN-10: 3-11-018580-6 (cloth : alk. paper)
ISBN-13: 978-3-11-018581-2 (pbk. : alk. paper)
ISBN-10: 3-11-018581-4 (pbk. : alk. paper)
1. Linguistics – Statistical methods. 2. Analysis of variance. I. Hout, Roeland van. II. Title.
P138.5.R543 2005
410'.2'1–dc22
2005019159

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

ISBN-13: 978-3-11-018580-5 hc.
ISBN-10: 3-11-018580-6 hc.
ISBN-13: 978-3-11-018581-2 pb.
ISBN-10: 3-11-018581-4 pb.

Cover design: Klein Kon Jung, Berlin
Printed in Germany.

Contents

Acknowledgements

We are indebted to our students for their comments on earlier drafts of the book. We thank dr. Han Oud of Radboud University of Nijmegen for the hours he was so kind to spend considering the pros and cons of maximum likelihood estimation in mixed models. We thank Lieselotte Toelle MA for correcting our English, Dr. Bert Cranen and drs. Joop Kerkhoff for the artwork. Needless to say, we are the ones responsible for any remaining errors.

Chapter 1
Language research and statistics

1.1. Statistics and analysis of variance in language research

Language research is based on data. Sometimes the data are quite subjective, like the appreciation of voices or accents, or introspective, like the intuitions of linguists on the well-formedness of utterances. In many situations an introspective approach is warranted and one does not need quantitative or statistical methods to corroborate the scientific argumentation. However there are many situations in which the language researcher needs to collect data through a survey, in a field study, in an experiment, or in a language corpus. Many linguistic subdisciplines use methods which are similar to the ways researchers in the social sciences obtain and analyze data.

Researchers in these subdisciplines want to generalize the outcomes to the population(s) from which they have taken one or more samples. The wish to make generalizations entails the use of inductive statistics, the branch of statistics which enables us to infer population characteristics. An example: "Speakers in dialect A exhibit phonetic process X more often than speakers of dialect B", on the basis of outcomes obtained in relatively small samples, for instance 110 speakers of dialect A and 80 speakers of dialect B. As is the case in psychology and sociology, sampling a population in such a way that the sample is representative of the population under consideration in gender, age, socio-economic status etc. is an important issue. Random sampling is a skill in itself, especially in survey research.

This book does not deal with data collection but with data analysis, and, more particularly, it deals with ANalysis Of VAriance, abbreviated ANOVA. This technique is the main instrument for social scientists and their linguistic colleagues to analyze the outcomes of research designs with more than two treatments or groups. Moreover, analysis of variance enables the researcher to assess the effects of more than one independent variable at the same time. When data are obtained from participants of two different groups at four points in time, we may want to know whether the outcomes of the two groups differ; we may also want to know whether their outcomes change at different rates.

Often students and researchers need to apply analysis of variance to their data although they may feel insecure about the basic principles of statistical testing. That is why in this book the treatment of analysis of variance is preceded by two chapters which explain the use of *t* tests, type I and type II errors and power analysis. These are fundamental concepts which constitute the basis of statistical testing. Special attention is paid to an important concept in experimental design: determining the size of the sample needed to detect specific hypothesized effects in (the) population(s).

This book gives a comprehensive treatment of ANOVA. Technical and more complete treatments can be found in Kirk (1995) and Winer, Brown, and Michels (1991), but these two textbooks are fairly hard to comprehend for those researchers in the field of language and speech behaviour who have only attended an introductory course in statistics, perhaps based on specific textbooks for language research like Butler (1985) and Woods, Fletcher, and Hughes (1986). Many language researchers seem to use ANOVA simply by following the HELP files of statistical packages like SPSS or books which are often obsolete as far as current developments are concerned. We want to explain to language researchers what they are doing when they use ANOVA and which options are available. In addition, we would like to inform them about developments in post hoc comparisons, power analysis, standards in reporting statistics, ways of dealing with missing data, and the pros and cons of the F1×F2 approach currently used in psycholinguistic research. Another problem with books on ANOVA is that they only deal with examples from the social sciences and are often either too simple or too complicated for the user of statistics.

