# Experiments : design, parametric and nonparametric analysis, and selection

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

Memorandum COSOR 92-15

# Experiments: Design, Parametric and Nonparametric Analysis, and Selection

P. van der Laan

Eindhoven, June 1992
The Netherlands

Eindhoven University of Technology
Department of Mathematics and Computing Science
Probability theory, statistics, operations research and systems theory
P.O. Box 513
5600 MB Eindhoven - The Netherlands

Secretariate: Dommelbuilding 0.03
Telephone: 040-47 3130

ISSN 0926 4493

# Experiments: Design, Parametric and Nonparametric Analysis, and Selection

*Paul van der Laan*
*Eindhoven University of Technology*
*Department of Mathematics and Computing Science*
*Eindhoven, The Netherlands*

**Summary**
Some general remarks for experimental designs are made. The general statistical methodology of analysis for some special designs is considered. Statistical tests for some specific designs under Normality assumption are indicated. Moreover, nonparametric statistical analyses for some special designs are given. The method of determining the number of observations needed in an experiment is considered in the Normal as well as in the nonparametric situation. Finally, the special topic of designing an experiment in order to select the best out of $k(\geq 2)$ treatments is considered.

# Experiments: Design, Parametric and Nonparametric Analysis, and Selection

## Contents

## 1. Introduction

In various fields of industrial investigation or biological and agricultural research the observational data are of a variable nature. Even when the conditions of an experiment are kept constant as much as possible the outcome or response variable will vary from one trial to another. This variability is indicated by the term:

experimental error

or error. A perfect repetition of a treatment to experimental units is in practice impossible. There is always some variation in the effect of the same treatment applied to different experimental units. This difference is due to heterogeneity of the material, errors of observation, but also the failure to repeat the treatment exactly. Division of the material and allocation to the different treatments in a random way is therefore of essential importance. A completely randomized design is a plan for collecting data in which a random sample is selected from each treatment, and the samples are normally supposed to be independent.
However, this kind of error is often only a small part of the variability. The main reasons for variability are often uncontrolled instruments or environmental factors like

- temperature
- humidity
- pressure
- pollution

etc. Statistical designs of experiments and corresponding statistical parametric or nonparametric methods of analysis can help us to draw conclusions from data with random fluctuations.
In order to achieve that the variability is of a random nature a good design is important. Replication and randomization are, besides the use of blocks, two principal aspects of designing an experiment. Often the responses in an experiment are subject to sources of variation in addition to the treatments under study. Suppose for instance that in a field experiment a research worker is studying the yield resulting from four varieties of wheat. If the experimental area is divided into a number of smaller areas of equal size, the blocks or replications, and each of these blocks are divided into a number of plots (units), then randomization within each block is carried out and we have in this case a so-called 'randomized block design'. A randomized block design is a plan for collecting data in which each of $k$ treatments is measured once in each of $b$ blocks. The order of the treatments within the blocks is random.
In this report general concepts and ideas of statistical aspects of experimental designs are considered and discussed. A good design of an industrial or scientific experiment is essential for correct and efficient investigation. A correct design makes it possible to draw correct and justified conclusions.
For answering questions data have to be collected from experimental units. For industrial experiments machines, ovens and other similar objects form experimental units. For agricultural experiments equal sized plots of land, a single or a group of plants, a single or a group of animals are used as experimental units.

About sixty years ago the main principles of experimental design were formulated at Rothamsted Experimental Station in United Kingdom. The pioneer in this field was Ronald A. Fisher. He published the results in his book 'The Design of Experiments' in 1935. In 1966 the 8-th edition appeared. During the period after the appearance of Fisher's book on experimental design an impressive stream of papers and books dealing with design of experiments and the corresponding parametric and nonparametric statistical analysis have been published. The corresponding parametric analysis is indicated by the name 'analysis of variance'. Why do we find so many uses of statistics in the analysis of a design? Because many of the decisions we make are based on uncertain data. Moreover, there is a need for greater efficiency. Finally, there is also a need for more complex experiments. Not only one factor is of interest, often a large number of factors and their possible interactions are important and have possibly influence on the response variable. Most people are poor at measuring and distinguishing between large number of factors. A statistical design in which the factors can simultaneously be varied is of great importance. Statistics is not just a set of techniques, it is an attitude of mind in approaching data. In particular it acknowledges the uncertainty and variability in data and data collection. Statistics is making decisions in the face of this uncertainty. Designing experiments in a statistically sound way and the corresponding parametric or nonparametric statistical analysis are of vital importance for good decision making.
Before collecting data and analysing these data it is important to give careful thought to the proper design of an experiment. A project has in general three phases:

- the experiment
- the design
- the analysis.

A first question is: 'What is the goal of the experiment?' The experiment includes a statement of the problem to be solved.
The purpose of an experiment is the responsibility of the research worker, not of the statistician. The research worker has to answer the question: 'What is he going to do with the results of the experiment?'
It is necessary to define the dependent or response variable and the independent variables or factors which may affect the response variable. Are the factors qualitative or quantitative? Can they held constant? Are the levels of the factors certain fixed values or are they a random choice from a population of all possible levels? Experimentation, designing a good experiment and an adequate statistical inference are essential features of general scientific and industrial methodology. A proper design, such that the interpretation of the data can be done in a valid way, is essential.
There are many books written on designing experiments. To mention a few: Kempthorne (1952), Cochran and Cox (1957), Federer (1955), Montgomery (1984), Cox (1966). The last mentioned book is well-known and is mainly dealing the fundamental concepts of with designing an experiment. It can be seen as a basic guide-line for experiments design. The other books are mainly dealing with the analysis of experiments.
It is the purpose of the next chapter of this report to present the fundamental concepts in the design of experiments. In later chapters also the analysis will be indicated. Chapter 3 is dealing with special designs as Incomplete Block Designs and Fractional replication. Chapter 4 is dealing with the required number of observations under the Normality assumption, whereas

Chapter 5 presents a nonparametric approach. Chapter 6 gives an analysis, parametric and nonparametric, of a number of problems. Chapter 7 presents the selection problem: design and analysis. Finally, in Chapter 8 a number of remarks are made.

## 2. General remarks

In this chapter we shall describe general concepts and general designs as Randomized Blocks, Latin Squares and Factorial Designs.
Large uncontrolled variation is common in biological sciences. For this reason effects of treatments under investigation are often masked by fluctuations outside the experimenter's control and the designing of experiments is an essential part of scientific research in order to make it possible to draw valid conclusions in an efficient way.
Let us take the following example. In an agricultural experiment one wants to compare different varieties (treatments in general). The experimental area is divided into plots of equal size (experimental units or units in general) and the varieties are assigned randomly to the plots. In general, neighbouring plots tend to give yield more alike than distant plots. It is possible that there is a systematic variation across the field. A different year or a different field may produce substantial different results. The uncontrolled variation is often large compared with the treatment effects. It is essential in most cases (or always (!)) to plan a good experiment in order to detect differences in treatment effects while so much variation associated with the experimental units is present.
For a good experiment it is required that the experimental units receiving treatment $T_1$ differ in no systematic way from the experimental units receiving treatment $T_2$. In this way it is possible to estimate the true treatment effect of $T_1$ minus $T_2$ without a systematic error. In this context the key word is

*randomization.*

Randomization makes it possible that possible differences in experimental units are randomly divided to the treatments. The influence of this error is measured by the so-called *standard error*. It is the magnitude of the random errors in the estimate of the treatment contrast. In fact, the standard error of the difference between two treatments $T_1$ and $T_2$ is equal to $\sigma(n_1^{-1} + n_2^{-1})^{\frac{1}{2}}$, where $n_i$ is the number of observations (experimental units) of $T_i (i = 1, 2)$. The standard error tends to zero if $n_1$ and $n_2$ tend to infinity. The number of experimental units is important in order to get a sufficiently small standard error. However, too many units has a disadvantage that small treatment effects may be detected, which are generally of no practical importance. Then the investment in time and energy was too large.
From a practical point of view an important remark is that the whole experiment must be relatively simple. The reason is the fact that in most cases the experiment has to be carried out by unexperienced people, unexperienced in statistical principles and therefore not conscious of the importance of a statistically correct execution of the experiment. For instance the fact that randomization is of vital importance, is not always realized.
We assume that the treatment effects add on to the experimental unit term and that the effect of a treatment is constant during the whole experiment.
Usually, when the numbers of units for treatments $T_1$ and $T_2$ are equal, the difference between $T_1$ and $T_2$ is estimated by the difference of the means of all observations on $T_1$ and $T_2$, respectively.
Moreover, it is of course not allowed that an experimental unit is affected by the treatment applied to the other experimental units. In other words, it is supposed that there is no interference between different experimental units. One has to be careful when an experimental unit is used different times (overflow effect), or if different experimental units are in physical or psychological contact (dependency). In agricultural experiments guard rows are left out of

consideration. In psychological experiments using a person several times, it is possible that the response is dependent on the whole sequence of situations that have preceded it.
The precision (or standard error) can be improved by taking more experimental units. An alternative is to improve the design of the experiments. This aspect can be illustrated by a simple (but practically important) situation, namely the comparison of just two treatments $T_1$ and $T_2$. The main aspect of this situation will be the design. The statistical analysis of the data is closely related to the design. That's the reason we shall also give the analysis. Using this situation we shall discuss the three basic principles of experimental design, namely

- replication
- randomization
- blocking

*Replication* is the repetition of the basic experiment. It provides an estimate of the experimental error. Then it is possible to investigate a possible statistical significance of a difference in treatments. If $\sigma^2$ is the variance of the data and there are $n$ replicates, then the variance of the sample mean is $\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n}$, so a more precise estimate of the mean of $Y$ is possible. In general: More replications increase the accuracy of estimates of the treatment effects or more replications result in more accurate conclusions. If a treatment is allocated to $r$ experimental units in an experiment, it is said to be replicated $r$ times (not: $r-1$). If in a design each of the treatments is replicated $r$ times, the design is said to have $r$ replications.

*Randomization* is of vital importance. Statistical methods require that the observations or errors are independently distributed. Randomization usually makes this assumption correct. By randomization we mean that the allocation of the experimental material as well as the order in which the individual runs or trials of the experiment are to be performed are *randomly* performed. See also the instructive Example 5.9 in Cox (1966).
By proper randomization, the effects of possible extraneous factors will also be averaged out. This aspect can easily be illustrated in the example to be given for the comparison of two treatments.
The randomization of treatments can be realized as follows. Random numbers are drawn with the aid of a table with random digits (random tables) or computer. For $t$ treatments we need $at$ random numbers, where $a$ equals the number of times a treatment occurs in the whole experiment (or in a part of the experiment). Ranking the random numbers in increasing order, produces a random permutation of the treatments. It is clear that for each experiment we need new random digits. Using the same random digits for different experiments is a very bad attitude.
To use a systematic pattern is also very dangerous. The same order $(T_1, T_2)$ for each block can be totally wrong. For instance, when $T_1$ and $T_2$ are measured after each other and a difference in time has influence. Also a system of ordering $(T_1,T_2),(T_2,T_1),(T_1,T_2),(T_2,T_1),...$ can be very dangerous, when this pattern coincides with some pattern in the uncontrolled variation in an agricultural field experiment.
As a conclusion, we can formulate that randomization makes it possible that unbiased estimates of the treatment effects, unbiased estimate of the error variance, and exact significance tests concerning the treatment effects can be obtained.
*Blocking* is a cornerstone of experimental design. It provides a method in order to increase

the precision of an experiment, or more precisely a larger precision with the same number of observations. Reducing the error can be done by making experimental units homogeneous. This can be achieved by forming the experimental units into several homogeneous groups, usually called blocks, allowing variation between the blocks. A block is a portion of the experimental material that is more homogeneous than the whole collection of material.
All these aspects can be illustrated in the next example of the comparison of just two treatments $T_1$ and $T_2$.
If there are 20 experimental units, then 10 units can be randomly allocated to $T_1$ and the rest to $T_2$. The effect of this uncontrolled variation on the error of the treatment comparison can be reduced by obtaining pairs of units as alike as possible. We suppose that there is no systematic difference between the first and the second unit in a pair. The two units in a pair are expected to give as nearly as possible identical observations in the absence of treatment differences. The correct procedure is to randomize the order of $T_1$ and $T_2$, independently for each pair. The succes of 'paired comparisons' depends on an efficient pairing of the units.

## 3. Special designs

Let us consider the following situation. One wants to investigate the systematic difference between two methods $T_1$ and $T_2$ for determining the fat content of rats. One has the impression that the two methods have the same standard deviation. One has a population of rats (adults, and of the same sex) for the experiment. Now, one has two different designs.

The first design $D_A$ is to draw randomly $2n$ rats from the population and allocate randomly $n$ rats to $T_1$ and the rest to $T_2$. This design is called a complete randomized design. The responses are indicated by

$$Y_{11}, Y_{12}, ..., Y_{1n} \quad (\text{for } T_1)$$

$$Y_{21}, Y_{22}, ..., Y_{2n} \quad (\text{for } T_2)\ ,$$

two mutually independent samples of size $n$.
Under the assumption of Normality with expectations $\alpha_1$ and $\alpha_2$ and with common variance $\sigma_A^2$ the suitable test statistic for testing the null hypothesis

$$H_0 : \alpha_1 - \alpha_2 = 0$$

against the alternative hypothesis

$$H_1 : \alpha_1 - \alpha_2 \neq 0$$

is

$$T_A = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot} - (\alpha_1 - \alpha_2)}{\left\{\frac{2}{n} S_A^2\right\}^{\frac{1}{2}}}\ ,$$

with $\bar{Y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$ $(i = 1,2)$ and $S_A^2 = \frac{1}{2n-2}\left\{\sum_{j=1}^{n}(Y_{1j} - \bar{Y}_{1\cdot})^2 + \sum_{j=1}^{n}(Y_{2j} - \bar{Y}_{2\cdot})^2\right\}$.
Under $H_0$ the statistic $T_A$ has a $t$-distribution with $2n - 2$ degrees of freedom.

The second design $D_B$ is the design with paired observations. From the population of rats a random sample of $n$ rats are drawn. Both treatments (in random order) are applied to each of the $n$ rats. Per rat the difference $V$ of treatment $T_1$ minus treatment $T_2$ is determined. We suppose that the $V_j$ $(j = 1, 2, ..., n)$ are independently Normally distributed with mean $\alpha_0 (= EV)$ and variance $\sigma_B^2$. The relevant test statistic for $H_0 : \alpha_0 = 0$ against $H_0 : \alpha_0 \neq 0$ is

$$T_B = \frac{\bar{V} - \alpha_0}{\left\{\frac{1}{n} S_B^2\right\}^{\frac{1}{2}}}\ ,$$

with

$$S_B^2 = \frac{1}{n-1}\sum_{j=1}^{n}(V_j - \bar{V})^2 \ .$$

The test statistic $T_B$ has under $H_0$ a $t$-distribution with $n-1$ degrees of freedom.
The numerators of both test statistics are of equal merit, for $\alpha_0 = \alpha_1 - \alpha_2$ and application of the prescription $\bar{V}$ to the $2n$ observations will give the same result as the prescription $\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}$.
But the denominators are totally different.
For the design $D_A$, the complete randomized design, there are two components in the variability of fat contents:

- error of measurement
- variability of fat content of the rats.

For the design $D_B$, with paired observations, only the error of measurement is present. By taking differences the influence of fat content has been eliminated.
Suppose the variance of the error of measurements is $\sigma_m^2$, then $\frac{1}{n}S_B^2$ has expectation $\frac{1}{n}\sigma_B^2 = \frac{1}{n}2\sigma_m^2$ for difference has variance $\sigma_m^2 + \sigma_m^2 = 2\sigma_m^2$. For design $D_A$ the denominator (besides the square root) $\frac{2}{n}S_A^2$ has expectation $\frac{2}{n}\sigma_A^2 = \frac{2}{n}(\sigma_m^2 + \sigma_p^2)$, with $\sigma_p^2$ the variance of the rat population.
We have

$$\frac{2}{n}(\sigma_m^2 + \sigma_p^2) > \frac{2}{n}\sigma_m^2 \ \text{ if and only if } \sigma_p^2 > 0 \ .$$

From this it follows that under the alternative hypothesis using $D_B$ the null hypothesis will be rejected sooner than using $D_A$. The influence of the difference in critical values (for $2n-2$ and $2n$ degrees of freedom) is not so important.
For the two designs $D_A$ and $D_B$ the confidence intervals for $\alpha_1 - \alpha_2$ with confidence level 0.95 have expected midpoints $\alpha_1 - \alpha_2 = \alpha_0$ and lengths

$$L_A = 2t_{2n-2;0.975}\left\{\frac{2}{n}S_A^2\right\}^{\frac{1}{2}}$$

and

$$L_B = 2t_{n-1;0.975}\left\{\frac{1}{n}S_B^2\right\}^{\frac{1}{2}} \ ,$$

respectively, with $P(T_m \leq t_{m;1-\gamma}) = 1-\gamma$, where $T_m$ has a $t$-distribution with $m$ degrees of freedom. In general, $2S_A^2 >> S_B^2$ and the difference between $t_{2n-2;0.975}$ and $t_{n-1;0.975}$ is not so large, which is illustrated in next table.

| $n$ | $t_{2n-2;0.975}$ | $t_{n-1;0.975}$ |
|---|---|---|
| 2 | 4.303 | 12.706 |
| 3 | 2.776 | 4.303 |
| 4 | 2.447 | 3.182 |
| 5 | 2.306 | 2.776 |
| 10 | 2.101 | 2.262 |
| 20 | 2.025 | 2.093 |
| 30 | 2.002 | 2.045 |
| 40 | 1.994 | 2.023 |
| $\infty$ | 1.960 | 1.960 |

If $S_A^2 > S_B^2$ then $D_A$ is less efficient than $D_B$ for $n > 2$. Certainly, for $n \geq 5$ or 10 the difference between the critical values becomes rather small.

**Example**
Using design $D_B$ the following numerical results are given

| Treatment | | Diff. |
|---|---|---|
| 1 | 2 | $v$ |
| 15.4 | 16.1 | -0.7 |
| 15.4 | 16.5 | -0.1 |
| 15.9 | 15.5 | 0.4 |
| 16.7 | 17.6 | -0.9 |
| 16.0 | 17.6 | -1.6 |
| 17.4 | 18.5 | -1.1 |
| 17.2 | 17.9 | -0.7 |
| 17.3 | 18.3 | -1.0 |
| 15.5 | 17.9 | -2.4 |
| 16.0 | 16.0 | 0.0 |

From the results it follows that

$$\bar{v} = -0.81$$

$$\sum_{j=1}^{10}(v_j - \bar{v})^2 = 5.929 ,$$

thus

$$s_B^2 = \frac{5.929}{9} = 0.659 .$$

We get for testing $H_0 : \alpha_0 = 0$ the test statistic

$$t_B = \frac{-0.81}{\sqrt{0.659/10}} = -3.155$$

and $H_0$ is rejected, because $t_{9;0.975} = 2.262$ and thus $t_{9;0.025} = -2.262$. An estimate for $\sigma_m^2$ is equal to $\frac{0.659}{2} = 0.330$.
Now we can estimate $\sigma_A^2$ as if the observations have been obtained using design $D_A$. The statistics $\Sigma(Y_{1j} - \bar{Y}_{1.})^2$ and $\Sigma(Y_{2j} - \bar{Y}_{2.})^2$ are now stochastically dependent (for $Y_{1j}$ and $Y_{2j}$ are belonging to the same rat), but the expectation of both sums is $9\sigma_A^2$. The unbiased estimate of $\sigma_A^2$ is equal to

$$\tfrac{1}{18}\{2836.33 - (162.8)^2/10 + 2965.19 - (171.9)^2/10\} = 1.349 .$$

Thus an estimate of $\sigma_p^2$ is equal to

$$1.349 - 0.330 = 1.019 .$$

From this it follows that ($\gamma = .05$):

| $D_A$ | $D_B$ |
|---|---|
| $\hat{\sigma}_m = .574$ | $\hat{\sigma}_m = .574$ |
| $\hat{\sigma}_p = 1.009$ | |
| $\hat{\sigma}_A = 1.161$ | $\hat{\sigma}_B = .812$ |
| $L_A = 2 * 2.101 * .519$ | $L_B = 2 * 2.262 * .257$ |
| $= 2.18$ | $= 1.16$ |
| 'Conf.int.' | Conf.int. |
| $(-1.90, .28)$ | $(-1.39, -.23)$ |

Substitution in $t_A$ should give 1.859 with probability of exceedance less than .10 assuming the $t$-distribution holds (which is not true).

An interesting question is 'What number of observations for $D_A$ leads to the same length of the confidence interval?' The requirement is (neglecting the difference in critical values):

$$\frac{2S_A^2}{S_B^2} = \frac{2 * 1.349}{0.659} = 4.09$$

as many observations for $D_A$. Thus 10 rats for $D_B$ and $2 * 40 = 80$ rats for $D_A$. The relative efficiency of $D_A$ (relative with respect to $D_B$) is estimated by $\frac{10}{40} = \frac{1}{4}$.

**Randomized Blocks**

A natural generalization of paired comparisons is to consider the situation with $t(> 2)$ treatments. If there are $t$ treatments we can make blocks of size $t$. The units in each block are expected to give as nearly as possible the same observations if the treatments are equivalent (in their effect). The order of treatments is randomized within each block. Each treatment occurs once in each block and the randomizations for the blocks are independent. The comparison of the treatments will take place within blocks, so the effect of variations between blocks is eliminated, so far as treatment comparisons are concerned. The general idea of grouping is frequently used in simple experiments as well in more complicated designs. As comparison of treatments takes place within blocks and the effect of constant differences between blocks

is eliminated, a good grouping of experimental units into block is of fundamental importance.

**Latin Squares and Graeco-Latin Squares**

A natural extension of the randomized block design is the Latin Square. In the randomized block design there is one system of grouping. It might happen that there are systematic differences between the units within blocks. Then we should have two sytems of grouping, into blocks and into order within blocks. We would wish to balance out both systematic variations. A restriction which limits the use of Latin Squares is that the number of blocks (rows) and the number of experimental units in a block (columns) are equal. Assume this number is equal to $k$. Then a Latin Square is a design in which $k^2$ units are divided in three ways into $k$ classes of $k$ elements, such that the divisions are pairwise orthogonal (proportional representation). In each row and each column each treatment is present once and only once.

The following example has three treatments indicated with the Latin characters $A$, $B$ and $C$ (e.g. industrial processes). There are two kinds of blocking: one corresponding with rows (e.g. Location) and one corresponding with columns (e.g. Days).

| | | Day | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Location | 1 | A | B | C |
| | 2 | B | C | A |
| | 3 | C | A | B |

The Latin Square can be used when the experimental units are simultaneously grouped in two ways. If the experimental units are grouped in three ways, then a Graeco-Latin Square is suitable. For $k = 3$ we have

| | | Day | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Location | 1 | $A, \alpha$ | $B, \beta$ | $C, \gamma$ |
| | 2 | $C, \beta$ | $A, \gamma$ | $B, \alpha$ |
| | 3 | $B, \gamma$ | $C, \alpha$ | $A, \beta$ |

For example: The Latin characters $A$, $B$ and $C$ correspond with observers and the Greek characters $\alpha, \beta$ and $\gamma$ correspond with three industrial processes. Each observer measures only one process on each day. The Greek characters are situated such that each observer occurs once in combination with each industrial process, whereas each observer measures once in each day and once at each location.

For both kinds of squares the arrangement of treatments (persons, processes, locations, days) should be determined by randomization.

**Concomitant observations**

In order to reduce error or to increase precision not only grouping into blocks can be used, but also the use of concomitant or supplementary observations. Then with each main observation for which we try to find the treatment effects, for each experimental unit we have one (or more) concomitant observations. The condition is that the value for any unit must be unaffected by the particular assignment of treatments to units actually used. This can be realized

by observing the concomitant variable before the treatment is applied, or before the effect of the treatment has had time to develop. Examples of concomitant variables are the yield of a variety of wheat on a plot in previous years, or the weight of the heart of an experimental animal use in a biological assay. Also the weight at the start of a diet experiment may be important for an efficient analysis. The proper technique of analysis for such an experiment is the analysis of covariance. For a simple model it is often possible to find the effects graphically.

**Factorial designs and Fractional factorial designs**

Up to now we have considered experiments for which one factor (e.g. variety) is investigated. There are situations for which we wish to investigate simultaneously the effect of several factors. For instance, in a production process we are interested in the influence of temperature, pressure, proportions of reactants, humidity, and the concentration of a chemical component on the yield per time unity. Each treatment is a combination of these five factors, or more accurately a combination of one level from each factor, where a factor level is an investigated fixation of the concerned factor. If $k$ factors are varied each on two levels, it results in $2^k$ treatments. The investigation of all these treatments in one experiment is called a factorial experiment, or a $2^k$-factorial experiment. Analogously, there are $3^k$, $2^3 * 3$, ...-factorial experiments. Let us consider the following agricultural experiment. We want to investigate the yield per plot of wheat in dependence of the following four factors: variety (3 levels: varieties $A, B$ and $C$), nitrogeneous fertilizer $N$ (2 levels: no, yes), phosphate $P$ (2 levels: no, yes) and potash $K$ (2 levels: no, yes). This is a factorial experiment with four factors, one factor at 3 levels and three factors at 2 levels. We have in short a $3 * 2^3$-factorial experiment.

We have assumed that each combination of factor levels is used the same number of times. Such an experiment is called a complete factorial experiment. A special class of incomplete experiments is the class of fractional factorial experiments. This class of fractional factorial experiments is of special use when the number of factors is not small, and a large number of observations (experimental units) is not attainable.

A simultaneous investigation of the 4 factors in the example mentioned before has the following essential advantage. Main effects of the factors as well as the interactions (two-factor, three-factor interaction,...) can be estimated, and tested under assumption of normality of error. Four- and higher order interactions are normally difficult to interpret. That's the reason that high order interactions are not taken into the model, but are also used for estimating the error variance. Besides this advantage of simultaneous investigation of two or more factors, is the economic advantage. A $2^4$-factorical experiment with 3 replications requires $2^4 * 3 = 48$ experimental units. A (not recommended) alternative is to use $\frac{48}{4} = 12$ units for the investigation of each factor. The main effect of each factor can be investigated by comparison of 6 units on the one (low) level with 6 units on the other (high) level. In the factorial experiment the estimation of a main effect is based on the comparison of 24 units on the low level with 24 units on the high level. So the accuracy is much larger, not to mention the fact that in the factorial experiment information concerning interactions can be obtained. That's to say the examination of the extent to which the effect of one factor is different for different levels of another factor. When interaction is present, the estimate we obtain of one factor if the levels chosen for the other two factors happen to be totally different from those of final practical interest, may be quite misleading. Moreover, if the goal is to find the best combination of treatment combinations, the investigation of factors separately does not produce relevant information. As a conclusion we can state that factorial experiments have, compared with the one factor at a time approach, the advantages of greater precision for estimating overall factor effects, and the possibility of getting information about possible

interactions. It is clear that a large number of factors results in a very large experiment, which is in general not desirable. *Fractional* factorial experiments can give a way out for moderate values of $k$.

Factors can be classified as follows. Firstly, factors can represent treatments applied to the units, and factors which represent a classification, outside the investigator's control, of the units into different types. Secondly, we can represent the factors as (1) quantitative factors (e.g. temperature, pressure), (2) specific qualitative factors (e.g. varieties, different production processes representing qualitatively different methodologies), and (3) sampled qualitative variables, for which the levels are sampled from a population of levels (e.g. 5 persons from a population, 7 monsters from a population of raw material). With quantitative factors the methodology of response curves or response surfaces relating the true treatment effect to the quantitative carrier variables defining the factor levels. The absense of interaction between two quantitative factors means that the response surface is oriented in such a way that the effect of changing one factor is the same for all levels of the other factor. With specific qualitative factors we can work with main effects and interactions. The main effect of a factor $T_1$ gives us the differences in mean observation between the different levels of $T_1$ averaged over all levels of the other factors. The two-factor interaction between $T_1$ and $T_2$ examines whether, averaged over all levels of the remaining factors, the difference between levels of $T_1$ is the same for all levels of $T_2$ or vice versa. Similarly, a three-factor interaction between $T_1, T_2$ and $T_3$ examines whether, averaged over all levels of the remaining factors, the two-factor interaction between $T_1$ and $T_2$ has the same pattern for all levels of $T_3$ or, equivalently, whether the two-factor interaction between $T_2$ and $T_3$ has the same pattern for all levels of $T_1$, or equivalently, whether the two-factor interaction between $T_1$ and $T_3$ has the same pattern for all levels of $T_2$. Analogously for more factor interactions. We have already remarked that these many-factor interactions are very rarely of direct practical use.