In Chapters 4 to 9 we cover the most widely used experimental designs step by step, showing the researcher how the analyses have to be executed. Chapters 9 and 10 are slightly more complicated than the others we admit, but we wanted to highlight more recent developments in the analysis of multigroup data: different estimation procedures, multilevel analysis, bootstrap and randomization tests. A specific section in Chapter 9 deals with missing data, a common phenomenon in psycho- and sociolinguistic research, which is more relevant than most researchers assume: The standard approach in reporting analysis of variance is not to deal with missing data.

The structure of a chapter is determined by the statistical concepts and analyses handled, but the following sections return in each chapter:

- Preview: Information is provided about what one is going to read.

- Technical sections with examples taken from linguistic research.
- Terms and concepts: Summary of the concepts presented in the chapter.
- Statistical reporting: Examples are presented, mostly based on suggestions made by the American Psychological Association (APA).
- Exercises.

1.2. Variables

There are a number of data collection methods in linguistics, such as survey research, experimental research and corpus research. In all these methods the *variable* concept plays a key role. A variable is a property or characteristic of a person, a condition, an object, or any other research element. These are defined by the research questions and the way in which they are made operational in the research procedures. In the examples we use in this book, the elements are often subjects, informants, language users or dialect speakers. We usually call them *participants*, to meet the more recent standard terminology in reporting research.

Often we want to know whether variable A affects scores obtained on variable B. We call variable A an independent variable, and variable B a dependent variable, a distinction which is particularly familiar in (quasi-)experimental research. The independent variable is not always under the control of the researcher. If, for instance, speakers of dialect A_1 live in rural areas, whereas speakers of dialect A_2 are mainly found in urban areas, the researcher cannot change this fact. Being a speaker of dialect A_2 is connected to being an urbanite. In genuinely experimental research the investigator has the independent variable(s) under full control. He/she can deliberately introduce four levels of noise (50, 60, 70 and 80 dB) in which participants have to identify specific speech segments, or specify the number of syllables (1, 2, 3) of carrier words which participants have to listen to. In the examples discussed there is at least one dependent and one independent variable.

Variables have values which they get as a result of a measurement procedure. Measurement is the assignment of numerals to objects or events according to rules, cf. Stevens (1946). The scale of measurement determines the amount of information contained in the data (Anderson, Sweeney, and Williams 1991: 19). There are four scales of measurement:

1. *Nominal scale* (also called 'categorical scale' or 'qualitative' scale). Language background is a good example. One cannot say that English

is a better or more sophisticated language background than Dutch. It is simply a different language, the way an adjective is different from a noun. A variable is nominal if it is used to label elements or observations in order to categorize or classify them. All transformations are allowed which leave classification unaffected: A, B and C can be transformed into $\gamma, \delta, \varepsilon$, into the numbers 1, 2 and 3, or the other way round, into 3, 2 and 1, and even into –5, 11236 and 432 etc. as long as the different labels represent different classes. Objects measured on nominal scales can only be distinguished from each other. They are equal or not equal, i.e. they belong to the same class or category (= equal) or not (= not equal); object K = object M or object K $\neq$ object M.

2. *Ordinal scale*. In addition to distinguishing objects (on the nominal scale), we can also rank objects which are measured on an ordinal scale (either object K > object M or object M > object K). Language competence is an example of an ordinal scale. Advanced learners of English know more about English than intermediate learners and the latter know more than beginners. Students with these labels differ in their competence of English, and can be ordered on the variable 'competence of English'. More information is available than information measured on a nominal scale. We do not know, however, whether the difference between advanced and intermediate students is equal to the difference between an intermediate and a beginner. Theoretically all monotone transformations are permitted, that is all transformation which leave the order unaffected. Strictly speaking a monotone transformation of scale values like 1, 2, 3, 4 (representing, for instance, degrees of harshness) into 1, 3, 37, 58 is permitted as the order is not affected. In practice, such a transformation is disturbing, for the numbers suggest that we do have more precise information. On an ordinal scale the intervals do not contain any information.
3. *Interval scale*. If the difference between measurement levels is known, the data are measured on an interval scale. For instance, the physical difference between the two pitch values 100 and 120 Hz is equal to the difference between 200 and 220 Hz (viz. 20 Hz). It is another question whether the same holds for perceived pitch differences. Linear transformations of the $F(X) = aX + b$ type are permitted as they leave the differences between scale values unaffected.
4. *Ratio scale*. An absolute requirement for a ratio scale is that a true zero

value is defined on the scale. A car which does not cost anything ('is for free') gets the value '0' on this scale. Ratios make sense: 400 kilos are twice as heavy as 200 kilos, and 200 kilos are twice as heavy as 100 kilos. This is not the case for data measured on an interval scale. We cannot say that the coronation of Charlemagne (800 AD in Aix-la-Chapelle) was twice as late as the 'Great Migration' (400 AD). We have a zero value in our calendar time, but it is an arbitrary zero value. The transformation $F(X) = aX$ is permitted here. Although it changes the absolute value of the difference, it does not change the relative value. In contrast to the interval scale adding b to the transformation is not permitted, as it would affect the absolute zero value.

The scale level on which variables are measured has consequences for the statistical technique that can be applied. It is customary to say that t tests and analysis of variance require strict interval data for the dependent variable. We do not think this is the case. The robustness of these techniques is amazing, but we postpone the discussion to Chapter 7.

A concrete example may illustrate the distinctions we make here. Let us assume that we have a nominal variable 'region'. If a researcher wants to know whether the nominal variable 'region' affects the duration of a vowel associated with a sentence accent (measured in milliseconds), we have the independent variable 'region' and the dependent variable 'duration'. Each measurement carried out on speakers from the region in question, constitutes a 'case'. In the following matrix we display nine *cases* rowwise, and two variables columnwise. The first column contains the values of the dependent variable, the second one the values of the independent variable.

Table 1.1. A data matrix with nine cases and two variables

case	region	duration
1	1	120.000
2	1	124.000
3	1	130.000
4	2	130.000
5	2	140.000
6	2	145.000
7	3	120.000
8	3	110.000
9	3	100.000

Most statistical packages have a standard data format: Cases are represented row-wise, variables column-wise. Cases and variables are defined by the research design, and a great variety of cases and variables is possible. In the examples we present in our book we restrict ourselves to straightforward research designs, where participants play the roles of cases. The data in Table 1.1 show this data format. The independent variable 'region' is nominal and distinguishes three regions. The dependent variable 'duration' is a ratio variable with a true zero point. There are three cases per region.

Finally, we would like to say a few words about corpus research, which has become very popular in the last two decades. The fact that large databases have become available is an important factor in the increasing use of corpora as sources of information, next to intelligent research tools. Very often the outcome variables in this kind of research consist of 'counts' and relative frequencies, often converted to percentages, like "X% of the sentences are shorter than 10 words in text type A, whereas it is Y% in text type B". Most of the time analysis of variance is not the appropriate tool for the analysis of this kind of data. We refer to Baayen (2001) for word frequency statistics, and to Oakes (1998) for statistics in corpus linguistics. For pitfalls in corpus research see Rietveld, Van Hout, and Ernestus (2004).

1.3. Designs

In this book we are going to deal with the statistical analysis of data obtained in a number of so-called research *designs*, which define the research variables and their status. Two relevant questions for the characterization of a design with one dependent and one independent variable are the following:

- *How many levels or values has the independent variable?* This question relates to the number of classes or categories distinguished in the nominal independent variable. If two levels are distinguished a *t* test can be applied. With more than two classes an ANOVA technique is required.
- *Is the independent variable a between-subject factor or a within-subject factor?* In studying language acquisition we can track the development over time by using different age groups, each group consisting of different children. Suppose we test them with a vocabulary test. By comparing the age groups we investigate vocabulary development; the differences we observe are differences between children in different age groups. So development is a between-participant factor, or in the classical terminol-

ogy we use, a between-subject factor. It concerns differences between groups. The other route is to track the development of a group of children over time. The same group is tested repeatedly over a longer period of time. Such a design is called longitudinal. The differences we observe are differences within the same children. Development now is a within-participant factor, or in the classical terminology, a within-subject factor.