For a sampled qualitative factor, the interaction of another contrast with it determines the error, when the other contrast is to be estimated for the whole (infinite) population of levels of this (sampled) qualitative factor.

In planning a factorial or fractional factorial experiment some practical steps can be considered. The first step is to make a list of factors of possible interest. This can be seen as a kind of brain storming. Then, in order to make the experiment practically realistic, one has to conclude which factors have to be included in the experiment. Then we have to consider at how many levels each factors should have to appear.

In order to investigate whether a quantitative factor has influence, the use of two levels (the difference as large as possible) is often sufficient. If one wants to have some estimate of the shape of the response curve, then three levels should be used. In this way it is possible to see whether the effect is non-linear. Then it can be seen whether an optimum is within the levels or outside the area of experimentation. More than three levels is for most situations not of practical interest. The response curve must be very complicated, will the use of the four or more levels be profitable.

In a small factorial experiment randomized blocks and Latin squares can be used in order to reduce the effect of a controlled variation.

The choice of the number of observations is of course an interesting topic. In the next chapter we will make some comments on it.

## 4. The number of observations: Normal distribution

The precision of the estimators of the treatment effects depends (among other things, like the design of the experiment and the error variation which is a function of the variability of the experimental material, the accuracy of the experimental work, and of the measurements) on the number of experimental units. If it is practically possible, given the design and the material, to increase the number of units (which is not always the case in practice), then we can make the precision of our estimates sufficiently high. The addition of 'sufficiently' means that we are in practice not interested in detecting differences which are not technically of interest. A difference that is statistically significant does not include a technically significant difference. Trying to get a too high precision is then waisting energy, time and money.
Suppose we have $t$ treatments $T_1, T_2, ..., T_t$. The true treatment effects are indicated by $\alpha_1, \alpha_2, ..., \alpha_t$ and we are interested in the contrasts (comparisons of the treatment effects) like $\alpha_1 - \alpha_2, \alpha_2 - \alpha_3, \frac{1}{3}(\alpha_1 + \alpha_2 + \alpha_3) - \frac{1}{2}(\alpha_4 + \alpha_5)$, etc. The sum of the coefficients of these linear combinations of the true treatment effects is equal to zero.
The error in the estimated contrast is the difference between the true and the estimated contrast. The average error is zero. As a measure of precision we define the standard error:

$$\{E(c_i - \gamma_i)^2\}^{\frac{1}{2}} ,$$

where $\gamma_i$ is the particular contrast and $c_i$ the corresponding estimator. If for example an unbiased estimator $c_1 = a_1 - a_2$ of $\gamma_1 = \alpha_1 - \alpha_2$ is build up by the mean of $n_1$ independent observations minus the mean of a different (independent) set of $n_2$ independent observations, then the standard error is equal to $\{n_1^{-1} + n_2^{-1}\}^{\frac{1}{2}}\sigma$, with $\sigma$ the residual standard deviation.

### The two-sample case

Let us consider in more detail the design of the comparison of two independent samples of independent observations. These observations correspond with two Normal random variables with parameters $(\alpha_1, \sigma_1^2)$ and $(\alpha_2, \sigma_2^2)$, respectively. Thus given are

$$Y_{11}, Y_{12}, ..., Y_{1n_1} \quad \text{and} \quad Y_{21}, Y_{22}, ..., Y_{2n_2} .$$

Assume that the total number of observations $n_1 + n_2 = 2n$ is given by limits of costs.
The problem is to determine $n_1$ and $n_2$, given $n_1 + n_2 = 2n$ ($n$ known integer larger than or equal to 1), such that

$$\text{var } (\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot})$$

is minimal.
If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown), then

$$\begin{aligned} \text{var } (\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) &= \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} \\ &= \left(\frac{1}{n_1} + \frac{1}{2n - n_1}\right)\sigma^2 \end{aligned}$$

and this is minimal for

$$\begin{aligned} -\frac{1}{n_1^2} + \frac{1}{(2n-n_1)^2} &= \frac{-(2n-n_1)^2 + n_1^2}{n_1^2(2n-n_1)^2} \\ &= \frac{4n(n_1-n)}{n_1^2(2n-n_1)^2} \\ &= 0 \,, \end{aligned}$$

which can easily be seen. From this it follows that $n_1 = n = n_2$ is the best choice for the sample sizes and var $(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) = \frac{2}{n}\sigma^2$.
If $\sigma_2^2 = f^2\sigma_1^2$ (with $f > 0$), then

$$\begin{aligned} \text{var}(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \\ &= \frac{\sigma_1^2}{n_1} + \frac{f^2\sigma_1^2}{2n-n_1} \\ &= \left(\frac{1}{n_1} + \frac{f^2}{2n-n_1}\right)\sigma_1^2 \end{aligned}$$

and this is minimal for

$$\begin{aligned} -\frac{1}{n_1^2} + \frac{f^2}{(2n-n_1)^2} &= \frac{-(2n-n_1)^2 + f^2 n_1^2}{n_1^2(2n-n_1)^2} \\ &= 0 \,, \end{aligned}$$

which can easily be verified.
The numerator is equal to zero for

$$\frac{2n-n_1}{n_1} = f$$

or

$$\frac{n_2}{n_1} = f \,.$$

The best choice for the sample sizes is

$$n_1 = \frac{2n}{f+1} \quad \text{and} \quad n_2 = 2n - \frac{2n}{f+1} = \frac{2nf}{f+1} \,.$$

If $n_1$ is not an integer, then rounding off into two directions furnishes the optimal integer $n_1$ (and $n_2 = 2n - n_1$).
For the variance of $\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}$ we get

$$\begin{aligned}\operatorname{var}(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) &= \left(\frac{1}{n_1} + \frac{f^2}{2n - n_1}\right)\sigma_1^2 \\ &= \left(\frac{f+1}{2n} + \frac{f^2(f+1)}{2nf}\right)\sigma_1^2 \\ &= \frac{(f+1)^2}{2n}\sigma_1^2 \,.\end{aligned}$$

For $f = 1$ we get the expression for $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Now we shall determine $n_1 = n_2 = n$ for two independent samples from distributions with equal known variance $\sigma^2$ such that the power of the $t$-test for two samples for testing

$$H_0 : \alpha_1 = \alpha_2$$

against

$$H_1 : \alpha_1 > \alpha_2$$

is at least $1 - \beta$.
The test statistic is

$$T = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot} - 0}{\sqrt{\frac{2}{n}\sigma^2}} \,,$$

which can be written as

$$\frac{(\bar{Y}_{1\cdot} - \alpha_1) - (\bar{Y}_{2\cdot} - \alpha_2) + \alpha_1 - \alpha_2}{\sigma\sqrt{\frac{2}{n}}}$$

$$= \chi + \delta \,,$$

where $\chi$ has (under $H_0$ as well as under $H_1$) a standard Normal distribution and $\delta = (\alpha_1 - \alpha_2)\sigma^{-1}\sqrt{\frac{n}{2}}$. The power of the test is equal to

$$P_{\alpha_1 - \alpha_2}(T \geq u_{1-\gamma}) = P(\chi + \delta \geq u_{1-\gamma}) \,,$$

with $u_{1-\gamma}$ defined by $P(\chi \leq u_{1-\gamma}) = 1 - \gamma$.
The power requirement

$$P(\chi + \delta \geq u_{1-\gamma}) = P(\chi \geq u_{1-\gamma} - \delta) \geq 1 - \beta$$

is fulfilled for

$$u_{1-\gamma} - \delta \leq u_\beta$$

or

$$\delta \geq u_{1-\gamma} - u_\beta = u_{1-\gamma} + u_{1-\beta} .$$

From this it follows that

$$(\alpha_1 - \alpha_2)\sigma^{-1}\sqrt{\frac{n}{2}} \geq u_{1-\gamma} + u_{1-\beta}$$

or

$$n \geq (u_{1-\gamma} + u_{1-\beta})^2 \left(\frac{\alpha_1 - \alpha_2}{\sigma\sqrt{2}}\right)^{-2} .$$

Let us consider the situation of unknown $\sigma$ and a two-sided alternative. It is possible to put a requirement on the confidence interval for $\alpha_1 - \alpha_2$.
A two-sided confidence interval for $\alpha_1 - \alpha_2$ equals

$$\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot} \pm t_{2n-2;1-\frac{1}{2}\gamma} S\sqrt{\frac{2}{n}} .$$

We can state the requirement

$$P\left(2t_{2n-2;1-\frac{1}{2}\gamma} S\sqrt{\frac{2}{n}} \leq L\right) \geq 1 - \beta .$$

The left-hand side can be written as (with $t := t_{2n-2;1-\frac{1}{2}\gamma}$)

$$P\left(4t^2\frac{2S^2}{n} \leq L^2\right)$$

$$= P\left(\frac{(2n-2)S^2}{\sigma^2} \leq \tfrac{1}{4}L^2 t^{-2}(n-1)n\sigma^{-2}\right)$$

$$= P\left(\chi^2_{2n-2} \leq \left(\frac{L}{\sigma}\right)^2 \frac{n(n-1)}{4t^2}\right) .$$

From the probability requirement it follows that

$$\left(\frac{L}{\sigma}\right)^2 \frac{n(n-1)}{4t^2} \geq \chi^2_{2n-2;1-\beta} .$$

It is not possible to solve for $n$, because $\chi^2_{2n-2;1-\beta}$ as well as $t_{2n-2;1-\frac{1}{2}\gamma}$ are functions of the sample size $n$. A trial and error method will be satisfactory. If $\sigma$ is known it is sufficient to express the length of the confidence interval as a factor times the known $\sigma$.

**One sample of paired observations**
We assume that one sample of $n$ independent paired observations

$$(Y_{11}, Y_{21}), (Y_{12}, Y_{22}), ..., (Y_{1n}, Y_{2n})$$

is given. Define $X_i = Y_{1i} - Y_{2i}$ $(i = 1, 2, ..., n)$. We assume that $X_1, X_2, ..., X_n$ are mutually independent, identically distributed Normal random variables with expectation $\alpha$ and known variance $\sigma^2$.
The problem is to determine the sample size $n$ such that the testing of

$$H_0 : \alpha = \alpha_0$$

against

$$H_1 : \alpha = \alpha_1 > \alpha_0 \ .$$

has a power of at least $1 - \beta$.
The test statistic

$$T = \bar{X}$$

is $N(\alpha, \frac{\sigma^2}{n})$.
Under $H_0$ we have

$$P_{\alpha_0}(\bar{X} \geq c) = \gamma$$

or

$$P_{\alpha_0}\left(\frac{\bar{X} - \alpha_0}{\sigma/\sqrt{n}} \geq \frac{c - \alpha_0}{\sigma/\sqrt{n}}\right) = \gamma \ .$$

Thus

$$\frac{c - \alpha_0}{\sigma/\sqrt{n}} = u_{1-\gamma}$$

or

$$c = \alpha_0 + \frac{\sigma}{\sqrt{n}} u_{1-\gamma} \ .$$

The power of the test is

$$P_{\alpha_1}(\bar{X} \geq \alpha_0 + \frac{\sigma}{\sqrt{n}} u_{1-\gamma})$$

$$= P_{\alpha_1}\left(\frac{\bar{X} - \alpha_1}{\sigma/\sqrt{n}} \geq \frac{\alpha_0 - \alpha_1}{\sigma/\sqrt{n}} + u_{1-\gamma}\right)$$

$$= P(\chi \geq -\delta + u_{1-\gamma})$$

with

$$\delta = \frac{\alpha_1 - \alpha_0}{\sigma/\sqrt{n}} .$$

From the probability requirement

$$P_{\alpha_1}(\chi \geq -\delta + u_{1-\gamma}) \geq 1 - \beta$$

it follows that

$$-\delta + u_{1-\gamma} = u_\beta .$$

Then

$$\delta - u_{1-\gamma} = -u_\beta = u_{1-\beta}$$

satisfies the requirement. From this it follows that

$$\frac{\alpha_1 - \alpha_0}{\sigma/\sqrt{n}} - u_{1-\gamma} = u_{1-\beta}$$

or

$$n = (u_{1-\beta} + u_{1-\gamma})^2 \left(\frac{\alpha_1 - \alpha_0}{\sigma}\right)^{-2} .$$

For given $\gamma$ and $\beta$ and the difference $\alpha_1 - \alpha_0$ expressed as a multiple times $\sigma$, the needed sample size can be found. In most cases rounding-off upwards will be necessary to get an integer value for the sample size $n$.
For a left-hand sided alternative the quantity $\delta$ is negative, but the same formula for $n$ (with only the absolute value of $\alpha_1 - \alpha_0$ of interest) can be used.
For two-sided testing of

$$H_0 : \alpha = \alpha_0$$

against the two-sided alternative

$$H_1 : \alpha \neq \alpha_0$$

using the test statistic $\bar{X}$, $H_0$ will be rejected if

$$|\bar{X} - \alpha_0| \geq u_{1-\frac{1}{2}\gamma}\frac{\sigma}{\sqrt{n}} .$$

The power of this test against an alternative $\alpha = \alpha_1$ is equal to

$$P_{\alpha_1}\left(|\bar{X} - \alpha_0| \geq u_{1-\frac{1}{2}\gamma}\frac{\sigma}{\sqrt{n}}\right)$$

$$= P_{\alpha_1}\left(|\frac{\bar{X} - \alpha_1 - (\alpha_0 - \alpha_1)}{\sigma/\sqrt{n}}| \geq u_{1-\frac{1}{2}\gamma}\right)$$

$$= P(|\chi + \delta| \geq u_{1-\frac{1}{2}\gamma})$$

$$= 1 - P(|\chi + \delta| < u_{1-\frac{1}{2}\gamma})$$

$$= 1 - P(-u_{1-\frac{1}{2}\gamma} - \delta < \chi < u_{1-\frac{1}{2}\gamma} - \delta) .$$

The requirement that the power has to be at least $1 - \beta$ leads to

$$P(-u_{\frac{1}{2}\gamma} - \delta < \chi < u_{1-\frac{1}{2}\gamma} - \delta) = \beta .$$

If $\delta$ is positive and sufficiently large, then $P(\chi \leq -u_{1-\frac{1}{2}\gamma} - \delta)$ can be neglected. Thus the requirement reduces to

$$P(\chi < u_{1-\frac{1}{2}\gamma} - \delta) = \beta .$$

From this it follows that

$$u_{1-\frac{1}{2}\gamma} - \delta = u_\beta$$

or

$$\delta = u_{1-\frac{1}{2}\gamma} - u_\beta = u_{1-\frac{1}{2}\gamma} + u_{1-\beta} .$$

If $\delta$ is negative and sufficiently large, then

$$\beta = P(-u_{1-\frac{1}{2}\gamma} - \delta < \chi < -\delta + u_{1-\frac{1}{2}\gamma})$$

$$\approx P(-u_{1-\frac{1}{2}\gamma} - \delta < \chi) \ .$$

Thus

$$-u_{1-\frac{1}{2}\gamma} - \delta = u_{1-\beta}$$

or

$$-\delta = u_{1-\beta} + u_{1-\frac{1}{2}\gamma} \ .$$

Thus combining the two results:

$$|\delta| = u_{1-\beta} + u_{1-\frac{1}{2}\gamma} \ ,$$

which gives

$$n = (u_{1-\beta} + u_{1-\frac{1}{2}\gamma})^2 \left(\frac{a_1 - a_0}{\sigma}\right)^{-2} \ .$$

Often a rounding-off to above is necessary to guarantee that the power is at least $1 - \beta$. The case that $\sigma$ is unknown and has to be estimated by $S^2$ (the pooled variance estimator; see section 3) can be dealed with in an analogous manner. The only difference is $u$ to be changed into $t_{n-1}$.