The distinctions related to these two questions were taken into account in the schedule in Table 1.2, where crossing the two questions delivers four basic designs.

Table 1.2. Four basic designs based on the properties of the independent variable

independent variable	two levels	more than two levels
between-subject factor	*design 1* *t* test independent samples	*design 3* one-way ANOVA
within-subject factor	*design 2* *t* test correlated samples	*design 4* repeated measures ANOVA

On the next pages we present schedules of the four basic designs. They are represented in the SPSS data matrix format. Research design 1 comprises two groups (equal numbers of cases are not required). In experimental conditions, the participants are assigned to the two groups in an a-select way. Table 1.3 distinguishes one independent variable, i.e. 'group' with two levels. The dependent variable is called 'dep'.

We did not assign each case (participant) a unique case label. In practice, we should do so, to be sure that the right data are assigned to the right participant. In SPSS a case automatically gets a case label. We assigned the levels 1 and 2 to the 'group' variable, but, as said above, the real values are not relevant as long as they are different. The values –103 and 3425 yield the same outcomes in the statistical analysis.

Design 2 is represented in Table 1.4. In this design we have a within-subject factor or variable, which means that the same participants are measured twice in two different situations or conditions. In fact we have one group of participants.

The data format in Table 1.4 looks completely different from the format in Table 1.3. In both tables, each case (= participant) is represented by one

Table 1.3. Design 1: SPSS data matrix, the independent between-subject factor or variable 'group' with two levels

group	dep
1	15
1	16
1	28
1	32
1	12
2	20
2	32
2	16
2	11
2	14

Table 1.4. Design 2: SPSS data matrix, the independent within-subject factor is represented by two variables

cond1	cond2
15	19
16	33
28	28
32	44
12	18

row. Participants in design 2 were measured twice, which is represented by two variables, the first one containing the scores obtained in the first condition 'cond1', and the second one the scores obtained in the second condition, 'cond2'. The constraint of having only two levels for the independent variable is expressed by just having two variables to represent the 'condition' factor.

Design 3 is an expansion of design 1 for the number of levels of the 'group' variable, which now distinguishes three values or levels. This can be seen in Table 1.5, which does not contain values for the dependent variable.

Design 4 is an expansion of design 2. Instead of two variables we now have three, which means that the participants were measured three times. The three measurements constitute the within-subject factor. Such designs are often labelled as repeated measures designs. The SPSS data matrix format is given in Table 1.6.

Table 1.5. Design 3: SPSS data matrix, the independent between-subject variable 'group' has three levels

group	dep
1	
1	
1	
1	
1	
2	
2	
2	
2	
3	
3	
3	
3	

Table 1.6. Design 4: SPSS data matrix, the independent within-subject factor 'time' is represented by three variables, 't1', 't2', and 't3'

t1	t2	t3

The subsequent step to expand the analytical possibilities is to increase the number of factors. Starting from the single between-subject and within-subject factor designs, we can add three designs:

- design 5: a multifactorial design for between-subject factors
- design 6: a multifactorial design for within-subject factors
- design 7: a multifactorial mix of between- and within-subject factors.

Design 5 is a so-called *completely randomized factorial design*. The two independent variables could be gender (2 levels) and experimental condition

(3 levels). Unequal numbers of participants often lead to problems, as we discuss later (Chapter 7). An example with the variables gender and experimental condition is given in Table 1.7.

Table 1.7. Design 5, SPSS data matrix, completely randomized factorial design with two independent variables 'gender' and 'cond' (= condition)

cond	gender	dep
1	1	
1	1	
1	2	
1	2	
2	1	
2	1	
2	2	
2	2	
3	1	
3	1	
3	2	
3	2	

The variable 'dep' in Table 1.7 is the dependent variable. One may add other independent variables, but the number of combinations will then quickly multiply. Adding a third facor with three levels results in 18 combinations. A large number of cells impedes the interpretation of the statistical outcomes.

Design 6 is a multifactorial within-subject design. All participants are measured in all factor combinations, as is shown in Table 1.8. There are two within-subject factors, 'time' (three levels) and 'condition' (two levels) which means 6 combinations. Consequently, there are six variables in Table 1.8. The subscript of factor 'c' changes more slowly than that of factor 't', moving from left to right.

Finally, we have design 7, the so-called *split-plot design*, with at least one within-subject factor and one between-subject factor. The latter refers to the independent variable which distinguishes participants, like gender or socio-economic status. In our example in Table 1.9 there is one between-subject factor 'therapy' with two levels (two kinds of therapy) and one within-subject factor, again 'time' with 3 levels: 't1', 't2', 't3'.

Table 1.8. Design 6: SPSS data matrix, with two independent within-subject factors 'time' (three levels, indicated by 't1', 't2', 't3') and 'condition' (two levels, indicated by 'c1', 'c2'), represented by six variables

c1t1	c1t2	c1t3	c2t1	c2t2	c2t3

Table 1.9. Design 7: SPSS data matrix, with a between-subject factor 'therapy' and an independent within-subject factor 'time' represented by three variables, 't1', 't2', and 't3'

therapy	t1	t2	t3
1			
1			
1			
1			
1			
1			
2			
2			
2			
2			
2			
2			

The three time variables in Table 1.9 could represent a pre-test, a post-test and a post-test after a longer period of time, to check whether the effect of a therapy remains or fades.

Other aspects can complicate the designs we have discussed. In the multifactorial designs presented above, all combinations of levels of all factors occur. The resulting design is called a crossed design. However, there are hierarchical designs as well, these are designs in which not all levels of one factor co-occur within all levels of another factor. An example is the factor denomination of hospitals (Christian and General) and the Hospitals themselves (3 Hospitals of each denomination). Each hospital is only listed in one denomination. We will discuss nesting in Chapter 5.

A final question here is whether the values of the independent variable represent a sample. Do the independent variables (factors) involve random samples of a large number of possible samples? If the answer is yes, we have to deal with a so-called random factor, if not, the factor is called a fixed factor. This distinction affects the way in which the data have to be analyzed, as we discuss in Chapter 5.

1.4. Statistical packages

Hardly anyone carries out statistical analyses by hand or by pocket calculator. It became the task of statistical computer packages, supplemented by dedicated software for less common statistical procedures. We mention SAS: Statistical Analysis System, SPSS : Statistical Package for the Social Sciences, MINITAB, S+, R. We do not want to express a preference for one of these, but we chose SPSS (version 12.0) to provide examples of analysis in this book, as we think it is the most widely known package in language research.

We will demonstrate how statistical analyses can be carried out in SPSS with the Point-And-Click (PAC) window system. We show how to use SPSS syntax in a number of cases, because the syntax approach offers extra options and possibilities in carrying out a statistical analysis.