A different approach is to put the requirement that

$$P\left(2t_{n-1;1-\frac{1}{2}\gamma}\frac{S}{\sqrt{n}} \leq L\right) \geq 1 - \beta \ .$$

The left-hand side can be written as follows

$$P\left(2t_{n-1;1-\frac{1}{2}\gamma}\frac{S}{\sqrt{n}} \leq L\right)$$

$$= P\left(4t^2\frac{S^2}{n} \leq L^2\right)$$

$$= P\left(\frac{(n-1)S^2}{\sigma^2} \leq \frac{n(n-1)L^2}{4t^2\sigma^2}\right)$$

$$= P\left(\chi^2_{n-1} \leq \frac{n(n-1)}{4t^2}\left(\frac{L}{\sigma}\right)^2\right) \ ,$$

where $t = t_{n-1;1-\frac{1}{2}\gamma}$. The quantities $t$ and $\chi^2_{n-1}$ depend on $n$, so it is not possible to solve explicitely for $n$, but a trial and error method will work in practice to determine the required $n$. Often a rounding-off to above will be necessary.

## 5. The number of observations: Nonparametric situation

In this chapter we shall discuss the problem of determining the number of observations in some nonparametric situations. In other words for some familiar nonparametric or distribution-free tests we shall determine the minimal sample size such that the tests have power of at least $1-\beta$ against alternatives that differ sufficiently from the hypothesis being tested. In our discussion we shall consider the Sign Test, the Wilcoxon Signed Rank Test and the Wilcoxon Two-Sample Rank Test.
Our starting point is a $\gamma$-level test with test statistic $T$. In the classical situation we had to determine the sample size such that the requirement is met that the power against a parameter value (which differs from the null hypothesis) is sufficiently large. In the nonparametric situation there is in general not such a parameter. That is the reason we have to use a different approach. This approach is presented by Noether (1987).
We suppose that $T$ is (approximately) distributed as $N(\mu(T), \sigma^2(T))$. Under $H_0$ we have $\mu = \mu_0(T)$ and $\sigma = \sigma_0(T)$.
For an upper-tailed test the critical region is defined by

$$T > \mu_0(T) + u_{1-\gamma}\sigma_0(T) ,$$

and the power against the alternative $H_a$ is given by

$$P_a(T \geq \mu_0(T) + u_{1-\gamma}\sigma_0(T))$$

$$= P\left(\chi \geq \frac{\mu_0(T) - \mu(T)}{\rho\sigma_0(T)} + u_{1-\gamma}\rho^{-1}\right)$$

with $\rho = \frac{\sigma(T)}{\sigma_0(T)}$ and $P(\chi \leq u_{1-\gamma}) = 1-\gamma$.
The requirement that the power of the test must be equal to $1-\beta$ is met if

$$\frac{\mu_0(T) - \mu(T)}{\rho\sigma_0(T)} + u_{1-\gamma}\rho^{-1} = u_\beta (= -u_{1-\beta})$$

or if

$$\frac{\mu(T) - \mu_0(T)}{\sigma_0(T)} = u_{1-\gamma} + \rho u_{1-\beta} .$$

With

$$Q(T) := \left\{\frac{\mu(T) - \mu_0(T)}{\sigma_0(T)}\right\}^2$$

the requirement is fulfilled if

$$Q(T) = (u_{1-\gamma} + \rho u_{1-\beta})^2 .$$

We assume that $\rho = 1$. This is for instance true for shift alternatives. For alternatives that are close to the null hypothesis this assumption is in general approximately correct. We get

$$Q(T) = (u_{1-\gamma} + u_{1-\beta})^2$$

and we can solve for the number of observations. For the one-sample problem discussed in the previous chapter we have the test statistic $T = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $Q(\bar{Y}) = \left(\frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right)^2$, and solving for $n$ gives the same result we got in the previous chapter.

**Sign Test**

Assume one sample of independent observations is given: $Y_1, Y_2, ..., Y_n$. The associated continuous random variable has median $m$. We want to test the null hypothesis

$$H_0 : m = m_0 ,$$

with $m_0$ known. Subtracting $m_0$ from the $n$ observations $H_0$ changes into

$$H_0 : m = 0 .$$

The Sign Test, which can be applied, has the following test statistic

$$T = \#\{Y > 0\} .$$

Defining

$$p = P(Y > 0)$$

then

$$\mu(T) = np, \ \mu_0(T) = \frac{1}{2}n$$

and

$$\sigma^2(T) = np(1-p), \ \sigma_0^2(T) = \frac{1}{4}n .$$

By the assumption of continuity we have $P_0(Y = 0) = 0$, and $P_0(Y > 0) = \frac{1}{2}$.
To meet the power requirement we get

$$\left(\frac{np - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}}\right)^2 = (u_{1-\gamma} + u_{1-\beta})^2$$

or

$$n = \frac{1}{4}\frac{(u_{1-\gamma} + u_{1-\beta})^2}{(p - \frac{1}{2})^2} .$$

Against a known alternative $H_a : p(> \frac{1}{2})$ the required sample size can be determined. A choice of $p$ can be based on past experience. A different possibility is to put $\frac{P(Y>0)}{P(Y<0)} = c$, then $\frac{p}{1-p} = c$ thus $p = \frac{c}{c+1}$.
For the Sign Test: $\rho = \frac{\sqrt{np(1-p)}}{\frac{1}{2}\sqrt{n}} = 2\sqrt{p(1-p)} < 1$ for $p \neq \frac{1}{2}$, so the determined sample size is conservative.
If information on $p$, and thus on $\rho$, is available, then sometimes an improved estimate can be given, using $\rho u_{1-\beta}$ instead of $u_{1-\beta}$ in the formula for $n$.

**Wilcoxon Signed Rank Test**
Given are $n$ independent observations

$$Y_1, Y_2, ..., Y_n$$

from a continuous symmetric random variable. Suppose the distribution is symmetric about the unknown location parameter $m$. We wish to test the null hypothesis

$$H_0 : m = 0$$

versus the alternative hypothesis

$$H_1 : m > 0 .$$

It is possible to apply the Wilcoxon Signed Rank Test. The test statistic $T$ is defined as follows. The absolute values of the $n$ observations are ranked in increasing order of magnitude. The test statistic $T$ is defined as the sum of the ranks associated with the positive observations. Under $H_0$ the statistic $T$ is asymptotically (for $n \to \infty$) Normally distributed with

$$ET = \frac{1}{2}\sum_{i=1}^{n} i = \frac{1}{4}n(n+1)$$

and

$$\text{var } T = \frac{1}{4}\sum_{i=1}^{n} i^2 = \frac{1}{24}n(n+1)(2n+1) .$$

These expressions can easily be derived by noticing that under $H_0$ the statistic $T$ can be written as $\sum_{i=1}^{n} Z_i R(Y_i)$ where $R(Y_i)$ is the rank of $Y_i$ (after ranking the absolute values of the $n$ observations) and the $Z_i$ are independent and identically distributed random variables with $P(Z_i = 1) = P(Z_i = 0) = \frac{1}{2}$ so that $E(Z_i) = \frac{1}{2}$ and var $(Z_i) = \frac{1}{4}$. Since $T$ is a linear combination of these variables, its exact mean and variance are easily determined under $H_0$. It can be seen that

$$T = \#\left\{\frac{Y_i + Y_j}{2} > 0\right\} \quad (i \leq j)$$

the number of positive Walsh averages, for a rank $R_1$ associated with the smallest positive observation $Y_{R_1}$ means that the $R_1$ Walsh averages $\frac{Y_1+Y_{R_1}}{2}, \frac{Y_2+Y_{R_1}}{2}, \ldots, \frac{Y_{R_1}+Y_{R_1}}{2} = Y_{R_1}$ are positive. If the rank of the second positive observation $Y_{R_2}$ is equal to $R_2$ then the number of Walsh averages is increased with $R_2$, etc.
Now, we can write

$$\mu(T) = \mu(\sum_{1 \leq i \leq j \leq n} \sum T_{ij}) \ ,$$

where

$$T_{ij} = \begin{cases} 1 & \text{if} \quad \dfrac{Y_i + Y_j}{2} > 0 \\ 0 & \text{if} \quad \dfrac{Y_i + Y_j}{2} < 0 \ . \end{cases}$$

We define

$$p_1 = P(Y_i > 0)$$

$$p_2 = P(Y_i + Y_j > 0) \quad (i \neq j)$$

so

$$E\{T_{ii}\} = p_1$$

and

$$\begin{aligned} E\{T\} &= nE\{T_{ii}\} + \tfrac{1}{2}n(n-1)E\{T_{ij}\} \\ &= np_1 + \tfrac{1}{2}n(n-1)p_2 \ . \end{aligned}$$

Under $H_0$ we get

$$
\begin{aligned}
p_1 &= \tfrac{1}{2} \\
p_2 &= P(Y_i + Y_j > 0) \\
&= \int_{-\infty}^{\infty} \int_{-v}^{\infty} f_Y(u) f_Y(v) du dv \\
&= \int_{-\infty}^{\infty} \{1 - F_Y(-v)\} f_Y(v) dv \\
&= \int_{-\infty}^{\infty} F_Y(v) f_Y(v) dv \\
&= \tfrac{1}{2} .
\end{aligned}
$$

From this it follows that

$$
\begin{aligned}
Q(T) &= \left\{ \frac{np_1 + \frac{1}{2}n(n-1)p_2 - \frac{1}{4}n(n+1)}{[\frac{1}{24}n(n+1)(2n+1)]^{\frac{1}{2}}} \right\}^2 \\
&\approx \frac{\{\frac{1}{2}n^2 p_2 - \frac{1}{4}n^2\}^2}{\frac{1}{12}n^3} \\
&= 3n(p_2 - \tfrac{1}{2})^2
\end{aligned}
$$

for sufficiently large $n$. The power requirement results in

$$
3n(p_2 - \tfrac{1}{2})^2 = (u_{1-\gamma} + u_{1-\beta})^2
$$

or

$$
n = \tfrac{1}{3}(u_{1-\gamma} + u_{1-\beta})^2 (p_2 - \tfrac{1}{2})^{-2} .
$$

In practice we can make a choice of $p_2$ based on the ratio $r = \frac{P(Y_1+Y_2>0)}{P(Y_1+Y_2<0)} = \frac{p_2}{1-p_2}$, then we get $p_2 = \frac{r}{r+1}$.
We see that the required sample size for the Sign Test is smaller than the size for the Wilcoxon Signed Rank Test if and only if (approximately)

$$
\tfrac{1}{4}(u_{1-\gamma} + u_{1-\beta})^2 (p_1 - \tfrac{1}{2})^{-2} < \tfrac{1}{3}(u_{1-\gamma} + u_{1-\beta})^2 (p_2 - \tfrac{1}{2})^{-2}
$$

or

$$
\frac{(p_1 - \frac{1}{2})^2}{(p_2 - \frac{1}{2})^2} > \tfrac{3}{4}
$$

or

$$\left|\frac{p_1 - \frac{1}{2}}{p_2 - \frac{1}{2}}\right| > \tfrac{1}{2}\sqrt{3} .$$

For a number of distributions $p_1$ and $p_2$ can be determined and the tests can be compared.