Chapter 2
Basic statistical procedures: one sample

2.1. Preview

In this chapter we review the basic principles of statistical testing. These principles are reviewed on the basis of the *one-sample* design with the dependent variable measured at the interval or ratio level. A one-sample design is a design in which a statistic of a sample drawn from a population, for instance a mean value, is compared with a value which is hypothesized to hold for the population. Obviously, a sample mean will hardly ever have the same value as the hypothesized value ($\bar{X}$ is not μ): If we hypothesize that the mean lifetime of lightbulbs is 2000 hrs, and we draw a random sample of 100 bulbs, the mean lifetime of that sample cannot be expected to be 2000 hrs, but a figure just above or below this, even if the mean liftetime in the population (μ) is 2000 hrs. The question is what degree of deviation of the observed value of the mean from the hypothesized value can be accepted without us having doubts on the hypothesized mean value. The concept of *sampling variability* is extremely important in this context (Section 2.2). In Section 2.3 we review the procedure of hypothesis testing and in Section 2.4 the well-known *t* distribution. Important concepts are *statistical power* and *effect size* (Section 2.5). In Section 2.6 we discuss the calculation of the *sample size needed* to detect a pre-defined effect size. These concepts tend to be (wrongly) neglected in linguistic research, but they are very important in medical research. An example is the test of a hypothesis that phoneticians only guess when they are asked to assess the age of a speaker, against the hypothesis that their ratings are correct in, for example, 75% of the cases.

2.2. Sampling variability

With continuous (interval or ratio) data, our interest is often focused on the mean(s) of our sample(s). However, the mean only provides a summary of the data we have collected. If we were to collect another sample of the same size, it is extremely unlikely that precisely the same mean value will be obtained. This phenomenon is called *sampling variability*.

Let us look at a simple situation, the reaction time for recognizing a Dutch word. The time taken to recognize the word 'rente' (= interest) was measured. The mean time taken to recognize the word was 523 ms. This is the mean of our sample of Dutch listeners, but we would like to be able to generalize this to all native listeners of Dutch, not just the ones we happen to have sampled. Our best guess of the mean time taken to recognise 'rente' in the population is the mean of our sample, 523 ms. This is all the information available, apart from the standard deviation, and we have no reason to expect it to be biased in any way.

Of course, we would like to know whether the value of 523 ms – denoted by $\bar{X}$ – is a good estimate of the population mean μ. In order to get an intuitive feeling of the quality of our estimate, we have to know more about the sampling variability. To that end we introduce the concepts of *sampling distribution* and *standard error*, *SE*. If we take a large number of samples of, for instance 30 observations, and calculate the mean of each sample, the sample means will be normally distributed around μ. Even if the distribution is skewed or uniform, the distribution of sample means is normal. This important fact is called the *Central Limit Theorem*, a theorem which plays a crucial role in statistics. The distribution of sample means also has a standard deviation, called the *standard error*, which can be estimated from the data itself. If we can assume that the size of the sample is smaller than or equal to 5% of the population size, the formula for the standard error of the sample means, $\bar{X}$ is (cf. Anderson, Sweeny, and Williams 1991: 231):

$$SE_{\bar{X}} = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (1)$$

where σ is the standard deviation of the population of our data, and n is the sample size. In most cases σ is unknown, and has to be estimated by the standard deviation of the sample: s. In our example, $s = 76.44$ and $n = 28$, thus the estimated standard error of the mean is $76.44/\sqrt{28} = 14.45$.

The formula for the standard error summarizes two factors which affect the stability of the estimates of the mean of a population μ: the variation in the population σ (estimated by s) and the size of the sample n. The larger the variation in the population is, the more we can expect sample means $\bar{X}$ to vary – that is why s is in the numerator of the formula; the larger n is, the smaller the fluctuations in sample mean values are.

The Central Limit Theorem tells us that a statistic like $\bar{X}$ is normally distributed with the mean μ and the standard deviation $\sigma/\sqrt{n}$ if the number

of observations is large enough (in theory the theorem holds when n approaches infinity, in practice $n > 30$ suffices). Supposing we have a normally distributed population, with $\mu = 250$ and $\sigma = 50$, we can randomly draw (by compute) 500 samples of size $n = 3$, of $n = 30$, and of $n = 100$. The distributions of the 500 samples for the three samples sizes are given in Figure 2.1.