**The two-sample test of Wilcoxon**
The test of Wilcoxon for two independent samples

$$Y_{11}, Y_{12}, ..., Y_{1m} \text{ and } Y_{21}, Y_{22}, ..., Y_{2n}$$

drawn from continuous distributions to test the null hypothesis

$$H_0 : F_{Y_1}(x) = F_{Y_2}(x) \text{ for all } x$$

against the alternative that $Y_2$-observations would tend to be larger than $Y_1$-observations. To detect this alternative we want to reject $H_0$ when the sum ranks of $Y_{21}, Y_{22}, ..., Y_{2n}$ in the combined sample are large in some sense. However, instead of measuring the tendency of the sum of $Y$-ranks to be large, we use the statistically equivalent test statistic of Mann-Whitney. The test statistic of Mann-Whitney $U$ is defined as

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij} ,$$

where the indicator random variables are defined as

$$D_{ij} = \begin{cases} 1 & \text{if } Y_{2j} > Y_{1i} \\ 0 & \text{if } Y_{2j} < Y_{1i} \end{cases}$$

for all $i$ and $j$. In words: $U$ is defined as the number of times a $Y_1$-observation precedes a $Y_2$-observation in the combined ordered arrangement of the two samples into a single sequence of $N = m + n$ variables increasing in magnitude.
We get

$$\begin{aligned} p &:= P(Y_2 > Y_1) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{v} f_{Y_2}(v) f_{Y_1}(u) du dv \\ &= \int_{-\infty}^{\infty} F_{Y_1}(v) f_{Y_2}(v) dv . \end{aligned}$$

Then

$$E\{U\} = \sum_{i=1}^{m}\sum_{j=1}^{n} E(D_{ij})$$

$$= mnp \ .$$

Under $H_0$ : '$F_{Y_1}(x) = F_{Y_2}(x)$ for all $x$' we get

$$p = \int_{-\infty}^{\infty} F_{Y_1}(v) f_{Y_1}(v) dv$$

$$= \tfrac{1}{2} \ .$$

The null and alternative hypothesis can be formulated more precisely as $p = \frac{1}{2}$ versus $p > \frac{1}{2}$. Thus

$$E_0\{U\} = \tfrac{1}{2}mn \ .$$

For the variance var $U$ we get

$$\text{var } U = \text{var}\{\sum_i \sum_j D_{ij}\}$$

$$= \sum_i \sum_j \text{var } D_{ij} + \sum_{i=1}^{m} \sum_{1 \le j \ne k \le n} \sum \text{cov}(D_{ij}, D_{ik}) +$$

$$+ \sum_{j=1}^{n} \sum_{1 \le i \ne h \le m} \sum \text{cov}(D_{ij}, D_{hj}) +$$

$$+ \sum_{1 \le i \ne h \le m} \sum \sum_{1 \le j \ne k \le n} \text{cov}(D_{ij}, D_{hk}) \ .$$

The random variables $D_{ij}$ are Bernoulli variables with

$$ED_{ij} = P(Y_2 > Y_1) = p$$

$$ED_{ij}^2 = 1^2 P(Y_2 > Y_1) = p$$

$$\text{var } D_{ij} = p - p^2 = p(1-p)$$

$$\text{cov}(D_{ij}, D_{hk}) = 0 \quad \text{for} \quad i \ne h \quad \text{and} \quad j \ne k$$

$$\text{cov}(D_{ij}, D_{ik}) = p_1 - p^2 \quad \text{for} \quad j \ne k$$

$$\operatorname{cov}(D_{ij}, D_{hj}) = p_2 - p^2 \quad \text{for} \quad i \neq h \ ,$$

where

$$\begin{aligned} p_1 &= P(Y_{2j} > Y_{1i} \cap Y_{2k} > Y_{1i}) \\ &= \int_{-\infty}^{\infty} \{1 - F_{Y_2}(x)\}^2 f_{Y_1}(x)\, dx \end{aligned}$$

and

$$\begin{aligned} p_2 &= P(Y_{1i} < Y_{2j} \cap Y_{1h} < Y_{2j}) \\ &= \int_{-\infty}^{\infty} F_{Y_1}^2(x) f_{Y_2}(x) dx \ . \end{aligned}$$

From this it follows

$$\begin{aligned} \operatorname{var} U = \ & mnp(1-p) + mn(n-1)(p_1 - p^2) + \\ & + mn(m-1)(p_2 - p^2) + 0 \\ = \ & mn\{p - p^2(N-1) + (n-1)p_1 + (m-1)p_2\} \ . \end{aligned}$$

When $F_{Y_2}(x) = F_{Y_1}(x)$, thus under $H_0$, it can easily be proved that $p_1 = \frac{1}{3} = p_2$. Then

$$\operatorname{var}_0 U = \tfrac{1}{12} mn(N+1) \ .$$

With

$$m = fN \quad \text{and} \quad n = (1-f)N$$

we find

$$\begin{aligned} Q(U) &= \frac{\{mnp - \frac{1}{2}mn\}^2}{\frac{1}{12}mn(N+1)} \\ &= \frac{12\{f(1-f)N^2 p - \frac{1}{2}f(1-f)N^2\}^2}{f(1-f)N^2(N+1)} \\ &\approx 12Nf(1-f)(p - \tfrac{1}{2})^2 \ . \end{aligned}$$

Thus the power requirement gives the next result

$$N = \tfrac{1}{12}(u_{1-\gamma} + u_{1-\beta})^2 (p - \tfrac{1}{2})^{-2} \{f(1-f)\}^{-1} \ .$$

An estimate of $r = \frac{P(Y_2 \geq Y_1)}{P(Y_2 < Y_1)}$ gives an estimate of $p = \frac{r}{r+1}$.

## 6. Analysis of variance and nonparametric analysis of some specific designs

In this chapter the analysis of some designs will be indicated. In many experiments the assumption of Normality is quite commonly made for the data analysis. However, there are experimental situations where this Normality assumption is not realistic. Nonparametric methods are methods for which the validity does not depend on the underlying distribution of the observations.The purpose of this chapter is to give an overview of a number of classical tests in some specific designs, side by side by a more or less comparable nonparametric analysis. For more detailed information we refer to Van der Laan and Verdooren (1987). In general we shall suppose that the observations have been drawn from continuous distributions, hence ties will occur with probability zero.

### 6.1. Two treatments with shift alternatives

*Under Normality*
Given are two independent samples of independent observations $X_1, X_2, ..., X_m$ and $Y_1, Y_2, ..., Y_n$ with

$$X_i \sim N(\mu_1, \sigma^2) \quad (i = 1, 2, ..., m)$$

and

$$Y_j \sim N(\mu_2, \sigma^2) \quad (j = 1, 2, ..., n)\,.$$

To test the null hypothesis

$$H_0 : \mu_1 = \mu_2$$

against $H_1 : \mu_1 - \mu_2 \neq 0, \mu_1 - \mu_2 > 0$ or $\mu_1 - \mu_2 < 0$ the $t$-test for two samples can be used.

*Nonparametric analysis*
Given are two independent samples of independent observations $X_1, X_2, ..., X_m$ and $Y_1, Y_2, ..., Y_n$ from populations with continuous distribution functions $F$ and $G$, respectively. To test the null hypothesis

$$H_0 : F \equiv G$$

against one- or two-sided alternatives the two-sample test of Wilcoxon can be applied. The test is sensitive to shift alternatives.
The condition of independency of the continuous observations is a sufficient one. The determination of the distribution of the test statistic $W$ of Wilcoxon under $H_0$ is only based on the property that the ranks of the $X$-observations can be considered as a random sample without replacement of size $m$ from the set $\{1, 2, ..., N = m + n\}$. From this it follows that the test of Wilcoxon can be applied in the situation where a treatment $A$ is applied to $m$ objects, randomly drawn from a population of $N$ objects, and a different treatment $B$ to the remaining $n$ objects.

## 6.2. More than two treatments with shift alternatives

*Under Normality*
Given are $k(>2)$ independent samples of independent observations $X_{i1}, X_{i2}, ..., X_{in_i}$ $(i = 1, 2, ..., k)$ from a classification $A$ with classes $A_1, A_2, ..., A_k$. Assume $X_{ij} \sim N(\mu_i, \sigma^2)$, $j = 1, 2, ..., n_i$. To test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

against $H_1$: 'at least one pair of $\mu$'s is unequal' the analysis of variance $F$-test can be used. The test statistic is

$$F = \frac{MSA}{MSE} \overset{H_0}{\sim} F_{N-k}^{k-1}$$

with $N = \sum_{i=1}^{k} n_i, MSA = \frac{SSA}{k-1}, MSE = \frac{SSE}{N-k}$ and the Sum of Squares of $A$ equals

$$SSA = \sum_{i=1}^{k} \left( \sum_{j=1}^{n_i} X_{ij} \right)^2 / n_i - \frac{1}{N} \left( \sum_{i,j} X_{ij} \right)^2$$

and the Sum of Squares of Error equals

$$SSE = \sum_{i,j} X_{ij}^2 - \sum_{i} \left( \sum_{j} X_{ij} \right)^2 / n_i \ .$$

$H_0$ is rejected for sufficiently large values.

*Nonparametric analysis*
Given are $k(>2)$ independent samples of independent observation $X_{i1}, X_{i2}..., X_{in_i}$ $(i = 1, 2, ..., k)$ drawn from populations with continuous distribution functions $F_i$. To test

$$H_0 : F_1 \equiv F_2 \equiv ... \equiv F_k \ ,$$

it is possible to apply the test of Kruskal-Wallis with test statistic

$$K = 12\{N(N+1)\}^{-1} \sum_{i} n_i (\bar{R}_{i.} - \bar{R}_{..})^2 \ ,$$

where $R_{ij}$ is the rank of $X_{ij}$ in the combined sample of $N$ observations. The test is sensitive to shift alternatives.

## 6.3. More than two treatments with ordered alternatives

*Under Normality*
For the case of $k$ independent samples we are interested in the comparison of $k(>2)$ ordered levels $z_1, z_2, ..., z_k$ of a quantitative treatment $A$. We assume $\mu_i = \varphi(z_i)$, $i = 1, 2, ..., k$, and wish to test

$$H_0 : \varphi(z) = \beta_0 + \beta_1 z$$

against the alternative hypothesis $H_1$: '$\varphi(z)$ is more complex than linear'.
We define the sum of squares for the linear trend of $A$, denoted by $SSL$, as

$$SSL = \left\{ \sum_i n_i z_i^2 - \frac{1}{N} \left( \sum_i n_i z_i \right)^2 \right\}^{-1} \left[ \sum_{i,j} z_i X_{ij} - \left( \sum_i n_i z_i \right) \left( \frac{1}{N} \sum_{i,j} X_{ij} \right) \right]^2 .$$

The test statistic is

$$\frac{SSA - SSL}{(k-2)MSE} ,$$

which is under $H_0$ distributed as $F_{N-k}^{k-2}$. Sufficiently large values lead to rejection of $H_0$.

*Nonparametric analysis*
It is possible to consider in the $k$-sample situation an ordered alternative

$$H_1 : F_1(x) \geq F_2(x) \geq ... \geq F_k(x) \quad \text{for all} \quad x ,$$

with at least one strict inequality sign for at least one $x$. Two possible tests are suggested: the test of Jonckheere-Terpstra and Chacko-Shorack, respectively. In the case of regular shifts (horizontal distances between the distribution functions are more or less equal) the test of Jonckheere-Terpstra is recommended. In the case of (strong) irregular shifts the test of Chacko-Shorack is preferable. The test statistic $T$ of Jonckheere-Terpstra is defined as

$$T = \sum_{i<j} T_{ij} ,$$

where $T_{ij}$ is the number of cases that an observation from sample $i$ is smaller than an observation from sample $j$ $(1 \leq i < j \leq k)$. For large values of $n_i$ is $T$ under $H_0$ approximately Normally distributed with

$$ET = \tfrac{1}{4}(N^2 - \sum_i n_i^2)$$

and

$$\text{var } T = \tfrac{1}{72}\{N^2(2N+3) - \sum_i n_i^2(2n_i+3)\} .$$

The test statistic of Chacko-Shorack can be determined as follows. Under $H_0$ one expects roughly $\bar{r}_{1\cdot} \leq \bar{r}_{2\cdot} \leq ... \leq \bar{r}_{k\cdot}$ (the lower case character $r$ is the outcome of the same capital character $R$; see Section 6.2). If this is not so for a pair, then the members of such a pair are put together (with sample size $n_i^*$) and again the average rank is computed. We continue with this operation, until

$$\bar{r}_{1\cdot}^* \leq \bar{r}_{2\cdot}^* \leq ... \leq \bar{r}_{l\cdot}^* ,$$

where $\bar{r}_{i\cdot}^*$ $(i = 1, 2, ..., l)$ are the average ranks of the ultimate groups with $n_i^*$ observations. The test statistic is

$$K^* = \frac{12}{N(N+1)} \sum_{j=1}^{l} n_j^* \{\bar{R}_{j\cdot}^* - \tfrac{1}{2}(N+1)\}^2 .$$

For large values of $n_1 = n_2 = ... = n_k$ we have approximately

$$P(K^* \geq c) = \sum_{i=2}^{k} p_{i,k} P(\lambda_{k-1}^2 \geq c) ,$$

where some $p_{i,k}$ can be found in the next table.

| $i$ \ $k$ | 3 | 4 | 5 | 6 |
|---|---|---|---|---|
| 2 | $\frac{1}{2}$ | $\frac{11}{24}$ | $\frac{5}{12}$ | $\frac{137}{360}$ |
| 3 | $\frac{1}{6}$ | $\frac{1}{4}$ | $\frac{7}{24}$ | $\frac{5}{16}$ |
| 4 | | $\frac{1}{24}$ | $\frac{1}{12}$ | $\frac{17}{144}$ |
| 5 | | | $\frac{1}{120}$ | $\frac{1}{48}$ |
| 6 | | | | $\frac{1}{720}$ |

## 6.4. Two treatments in randomized blocks of size two (orthogonal design)

*Under Normality*

In a block $B_j$ of size 2 $(j = 1, 2, ..., b)$ two treatments $A_1$ and $A_2$ are assigned at random to the units of the block. We assume

$$Y_{ij} = \alpha_i + \beta_j + E_{ij} \quad (i = 1, 2; j = 1, 2, ..., b) ,$$

where $\alpha_i$ is the expected effect of $A_i$, $\beta_j$ is the expected effect of $B_j$ and $E_{ij}$'s are independent $N(0, \sigma^2)$ random variables.
To test

$$H_0 : \alpha_1 = \alpha_2$$

against $H_1 : \alpha_1 \neq \alpha_2$ the test statistic $\frac{MSA}{MSE}$ can be used, where $MSA = \frac{SSA}{2-1}$ with $SSA = b^{-1} \sum_{i=1}^{2} \left( \sum_j Y_{ij} \right)^2 - (2b)^{-1} \left( \sum_{i,j} Y_{ij} \right)^2$ and $MSE = \frac{SSE}{b-1}$ with

$$SSE = \sum_{i,j} Y_{ij}^2 - \tfrac{1}{2}\sum_j \left(\sum_i Y_{ij}\right)^2 - SSA\ .$$

Under $H_0$ $\frac{MSA}{MSE}$ is distributed as $F_{b-1}^1$. Large values lead to rejection of $H_0$.

*Nonparametric analysis*
Assume $b$ independent pairs of observations $(X_1,Y_1),(X_2,Y_2),...,(X_b,Y_b)$ are given. A pair is for instance the result of two treatments applied in a block. To test

$$H_0 : P(X_i > Y_i) = P(X_i < Y_i) = \tfrac{1}{2} \text{ for all } i$$

against $H_1$ : $H_0$ is not true, the Sign test can be used. In general a more powerful test is the Wilcoxon Signed Rank test. With this test the null hypothesis $H_0$: '$Z_i = X_i - Y_i (i = 1,2,...,b)$ is symmetrically distributed around zero.' can be tested.

### 6.5. More than two treatments in randomized blocks (orthogonal design)

*Under Normality*
Given are $b$ blocks $B_j (j = 1,2,...,b)$ of size $\sum_{i=1}^{t} m_{ij}$. In each block $t(> 2)$ treatments $A_1, A_2, ..., A_t$ are applied at random, treatment $A_i$ in block $B_j$ is allotted at random to $m_{ij}$ units. We consider an orthogonal design, thus the relation $m_{ij} = m_{i.}m_{.j}/m_{..}$ holds. The model is

$$Y_{ijk} = \alpha_i + \beta_i + E_{ijk} \quad (i = 1,2,...,t; j = 1,2,...,b; k = 1,2,...,m_{ij})$$

with $\alpha_i$ the expected effect of $A_i$, $\beta_j$ that of $B_j$ and $E_{ijk}$ are i.i.d. $N(0,\sigma^2)$ r.v.
To test $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_t$ the test statistic is

$$T = \frac{MSA}{MSE}$$

with

$$MSA = \frac{SSA}{t-1}$$

$$SSA = \sum_i \left(\sum_{j,k} Y_{ijk}\right)^2 / m_{i.} - \frac{1}{m_{..}} \left(\sum_{i,j,k} Y_{ijk}\right)^2$$

$$MSE = (m_{..} - t - b + 1)^{-1} SEE$$

$$SSE = \sum_{i,j,k} Y_{ijk}^2 - \sum_j \left(\sum_{i,k} Y_{ijk}\right)^2 / m_{.j} - SSA\ .$$

Under $H_0$ the statistic $T$ is distributed as $F^{t-1}_{m..-t-b+1}$. Large values of $T$ lead to rejection of $H_0$.

*Nonparametric analysis*
To test the null hypothesis of no treatment effect, the observations $Y_{ijk}$ within block $B_j$ are ranked in increasing order of magnitude with ranks $1, 2, ..., m_{.j}$ $(j = 1, 2, ..., b)$. These ranks are $R_{ijk}$. The test statistic is defined as

$$Q_o = 12N \left\{ \sum_j m^2_{.j}(m_{.j}+1) \right\}^{-1} \sum_i m^{-1}_{i.} \left\{ R_{i..} - \tfrac{1}{2} \sum_j m_{ij}(m_{.j}+1) \right\}^2$$

with $R_{i..}$ the sum of the ranks for treatment $A_i$. $Q_o$ has under $H_0$ asymptotically a $\chi^2_{t-1}$-distribution. The condition of the approximation and the treatment of ties can be found in Benard and Van Elteren (1953).
If $m_{ij} = 1$ for all $i$ and $j$, the test statistic $Q_o$ simplifies to the test statistic of Friedman.
An alternative procedure based on standardized ranks has been proposed by De Kroon and Van der Laan (1983). They suggested as test statistic

$$\tilde{Q} = \sum_i m_{i.}(\bar{\tilde{R}}_{i..})^2$$

with

$$\bar{\tilde{R}}_{i..} = \frac{1}{m_{i.}} \sum_{j,k} \tilde{R}_{ijk}$$

and

$$\tilde{R}_{ijk} = \{R_{ijk} - \tfrac{1}{2}(m_{.j}+1)\}\{12^{-1} m_{.j}(m_{.j}+1)\}^{-\frac{1}{2}} .$$

Under $H_0$ the statistic $\tilde{Q}$ has asymptotically (for $m_{.j} \to \infty$) a $\chi^2_{t-1}$-distribution.

## 6.6. More than two treatments in randomized blocks and ordered alternatives (orthogonal design)

*Under Normality*
As described in Section 6.3 we are interested in the investigation of a quantitative treatment $A$ with levels $z_1, z_2, ..., z_t$. We consider a randomized block design with one observation per cell and the model

$$X_{ij} = \mu + \alpha_i + \beta_j + E_{ij}, \ (i = 1, 2, ..., t; j = 1, 2, ..., b) ,$$

where $\mu$ is the general mean, $\alpha_i$ is the deviation from $\mu$ for treatment $A_i$, $\beta_j$ that of block $B_j (\sum_i \alpha_i = 0 = \sum_j \beta_j)$ and $E_{ij}$ are i.i.d. $N(0, \sigma^2)$ r.v..
To test $H_0$: '$\phi(z) = \gamma_{0j} + \gamma_1 z$ in block $j$' against $H_1$: '$\phi(z)$ is more complex than linear', we start with the computations as described in Section 6.5 for $m = 1$.
The test statistic is

$$\frac{SSA - SSL}{(t-2)MSE}$$

with

$$SSL = \{\sum_i bz_i^2 - \frac{1}{bt}(\sum_i bz_i)^2\}^{-1}\{\sum_{i,j} z_i X_{ij} - \frac{1}{bt}(\sum_i bz_i)(\sum_{i,j} X_{ij})\}^2 \quad ,$$

and has under $H_0$ a $F^{t-2}_{(t-1)(b-1)}$-distribution. Large values lead to rejection of $H_0$.

*Nonparametric analysis*
In Friedman's block design with one observation per cell, we assume

$$X_{ij} = \mu + \alpha_i + \beta_j + E_{ij}$$

($\Sigma\alpha_i = 0 = \Sigma\beta_j$) with $E_{ij}$ independently and continuously distributed random variables which are identically distributed in each block. For testing $H_0$ of no treatment effect against $H_1 : \alpha_1 \leq \alpha_2 \leq ... \leq \alpha_t$ with at least one strict inequality sign, the test statistic of Page (1963)

$$L = \sum_i iR_i.$$

can be used. Ranking within blocks, the same as for Friedman's test. $H_0$ is rejected when $L$ is sufficiently large. For large values of $b$ the statistic $L$ has approximately a Normal distribution with

$$EL = \tfrac{1}{4}bt(t+1)^2$$

and

$$\text{var } L = \frac{1}{144(t-1)}b(t^3 - t)^2 \ .$$

## 6.7. More than one treatment in a Balanced Incomplete Block Design (BIBD)

*Under Normality*
It is in practice possible that a block can only contain at most $k$ experimental units. A possible reason may be that blocks with a size larger than $k$ are not homogeneous enough. If $t > k$ we need a so-called incomplete block design. Suppose that for $t$ treatments in blocks of size $k$ a design can be constructed for which

- every treatment occurs on $r$ units
- every pair of treatments occurs in $\lambda$ of the $b$ blocks.

If $b$ blocks are required, then $rt = bk$ and $\lambda(t-1) = r(k-1)$. Such a design (if it exists) is called a Balanced Incomplete Block Design. Randomization of blocks, treatments and units is necessary.
To test $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_t$ we use the test statistic

$$T = \frac{MSA}{MSE} .$$

Let $T_i$ be the total of the $Y_{ij}$'s of treatment $A_i$, $B_i^*$ be the total of the blocks in which treatment $A_i$ occurs. Then we define

$$Q_i = \frac{1}{k}(kT_i - B_i^*)$$

and

$$MSA = \frac{SSA}{t-1}$$

$$SSA = \frac{1}{\lambda tk}\sum_i (kQ_i)^2$$

$$MSE = (tr - t - b + 1)^{-1} SSE$$

$$SSE = \sum_{i,j} Y_{ij}^2 - \frac{1}{k}\sum_j (\text{Block } j)^2 - SSA ,$$

where Block $j$ is the total of the $k$ observations in block $j$. $T$ has under $H_0$ a $F_{tr-t-b+1}^{t-1}$-distribution.
Large values of $T$ lead to rejection of $H_0$.

*Nonparametric analysis*
In a BIBD the rank test of Durbin can be applied, which is based on the test statistic

$$D = \frac{12(t-1)}{rt(k^2-1)}\sum_i \{R_i - \tfrac{1}{2}r(k+1)\}^2$$

with $R_i$ the sum of the ranks for $A_i$ after ranking within blocks where $A_i$ occurs.
The asymptotic distribution of $D$ under $H_0$ is $\chi_{t-1}^2$. The exact distribution can be found in Van der Laan and Prakken (1972).

### 6.8. Interaction in a two-way layout

*Under Normality*
Assume we have two factors $A$ and $B$, with $A$ on $s$ levels and $B$ on $t$ levels. A complete factorial experiment of $A$ and $B$ is performed, where each treatment combination has been executed $m(>1)$ times. The design is a completely randomized design with $stm$ units.
Let $X_{ijk}$ be the $k$-th observation of the treatment combination $(A_i, B_j)$ with the model

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk}$$

$$(i = 1,2,...,s; j = 1,2,...,t; k = 1,2,...,m)\,,$$

where $\mu$ is the general mean, $\alpha_i$ is the deviation from the general mean for $A_i, \beta_j$ that for $B_j$ and $\gamma_{ij}$ is the deviation from the general mean due to the non-additivity effect or interaction effect of $A_i$ and $B_j$. Then the following holds:
$\sum_i \alpha_i = 0 = \sum_j \beta_j$ and $\sum_i \gamma_{ij} = 0 = \sum_j \gamma_{ij}$ (for all $j$ and $i$, respectively). The error terms $E_{ijk}$ are i.i.d. $N(0,\sigma^2)$ random variables.
To test

$$H_0 : \gamma_{ij} = 0 \text{ for all } i \text{ and } j$$

the test statistic is

$$T = \frac{MSAB}{MSE}$$

with

$$MSAB = \frac{SSAB}{(s-1)(t-1)}$$

$$SSAB = \frac{1}{m}\sum_{i,j}(\sum_k X_{ijk})^2 - \frac{1}{tm}\sum_i(\sum_{j,k} X_{ijk})^2 -$$

$$-\frac{1}{sm}\sum_j(\sum_{i,k} X_{ijk})^2 + \frac{1}{stm}(\sum_{i,j,k} X_{ijk})^2$$

$$MSE = \frac{1}{st(m-1)}SSE$$

$$SSE = \sum_{i,j,k} X_{ijk}^2 - \frac{1}{m}\sum_{i,j}(\sum_k X_{ijk})^2\,.$$

$T$ has under $H_0$ a $F^{(s-1)(t-1)}_{st(m-1)}$-distribution with a righthand-sided critical region.

*Nonparametric analysis*
(Testing against rank-interaction)
The concept of rank-interaction, a nonparametric concept of interaction, has been introduced by De Kroon and Van der Laan (1981). In each cell $(i,j)$, a combination of the $i$-th level of factor $A$ and the $j$-th level of factor $B$, are given $m(>1)$ observations. The error terms $E_{ijk}$ are independently and identically distributed random variables with continuous distribution functions and $E(E_{ijk}) = 0 (i = 1,2,...,s; j = 1,2,...,t$ and $k = 1,2,...,m)$. We wish to

$$H_0 : \beta_1 + \gamma_{i1} = \beta_2 + \gamma_{i2} = ... = \beta_t + \gamma_{it} \text{ for } i = 1,2,...,s\,.$$

The choice of a test procedure depends on the alternative hypothesis one is interested in. Depending on which alternative is thought to be important in the practical problem, one can choose one of three statistics $T, T_1$ and $T_2$. As an omnibus test we have

$$T = \sum_{i=1}^{s} K_i ,$$

where $K_i$ is the Kruskal-Wallis statistic computed for the classification $B$ within class $i$ of factor $A$. The test statistc $T_1$ is defined to be the Friedman test statistic for the case of $s$ blocks of ranking over $t$ classes of factor $B$, each class containing $m$ observations. $T_1$ will be used if differences between the $\beta_j$'s are interesting. The test statistic $T_2 = T - T_1$ can be used if one is mainly concerned with rank-interaction. Roughly speaking rank-interaction can be understood as the phenomenon where the ranks of the response variable $X$ for the levels of factor $B$ are different for some levels of factor $A$. In other words not all deviations from zero of the $\gamma_{ij}$'s are of interest, but only those deviations that give discordance between the rankings within the levels of factor $A$. If the rankings of

$$\beta_1 + \gamma_{i1}, \beta_2 + \gamma_{i2}, ..., \beta_t + \gamma_{it} \quad (i = 1, 2, ..., s)$$

are not identical for different values of $i$, we say that rank-interaction $B^*(A)$ exists. If these rankings are identical ("concordance"), rank-interaction $B^*(A)$ is said not to exist.
It is also possible to use the statistics $T_1$ and $T_2$ simultaneously, each at significance level $\alpha/2$, rejecting $H_0$ if at least one of the outcomes of $T_1$ or $T_2$ is larger than or equal to the corresponding critical value. In this way it is possible to carry out an omnibus test with the possibility of detecting simultaneously which component of $H_0$ is not true. This procedure is somewhat, probably slightly, conservative. For critical values for all three tests and further details of the various procedures we refer to De Kroon and Van der Laan (1981). Extensive tables with critical vlaues can be found in Van der Laan (1987). For power comparisons see De Kroon and Van der Laan (1984).