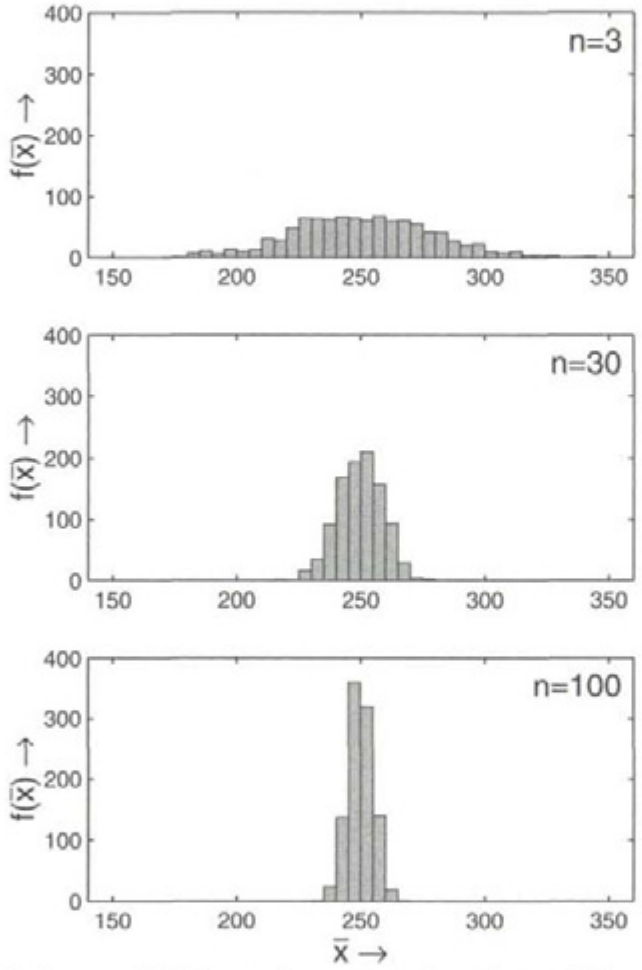

Figure 2.1. Distributions of 500 sample means for three different sample sizes: $n = 3, n = 30, n = 100$, with samples drawn from a normally distributed population with $\mu = 250$ and $\sigma = 50$

The *SE* of the sampling distribution with $n = 3$, is $50/\sqrt{3} = 28.90$; the *SE* for samples with $n = 30$ is $50/\sqrt{30} = 9.13$. For sample size $n = 100$, we get $50/\sqrt{100} = 5$. The effect of n becomes quite clear: The larger n is, the smaller the *SE* of the sampling distribution becomes. This is shown by the three sample distributions in Figure 2.1. At the same time, we see that the

samples have a normal distribution, particularly with larger sample sizes.

Figure 2.1 shows that the standard error of the sample mean decreases if the sample size increases. Let us take a closer look at this relationship. Given the sample mean, the standard error and the sample size are connected by the square root function. Some examples are given in Table 2.1, in which a standard deviation of 100 is assumed in the population.

Table 2.1. Relationship between sample sizes and the standard error, given a standard deviation of 100

sample size	square root sample size	standard error
1	1	100
4	2	50
16	4	25
64	8	12.50
256	16	6.25
1024	32	3.125

Starting at the value 1 the sample sizes are quadrupled. Going from sample size 1 to 4, the standard error decreases with $1/\sqrt{4} = 1/2$, which entails a drop of the standard error from 100 to 50. Quadrupling the sample of 4 to 16 again returns a drop by 1/2. The conclusion is obvious. Within the range of smaller sample sizes, increasing the sample size has a large impact on the standard error. For larger samples, there is still an impact, but it is less pronounced.

2.3. Hypothesis testing: one sample

When an $\bar{X}$ – a sample mean – is available, we might like to know whether its value is in accordance with an expectation. Especially in industrial settings people might have expectations about the value of μ in the population, and confront these with the actual sample means (for instance of the mean lifetime of light bulbs). If the observed sample mean substantially deviates from the expectation, then one has to revise the expectation (or sue the manufacturer...). In this case the standard normal distribution – also called the z distribution – offers help. As you may remember from basic statistics, a transformation of a raw score into a z score yields a variable with 'standard properties'. The resulting variable has a mean of 0, and a standard deviation of 1. This is also written as $N(0, 1)$. The equation is:

$$z = \frac{X - \mu}{\sigma} \qquad (2)$$

We can also standardize statistics like $\bar{X}$ or the difference between two means: $\bar{X}_1 - \bar{X}_2$. The only thing we need is the appropriate term for the denominator: SE, which takes different forms for different statistics. For the statistic $\bar{X}$ the z value is:

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \qquad (3)$$

You may ask what standardization is for? Well it enables us to use tables to find out whether the deviation of our sample mean from the hypothesized population mean is a probable one. Let us assume that we have a hypothesis about the mean μ of the reaction times mentioned above, for example 490 ms. The sample mean is 523 ms, the standard deviation 76.44, and the sample size 28. Filling in these values we get:

$$z = \frac{523 - 490}{76.44/\sqrt{28}} = 2.28 \qquad (4)$$

We just have to use the well-known tables of the standard normal distribution to find out that the probability of obtaining a z equal to or larger than 2.28 is 0.0113: $p(z \geq 2.28) = 0.0113$. Below we reproduce two sections of this kind of table which can be organized in two different ways:

1. The upper part (= a) in Table 2.2 shows the probability of obtaining z values between 0 and the observed z value (= Z) ($p(0 \leq z \leq Z)$) (see section a in Table 2.2); in order to obtain the probability of a z value equal to or larger than the actual Z value, we subtract the probability found from 0.50, as the probability surface left and right from $z = 0$ equals 0.50; cf. Figure 2.2, the distribution in the right panel.
2. The lower part (= b) in Table 2.2 shows the probability of obtaining z values equal to or larger than the actual Z value; cf. Figure 2.2, the distribution in the left panel.

Referring to our example, with $z = 2.28$, we obtain in part (a) of Table 2.2 at the intersection of $z = 2.2$ and 0.08: 0.4887; 0.5000 – 0.4887 = 0.013. In part (b) we get this result straight away: 0.013.

Table 2.2. Sections of tables of the standardized normal distribution

a	$p(0 \leq z \leq Z)$									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
b	$p(z \geq Z)$									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110

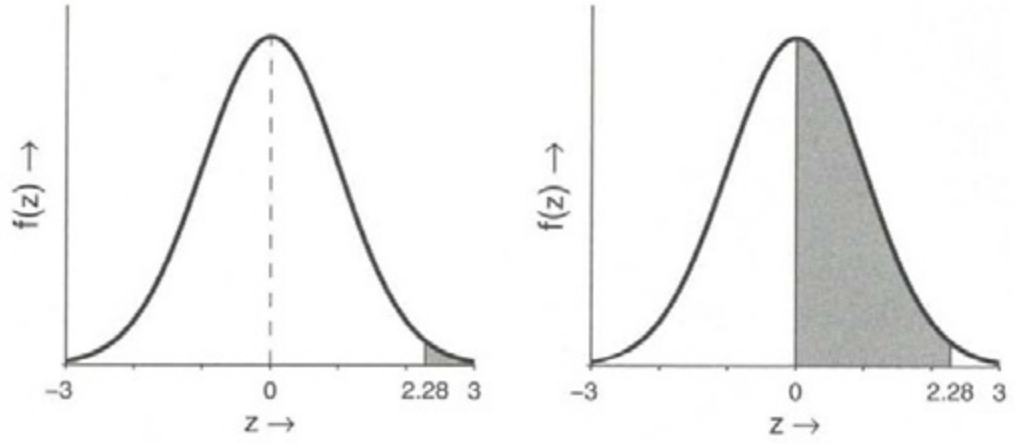

Figure 2.2. The z distribution, showing the probability areas, p, for $z \geq 2.28$ (left panel) and $0 \leq z \leq 2.28$ (right panel)

In fact we tested two hypotheses:

H_0 The mean recognition time of 'rente' does not differ from 490 ms; the difference found is simply due to sampling variability.

H_1 The mean recognition time of 'rente' differs from 490 ms.