## 7. Design and analysis of selection experiments

In this chapter the design and analysis of selection experiments will be discussed. In practice we are often confronted with the problem of selection. Especially in the field of testing varieties selection is an essential feature. But also in biology, pharmacology, industry, etc. a large number of problems are in fact selection problems. For all kind of selection problems in practice a quantitative methodology of selection is needed. Statistical estimation and hypotheses testing provide a methodology which can help us to analyse the observations. This formulation, in terms of estimating parameters and testing statistical hypotheses, does not exactly suit the objectives of an experimenter in a number of situations.
If an agricultural experimenter is investigating a number $k$ of wheat varieties, characterized by the population mean yield $\theta$ per plot, he usually wants to select the best variety. The goal of the experiment is often not to accept or reject the homogeneity hypothesis (equality of the population means) but to select the best variety, where best is associated with the maximum value of $\theta$. In such an agricultural experiment one would expect that the wheat varieties are essentially (genetically) different. So one would expect to reject the homogeneity hypothesis if the sample sizes are sufficiently large. Then we are faced with a result that can not be a final decision. This makes the test of homogeneity not always realistic. Of course, the method of multiple comparison and simultaneous confidence intervals can give additional information. However, selection theory gives in a number of situations a more realistic formulation of the problem.
We suppose that the $k$ populations are described by qualitative variables. Some examples are the following. In an agricultural experiment $k$ varieties of wheat are given. A comparison is made on the basis of the response: yield per acre. In medicine $k$ types of drug are given. The goal is to select the drug with the maximal number of hours without pain. In chemical engineering $k$ types of catalyst are investigated with as response variable the number of gallons per day.
It is also possible that the populations are described by quantitative variables, for instance amount of fertilizer, amount of drug, temperature of a reaction, respectively, with the same response variable as before. Whereas statistical selection procedures are developed for finding the best population for qualitative populations, the response surface analysis is an adequate technique to find the optimal doses for quantitative variables. For more detail we refer to Box, Hunter and Hunter (1978).
Assume $k(\geq 2)$ independent Normal random variables $X_1, X_2, ..., X_k$ are given. These variables are associated with the $k$ populations and may be sample means. The unknown means of the $k$ populations are denoted by $\theta_1, \theta_2, ..., \theta_k$. The experiments design can be a complete randomized design with $n$ plots or a randomized complete block design with block size $k$ and the plots randomly associated to the $k$ populations. The goal is to select the population with mean $\theta_{[k]}$, where $\theta_{[1]} \leq \theta_{[2]} \leq ... \leq \theta_{[k]}$ denote the ordered values of $\theta_1, \theta_2, ..., \theta_k$.
There are two main approaches in the field of selection methodology. These main approaches are introduced by Robert Bechhofer and Shanti Gupta, respectively. The approach of Bechhofer is indicated by "Indifference Zone Approach" and the method of Gupta is indicated by "Subset Selection". The basic papers are Bechhofer (1954) and Gupta (1965), but afterwards they have published quite a lot of papers in the field of selection. An overview of selection methods is given in Gupta and Panchapakesan (1979) (which contains already about 750 references), Gupta (ed., 1977), Rizvi (ed., 1985, 1986), Dudewicz and Koo (1982) and Van der Laan and Verdooren (1989).
The subset selection procedure selects a subset, non-empty and as small as possible, of the $k$

populations considered in order to include the best population with the probability requirement that the probability of a Correct Selection is at least $P^*$. A Correct Selection (CS) means in this context that the best population is an element of the selected subset. The subset selection approach has certain advantages in practice. It can be applied to analyse the data after the experiment has been realized. In the context of this report we are interested in designing an experiment. That is the reason we shall concentrate on the approach of Bechhofer: The Indifference Zone approach.

The Indifference Zone approach is an important approach for designing an experiment. The goal is to indicate the best population. It provides a value for the common sample size needed to meet certain probability requirements. A Correct Selection (CS) means in this context that the best population is indicated. The probability requirement is that the probability of a $CS$ is at least $P^*$, whenever the best population is at least $\delta^*$ away from the second best. The minimal probability $P^*$ can only be guaranteed if the common sample size $n$ is large enough. In the next section this procedure will be described in more detail.

**Bechhofer's approach to selection: Indifference Zone procedures**

Assume $k(\geq 2)$ independent populations denoted by $G = (\pi_1, \pi_2, ..., \pi_k)$ are given. The related independent random variables are denoted by $Y_1, Y_2, ..., Y_k$. The random variable $Y_i$ has cumulative distribution function $F(y; \theta_i)$ with the unknown real-valued parameter $\theta_i \in \Theta (i = 1, 2, ..., k)$. Let $\Omega$ denote the parameter space $\{\theta : \theta = (\theta_1, \theta_2, ..., \theta_k); \theta_i \in \Theta, i = 1, 2, ..., k\}$.
The ranked parameter values are indicated by $\theta_{[1]} \leq \theta_{[2]} \leq ... \leq \theta_{[k]}$. The population associated with $\theta_{[i]}$ will be denoted by $\pi_{(i)}$. Usually in ranking the population $\pi_j$ is better than $\pi_i$ if $\theta_j > \theta_i$. Then we define the population $\pi_{(k)}$ associated with $\theta_{[k]}$ as the best population. The $t (1 \leq t < k)$ best populations are the populations $\pi_{(k-t+1)}, \pi_{(k-t+2)}, ..., \pi_{(k)}$. If there are more than $t$ contenders because of ties, it is assumed that $t$ of these are appropriately tagged.
The goal of selection considered in this paper is to select an unordered set of $t$ populations associated with the set $\{\theta_{[k-t+1]}, \theta_{[k-t+2]}, ..., \theta_{[k]}\}$. In this context a correct selection means that the $t$ best populations are selected.
Let $\delta_{ji}$ denote a measure of distance between the populations $x_{(i)}$ and $\pi_{(j)}$ with $1 \leq i < j \leq k$. In case $\theta$ is a location parameter, $\delta_{ji}$ is usually defined as $\theta_{[j]} - \theta_{[i]}$.
The probability requirement that the probability of selecting the $t$ best populations is at least $P^*$ can be written as follows. Denoting the probability of a correct selection (CS) using a selection procedure $R$ by $P(CS|R)$ or $P(CS)$, one can write the probability requirement as $P(CS) \geq P^*$ for $\theta \in \Omega(\delta^*) = \{\theta : \delta_{k-t+1,k-t} \geq \delta^*\}$.
The experimenter has to specify positive constants $\delta^*$ and $P^*$, where $P^* \in \left( \binom{k}{t}^{-1}, 1 \right)$.
Usually the selection rule is based on sufficient statistics for $\theta_1, \theta_2 ..., \theta_k$. These sufficient statistics are based on samples of common size $n$ from the $k$ populations.
The general problem in this context is to determine the smallest common sample size $n$ for which

$$\inf_{\Omega(\delta^*)} P(CS) \geq P^* .$$

This condition is called the $P^*$-condition for the probability requirement. The infimum of the $P(CS)$ is evaluated over the subset $\Omega(\delta^*)$ of the parameter space $\Omega$. $\Omega(\delta^*)$ is the so-called preference zone and $\Omega^c(\delta^*)$ is called the indifference zone.

Let us consider the situation of $k$ Normal populations with common known variance $\sigma^2$ in more details. The selection rule $R$ is based on the sufficient statistic $\bar{Y}_i$ for $\theta_i$, where $\bar{Y}_i$ is the mean of a sample of $n_i$ independent observation from $\pi_i (i = 1, 2, ..., k)$. The goal is to partition the set $G$ into 2 subsets $G_1$ and $G_2$ with $G = G_1 \cup G_2, G_1 \cap G_2 = \emptyset, G_1 = \{\pi_{(k-t+1)}, \pi_{(k-t+2)}, ..., \pi_{(k)}\}$ and $G_2 = \{\pi_{(1)}, \pi_{(2)}, ..., \pi_{(k-t)}\}$.
The selection rule $R$ is defined as follows. We determine a subset $S \subset G$ of size $t$ on the basis of the sample means. Include in $S$ the populations associated with

$$\bar{Y}_{[k-t+1]}, \bar{Y}_{[k-t+2]}, ..., \bar{Y}_{[k]} ,$$

where $\bar{X}_{[1]} \leq X_{[2]} \leq ... \leq \bar{X}_{[k]}$ are the ranked sample means. The $P^*$-condition is

$$P(CS) = P[S = G_1 | \theta \in \Omega(\delta^*)] = P[S = G_1 | \delta_{k-t+1,k-t} \geq \delta^*] \geq P^* .$$

We have for $n_i = n$ and $\bar{Y}_{(i)}$ is the sample mean associated with $\pi_{(i)}$:

$$P(S = G_1 | \theta \in \Omega(\delta^*)] = P[\max_{1 \leq i \leq k-t} \bar{Y}_{(i)} < \min_{k-t+1 \leq i \leq k} \bar{Y}_{(i)} | \theta \in \Omega(\delta^*)]$$

$$\geq t \int_{-\infty}^{\infty} \Phi^{k-t}(z + \tau)\{1 - \Phi(z)\}^{t-1} d\Phi(z) ,$$

where $\Phi$ is the standard Normal distribution function and

$$\tau = n^{\frac{1}{2}} \delta^* \sigma^{-1} .$$

The minimum of the $P(CS)$ for $R$ is attained for the so-called Least Favourable Configuration (LFC) in $\Omega(\delta^*)$, given by

$$\theta_{[1]} = ... = \theta_{[k-t]} = \theta_{[k-t+1]} - \delta^* = ... = \theta_{[k]} - \delta^* .$$

The minimum sample size required is the smallest integer $n$ for which the $P^*$-condition is fulfilled.
For the LFC we have

$$P_{LFC}(CS) = t \int_{-\infty}^{\infty} \Phi^{k-t}(z + \tau)\{1 - \Phi(z)\}^{t-1} d\Phi(z) ,$$

which is an increasing function of $\tau$ and tends to 1 as $\tau$ tends to infinity. Hence, there is a unique smallest value $\tau$ meeting the probability requirement. This value is the solution of the equation.

$$P_{LFC}(CS) = P^* .$$

For the special case $t = 1$, the rule $R$ selects the population associated with $\bar{Y}_{[k]}$. Then for the $P^*$-condition we have

$$P[S = G_1 | t = 1, \theta \in \Omega(\delta^*)] = P[\max_{1 \leq i \leq k-1} \bar{Y}_{(i)} < \bar{Y}_{(k)} | \theta \in \Omega(\delta^*)]$$

$$\geq \int_{-\infty}^{\infty} \Phi^{k-1}(z + \tau) d\Phi(z) \geq P^* .$$

In order to meet the probability requirement in the case $t = 1$ we have to choose:

$$n = \left(\frac{\tau\sigma}{\delta^*}\right)^2 .$$

In order to be sure that the common sample size is large enough to satisfy the probability requirement, the computed value of $n$ is rounded upward if it is not an integer. The quantity $\tau$, which depends on $k$ and $P^*$, can be computed by numerical integration. Some values can be found in next table.

| $k$ \ $P^*$ | .90 | .95 |
|---|---|---|
| 2 | 1.812 | 2.326 |
| 3 | 2.230 | 2.710 |
| 4 | 2.452 | 2.916 |
| 5 | 2.600 | 3.055 |
| 6 | 2.710 | 3.159 |
| 7 | 2.797 | 3.242 |
| 8 | 2.869 | 3.310 |
| 9 | 2.930 | 3.368 |
| 10 | 2.983 | 3.418 |
| 25 | 3.391 | 3.810 |

Extensive tables for $\tau$ can be found in for instance Gibbons, Olkin and Sobel (1977), Table A1, for various values of $P^*$ and $k$.
An important function characterizing the "power" of a selection procedure is the Operating Characteristic curve (OC curve), defined as the $P(CS)$ for the Generalized $LFC$: $\theta_{[1]} = \theta_{[k-1]} = \theta_{[k]} - \delta$, so $P(CS)$ is a function of $\delta$ (besides $\theta_{[k]}$ and $n$).
The following confidence statement can be made after the experiment has been conducted:

$$0 \leq \theta_{[k]} - \theta_s \leq \delta^* ,$$

where $\theta_s$ is the unknown population mean associated with the selected population. This statement can be made provided only $\delta^* n^{\frac{1}{2}} \sigma^{-1} = \tau$.

### Normal populations with unknown $\sigma^2$

Given are $k$ Normal populations with unknown means and unknown equal variances, where

the common variance $\sigma^2$ is considered as a nuisance parameter. The goal is to select the best population (associated with $\theta_{[k]}$). Without information about $\sigma^2$, it can be seen by increasing $\sigma^2$ that there is no common fixed sample size large enough such that the probability requirement will hold for all possible values of $\sigma^2$. For if the true value of $\sigma^2$ is sufficiently large, the $P(CS)$ will be arbitrarily close to $k^{-1}$, which is smaller than any reasonable value of $P^*(k^{-1} < P^* < 1)$.

It is possible to use a good estimate of $\sigma^2$ by pooling the sample variances and use this estimate as an approximation to $\sigma^2$. Bechhofer, Dunnett and Sobel (1954) and Dunnett and Sobel (1954) use a two-stage procedure. The first stage is used to estimate $\sigma^2$. The two stages together are used to reach a final decision. In Gibbons, Olkin and Sobel (1977) tables can be found that make the execution possible.

The case of unequal unknown variances is investigated by Dudewicz and Dalal (1975). Bechhofer (1960) and Bechhofer, Santner and Turnbull (1977) discuss two-factor experiments.

Complete factorial experiments are considered in Dudewicz and Taneja (1980), Lun (1977) and Bechhofer (1977). Selection problems in balanced complete and incomplete block designs are considered by Rasch (1978).

## 8. Final remarks

In 'Design and Analysis of experiments' by Doornbos (1990) one can find a number of general designs and the corresponding analyses.
The possibility of combining nonparametric tests for the analysis of factorial designs is discussed in Van der Laan and Weima (1983).

## Literature


AFSARINEJAD, K. (1983). Balanced repeated measurements designs. *Biometrika* **70**, 199-204.

AGRESTI, A. (1984). *Analysis of Ordinal Categorical Data.* John Wiley & Sons, Inc., New York.

ALTHAM, P.M.E. (1971). The analysis of matched proportions. *Biometrika* **58**, 561-576.

ANSCOMBE, F.J. and TUKEY, J.W. (1963). The examination and analysis of residuals. *Technometrics* **5**, 141-160.

BECHHOFER, R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Statist.* **25**, 16-39.

BECHHOFER, R.E. (1960). A multiplicative model for several factors. *J. Amer. Statist. Assoc.* **55**, 245-264.

BECHHOFER, R.E. (1977). Selection in factorial experiments. *Proc. of the 1977 Winter Simulation Conference* (Eds. H.J. Highland, R.G. Sargent and J.W. Schmidt) held at the National Bureau of Standards, Gaithersburg, Maryland, 65-70.

BECHHOFER, R.E., DUNNETT, C.W. and SOBEL, M. (1954). A two-sample multiple decision procedure for ranking means of normal populations with common unknown variance. *Biometrika* **41**, 170-176.

BECHHOFER, R.E., SANTNER, T.J. and TURNBULL, B.W. (1977). Selecting the largest interaction in a two-factor experiment. *Statist. Decision Theor. and Related Topics - II* (Eds. S.S. Gupta and D.S. Moore), Acad. Press, New York, 1-18.

BENARD, A. and ELTEREN, PH. VAN (1953). A generalization of the method of $m$ rankings. Kon. Nederl. Akad. van Wetensch., *Proceed. Series A* **56**, 358-369.

BERENBLUT, I.I. and WEBB, G.I. (1974). Experimental design in the presence of autocorrelated errors. *Biometrika* **61**, 427-437.

BISHOP, S.H. and JONES, B. (1984). A review of higher-order cross-over designs. *Journal of Applied Statistics* **11**, 29-50.

BOX, G.E.P. (1957). Evolutionary Operation: A Method for Increasing Industrial Productivity. *Applied Statistics* **6**, 81-101.

BOX, G.E.P. and DRAPER, M.R. (1969). *Evolutionary Operation.* John Wiley & Sons, Inc., New York, London.

BOX, G.E.P. and DRAPER, N.R. (1987). *Empirical Model Buidling and Response Surfaces.* John Wiley & Sons, Inc., New York, London.

BOX, G.E.P., HUNTER, W.G. and HUNTER, J.S. (1978). *Statistics for Experimenters.* John Wiley & Sons, New York.

BROEMELING, L.D. (1985). *Bayesian Analysis of Linear Models.* Marcell Dekker, New York.

CHENG, C.S. and WU, C.F. (1980). Balanced repeated measurements designs. *Annals of Statistics* **8**, 1272-1283. Corrigendum (1983) **11**, 349.

COCHRAN, W.G. and COX, G.M. (1957). *Experimental designs.* 2nd ed., John Wiley & Sons, Inc., New York-London.

CONOVER, W.J. and IMAN, R.L. (1981). Rank transformations as a bridge between parametric and non-parametric statistics. *The American Statistician* **35**, 124-129.

CORSTEN, L.C.A. (1984). *Ontwerp en analyse van experimenten (in Dutch).* Vakgroep Wiskunde, Landbouwhogeschool Wageningen.

COX, D.R. (1966). *Planning of Experiments.* John Wiley & Sons, Inc., New York-London-Sydney.

COX, D.R. (1984). Interaction. *International Statistical Review* **52**, 1-31.

DAS, M.N. and GIZI, N.C. (1979). *Design and Analysis of Experiments.* Wiley Eastern Limited, New Delhi, Bangalore, Bombay, Calcutta.

DAVIES, O.L. (ed.) (1963). *Design and analysis of industrial experiments.* 2nd. ed., Oliver and Boyd, Edinburgh, London.

DAVIS, A.W. and HALL, W.B. (1969). Cyclic change-over designs. *Biometrika* **56**, 283-293.

DOORNBOS, R. (1990). *Design and analysis of experiments.* Report Department of Mathematics and Computing Science, Eindhoven University of Technology.

DUDEWICZ, E.J. and DALAL, S.R. (1975). Allocation of observations in ranking and selection with unequal variances. *Sankhya B* **37**, 28-78.

DUDEWICZ, E.J. and KOO, J.O. (1982). The complete categorized quide to statistical selection and ranking procedures. Series in Math. and Management Sciences, Vol. 6, *Amer. Sciences Press*, Inc., Columbus, Ohio.

DUDEWICZ, E.J. and TANEJA, V.S. (1980). Ranking and Selection in designed experiments: complete factorial experiments. *Techn. Report 210*, Ohio State Univ.

DUNNETT, C.W. and SOBEL, M. (1954). A bivariate generalization of Student's $t$ distribution, with tables for certain special cases. *Biometrika* **41**, 153-169.

EVERETT, B.S. (1977). *The Analysis of Contingency Tables.* Chapman and Hall, London.

FEDERER, W.T. (1955). *Experimental Design, Theory and Application.* MacMillan and Co.

FEDEROV, V.V. (1972). *Theory of optimal experiments*, Academic Press, New York and London.

FINNEY, D.J. (1960). *An introduction to: The theory of experimental design.* The University of Chicago Press, Chicago.

FISHER, R.A. (1966). *The Design of Experiments.* 8th ed., Hafner Publishing Company, New York.

FLEISS, J.L. (1986). *The Design and Analysis of Clinical Experiments.* John Wiley & Sons, Inc., New York.

FLETCHER, D.J. (1987). A new class of change-over designs for factorial experiments. *Biometrika* **74**, 649-654.

FLETCHER, D.J. and JOHN, J.A. (1985). Changeover designs and factorial structure. *Journal of the Royal Statistical Society B* **47**, 117-124.

GENTLEMAN, J.F. and WILK, M.B. (1975). Detecting outliers in a two-way table: I. Statistical behavior of residuals. *Technometrics* **17**, 1-14.

GIBBONS, J.D. (1985). *Nonparametric Statistical Inference.* Marcell Dekker, New York.

GIBBONS, J.D. OLKIN, I. and SOBEL, M. (1977). *Selecting and Ordering populations: A new statistical methodology.* John Wiley & Sons, Inc., New York.

GUPTA, S.S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics* **7**, 225-245.

GUPTA, S.S. (Ed.) (1977). Ranking and selection. Special Issue, *Comm. Statist., Theor. Meth.* **A6**.

GUPTA, S.S. and PANCHAPKESAN, S. (1979). *Multiple decision procedures: Theory and methodology of selecting and ranking populations.* John Wiley & Sons, New York-London.

HAALAND, P.D. (1989). *Experimental Design in Biotechnology.* Marcell Dekker, New York.

HEDAYAT, A. and AFSARINEJAD, K. (1975). Repeated measures designs, I. in *A Survey of Statistical Design and Linear Models* (ed. J.N. Srivastava). North-Holland, Amsterdam.

HEDAYAT, A. and AFSARINEJAD, K. (1978). Repeated measures designs, II. *Annals of Statistics* **6**, 619-628.

HERZBERG, A.M. and COX, D.R. (1969). Recent work on the design of experiments: A bibliography and a review. *J. R. Statist. Soc. A* **132**, 29-67.

HICKS, C.R. (1964). *Fundamental concepts in the design of experiments.* Holt, Rinehart and Winston, New York and London.

HOLLANDER, M. and WOLFE, D.A. (1973). *Nonparametric Statistical Methods.* John Wiley & Sons, Inc., New York.

JOHN, J.A. (1987). *Cyclic Designs.* Chapman and Hall, London and New York.

JOHN, P.W.M. (1971). *Statistical Design and Analysis of Experiments.* The MacMillan Company, New York, London.

JONES, B. and KENWARD, M.G. (1990). *Design and Analysis of Cross-Over Trials.* Chapman and Hall, London-New York.

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments.* John Wiley & Sons, Inc., New York-London.

KHURI, A.I. and CORNELL, J.A. (1987). *Response Surfaces: Design and Analyses.* Marcell Dekker, New York.

KISH, L. (1987). *Statistical Design for Research.* John Wiley & Sons, Inc., New York.

KOCH, G.G. (1972). The use of non-parametric methods in the statistical analysis of the two-period change-over design. *Biometrics* **28**, 577-584.

KOCH, G.G., AMARA, I.A., STOKES, M.E. and GILLINGS, D.B. (1980). Some views on parametric and non-parametric analysis for repeated measurements and selected bibliography. *International Statistical Review* **48**, 249-265.

KROON, J.P.M. DE and LAAN, P. VAN DER (1981). Distribution-free test procedures in two-way layouts; a concept of rank-interaction. *Statistica Neerlandica* **35**, 189-213.

KROON, J.P.M. DE and LAAN, P. VAN DER (1983). A generalization of Friedman's rank statistic. *Statistica Neerlandica* **37**, 1-14.

KROON, J.P.M. DE and LAAN, P. VAN DER (1984). A comparison of the powers of a generalized Friedman's rank test and an aligned rank procedure based on simulation. *Statistica Neerlandica* **38**, 189-198.

LAAN, P. VAN DER (1987). Extensive tables with critical values of a distribution-free test for rank-interaction in a two-way layout. *Biuletyn Oceny Odmian* **12**, 195-202.

LAAN, P. VAN DER and PRAKKEN, J. (1972). Exact distribution of Durban's distribution-free test statistic for Balanced Incomplete Block Designs, and comparison with the chi-square and $F$-approximation. *Statistica Neerlandica* **26**, 155-164.

LAAN, P. VAN DER and VERDOOREN, L.V. (1987). Classical analysis of variance methods and nonparametric counterparts. *Biometrical Journal* **29**, 635-665.

LAAN, P. VAN DER and VERDOOREN, L.V. (1989). Selection of populations: An overview and some recent results. *Biometrical Journal* **31**, 383-420.

LAAN, P. VAN DER and WEIMA, J. (1983). Application of the combination of some non-parametric tests for simple factorial designs (in Dutch). *Kwantitatieve Methoden* **10**, 121-138.

LOUIS, T.A. (1988). General methods for analysing repeated measures. *Statistics in Medicine* **7**, 29-45.

LUN, F.W. (1977). Selection in factorial experiments. Project Report, *Statistics 828*, Ohio State Univ.

MATTHEWS, J.N.S. (1988). Recent developments in crossover designs. *International Statistical Review* **56**, 117-127.

MEAD, R. (1988). *The design of experiments: Statistical principles for practical applications.* Cambridge University Press, Cambridge, New York.

MENDENHALL, W. (1968). *Introduction to Linear Models and the Design and Analysis of Experiments.* Wadsworth Publishing Company, Inc., Belmont, California.

MONTGOMERY, D.C. (1984). *Design and Analysis of Experiments.* 2nd ed., John Wiley & Sons, Inc., New York.

MYERS, R.H. (1971). *Response Surface Methodology.* Allyn and Bacon, Inc., Boston.

MYERS, R. (1976). *Response Surface Methodology.* Edwards.

NOETHER, G.E. (1987). Sample size determination for some common nonparametric tests. *J. Amer. Statist. Assoc.* **82**, 645-647.

OTTESTAD, P. (1970). *Statistical Models and their Experimental Application.* Griffin's Statistical Monographs and Courses, London.

PATEL, H.I. (1986). Analysis of repeated measures designs with changing covariates in clinical trials. *Biometrika* **73**, 707-715.

PATTERSON, H.D. (1973). Quenouille's change-over designs. *Biometrika* **60**, 33-45.

PIGEON, J.G. and RAGHAVARAO, D. (1987). Crossover designs for comparing treatments with a control. *Biometrika* **74**, 321-328.

PLACKETT, R.L. (1981). *The Analysis of Categorical Data.* Griffin, London.

POCOCK, S.J. (1983). *Clinical Trials.* John Wiley & Sons, Inc., New York.

QUENOUILLE, M.H. (1953). *The Design and Analysis of Experiment.* Griffin's Statistical Monographs & Courses, London.

RAGHAVARAO, D. (1971). *Construction and Combinatorial Problems in Design of Experiments.* John Wiley & Sons, Inc., New York.

RAGHAVARAO, D. (1989). Crossover designs in industry, in *Design and Analysis of Experiments, With Applications to Engineering and Physical Sciences* (ed. S. Gosh), Marcell Dekker, New York.

RAKTOE, B.L., HEDAYAT, A. and FEDERER, W.T. (1981). *Factorial Designs.* John Wiley & Sons, Inc., New York.

RASCH, D. (1978). Selection problems in Balanced Block Designs. *Biometrical Journal* **20**, 275-278.

RIZVI, M.H. (Ed.) (1985, 1986). Modern Statistical Selection. Part I and II. Proceedings of the Conference "Statistical ranking and selection - Three decades of development". Univ. of California at Santa Barbara, Dec. 1984, *Amer. J. of Math. and Management Sciences*, Vol. 5 (Nos. 3 & 4) and Vol. 6 (Nos. 1 & 2).

SANTNER, T.J. and TAMHANE, A.C. (1984). *Design of Experiments (Ranking and Selection).* Marcell Dekker, New York-Basel.

SCHEFFE, H. (1959). *The Analysis of Variance.* John Wiley & Sons, Inc., London-New York.

SEARLE, S.R. (1971). *Linear Models.* John Wiley & Sons, Inc., New York, London.

SEN, M. and MUKERJEE, R. (1987). Optimal repreated measurements designs under interaction. *Journal of Statistical Planning and Inference* **17**, 18-91.

STEFANSKY, W. (1972). Rejecting outliers in factorial designs. *Technometrics* **14**, 469-479.

TUKEY, J.W. (1951). Quick and dirty methods in Statistics, Part II, Simple analysis for standard designs. Proceedings of the fifth annual Convention. *American Society for Quality Control*, 189-197.

WARE, J.H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician* **39**, 95-101.

List of COSOR-memoranda - 1992

| Number | Month | Author | Title |
|---|---|---|---|
| 92-01 | January | F.W. Steutel | On the addition of log-convex functions and sequences |
| 92-02 | January | P. v.d. Laan | Selection constants for Uniform populations |
| 92-03 | February | E.E.M. v. Berkum<br>H.N. Linssen<br>D.A. Overdijk | Data reduction in statistical inference |
| 92-04 | February | H.J.C. Huijberts<br>H. Nijmeijer | Strong dynamic input-output decoupling:<br>from linearity to nonlinearity |
| 92-05 | March | S.J.L. v. Eijndhoven<br>J.M. Soethoudt | Introduction to a behavioral approach<br>of continuous-time systems |
| 92-06 | April | P.J. Zwietering<br>E.H.L. Aarts<br>J. Wessels | The minimal number of layers of a perceptron that sorts |
| 92-07 | April | F.P.A. Coolen | Maximum Imprecision Related to Intervals of Measures and Bayesian Inference with Conjugate Imprecise Prior Densities |
| 92-08 | May | I.J.B.F. Adan<br>J. Wessels<br>W.H.M. Zijm | A Note on "The effect of varying routing probability in two parallel queues with dynamic routing under a threshold-type scheduling" |
| 92-09 | May | I.J.B.F. Adan<br>G.J.J.A.N. v. Houtum<br>J. v.d. Wal | Upper and lower bounds for the waiting time in the symmetric shortest queue system |
| 92-10 | May | P. v.d. Laan | Subset Selection: Robustness and Imprecise Selection |
| 92-11 | May | R.J.M. Vaessens<br>E.H.L. Aarts<br>J.K. Lenstra | A Local Search Template<br>(Extended Abstract) |
| 92-12 | May | F.P.A. Coolen | Elicitation of Expert Knowledge and Assessment of Imprecise Prior Densities for Lifetime Distributions |
| 92-13 | May | M.A. Peters<br>A.A. Stoorvogel | Mixed $H_2/H_\infty$ Control in a Stochastic Framework |

| Number | Month | Author | Title |
|---|---|---|---|
| 92-14 | June | P.J. Zwietering<br>E.H.L. Aarts<br>J. Wessels | The construction of minimal multi-layered perceptrons: a case study for sorting |
| 92-15 | June | P. van der Laan | Experiments: Design, Parametric and Nonparametric Analysis, and Selection |