Distribution-Free Statistics: An Application-Oriented Approach

JOACHIM KRAUTH

Distribution-Free Statistics: An Application-Oriented Approach

JOACHIM KRAUTH

Psychological Institute, University of Düsseldorf

1988

ELSEVIER

AMSTERDAM · NEW YORK · OXFORD

ISBN Volume: 0-444-80934-1 (Hardback)
0-444-80988-0 (Paperback)
ISSN Series: 0921-0709

Published by:
Elsevier Science Publishers B.V. (Biomedical Division)
P.O. Box 211
1000 AE Amsterdam
The Netherlands

Sole distributors for the U.S.A. and Canada:
Elsevier Science Publishing Company, Inc.
52 Vanderbilt Avenue
New York, NY 10017
U.S.A.

Library of Congress Cataloging in Publication Data

Krauth, Joachim, 1941-
Distribution-free statistics : an application-oriented approach / Joachim Krauth.
p. cm. -- (Techniques in the behavioral and neural sciences ; v. 2)
Includes index.
ISBN 0-444-80934-1
ISBN 0-444-80988-0 (pbk.)
1. Psychometrics. 2. Nonparametric statistics. 3. Statistical hypothesis testing. I. Title. II. Series.
BF39.K689 1987
519.5--dc19

PRINTED IN THE NETHERLANDS

Techniques in the Behavioral and Neural Sciences

Volume 2

Series Editor

JOSEPH P. HUSTON

Düsseldorf

ELSEVIER

AMSTERDAM · NEW YORK · OXFORD

Preface

Even today most researchers in the behavioral and neurosciences use so-called parametric significance tests for evaluating their data. The most commonly used tests of this kind are t-tests for two independent samples, for two paired samples, and for the product moment coefficient. For designs with more than two independent samples or with repeated measurements, F-tests of analysis of variance are used. A glance at any journal in the fields of the behavioral or neurosciences will reveal that most sets of data are analyzed using one of these methods. This seems strange, for most researchers will probably remember from their statistics courses that these procedures will yield correct statistical decisions only if certain very restrictive assumptions are fulfilled.

One of these assumptions is that of univariate or even multivariate normal distributions of the data. Examination of real data reveals that the assumption of a normal distribution is not justified in the majority of cases. Empirical distributions are seldom symmetrical, a necessary assumption for normality. Furthermore, empirical distributions tend to have heavier tails than normal distributions, i.e., observations in the tails have a greater probability of occurrence than is the case for normal distributions.

One argument often used to justify the use of t-tests and F-tests of analysis of variance is that these tests are quite robust with respect to violations of the normality assumption. That is, they will still result in a valid statistical procedure even if the normality assumption is violated. This argument is mainly based on some old studies of robustness such as the Norton study described in Lindquist (1953) or the study of Boneau (1960). Two problems exist with regard to the interpretation of and generalization from the results of such studies. First, it is always the case for such studies that only a few of the infinite number of possible situations can be investigated. Second, robustness depends on the criterion which one uses to define robustness. The same results may be interpreted as a proof of robustness by one researcher and as a disproof

of robustness by another researcher with a stricter criterion. Extensive computer studies have shown that there are many situations for which the arithmetic mean is not a robust estimate of the population mean, and t-tests and F-tests are based on such means. Simple examples which show that situations exist for which t-tests are not robust are given in Krauth (1983).

In addition to the assumption of normality, another assumption of parametric procedures that is often not taken into consideration when the decision is made to use a parametric test is the assumption that the data are at least on an interval-scale level. In the literature one can find many examples of studies for which the assumption of interval-scale data is not met. For example, rating scales are commonly used to measure the type of behavior a patient or a laboratory animal exhibits. The values which are assigned to the different ordered categories of the scale are arbitrarily chosen and could be replaced with different values without altering the meaning of the data. Although the meaning of the data would not change, the result of the statistical test might change if parametric tests such as F-tests or t-tests were used to analyze the data.

If distribution-free tests are used, it is not necessary to make any assumptions about normality. Furthermore, it is possible to use distribution-free tests for data that contain very little numerical information. This means, for example, that for data from rating scales, the order of the measurements is the sole information which is used in the significance tests while the values of the measurements, which were assigned in a more or less arbitrary fashion, have no influence on the statistical results.

One of the earliest and most celebrated books on distribution-free tests is that of Sidney Siegel (1956). Few people using distribution-free methods in the behavioral sciences will not be familiar with Siegel's book. Much of the book's success can be attributed to the fact that it is written in such a comprehensible fashion that even people with a weak statistical background will have little or no difficulty in understanding and correctly applying the tests. This book is so well written that it has not yet been surpassed in popularity.

In a sense Siegel has written the ultimate book on distribution-free tests for the behavioral sciences, and one might ask why anybody would try to write a new book on this subject. The main reason is that since 1956 extensive research has taken place in the field of nonparametric statistics, the results of which naturally could not be considered by Siegel. Thus many studies investigating the small- and large-sample behavior of distribution-free tests have been performed. The efficiency of old and new distribution-free tests has been investigated and compared with that of parametric and distribution-free competitors, and optimality properties have been proven. A second reason is that Siegel's book does not present tests for some commonly occurring experimental situations. For example, tests for censored data, for ordered categories, for interaction in a factorial design, for trend, and for heterogeneity of dependent samples are missing. Of course, it would be nearly impossible to present a distribution-free test for every different experimental situation. An attempt to do so would result in an unwieldy encyclopedia containing many tests for situations which seldom occur in reality. Since

people disagree about which situations have small or large probabilities of occurrence, the selection of the tests which make up such a book will always have a subjective element.

In contrast to Siegel's book, we omitted the one-sample tests and, in particular, the so-called goodness-of-fit tests which are used for testing whether the distribution of a population differs from a theoretically given distribution; for example, a normal distribution. Goodness-of-fit tests are sometimes proposed for use as preliminary tests for parametric tests. If significant departures from normality are found, then parametric tests should not be used. However, if no significant departures from normality are found, it is sometimes argued that one can then use parametric tests. This conclusion is erroneous because it does not take into account the possibility of a type II error. We therefore do not recommend the use of goodness-of-fit tests as preliminary tests.

While omitting one sample tests, we introduced, in contrast to Siegel, tests for dependent samples. In Siegel's book the term 'related samples' is used for two different situations, one for which independent matched samples are considered, and the other for the situation where each subject serves as its own control. In the latter case, two or more scores are obtained from the same subject. These measurements might be dependent and, if so, the multivariate distribution of the measurements for a subject might be asymmetric. If this does occur, all tests described in Siegel (1956) for two or k related samples, with the exception of McNemar's test, can yield significant results, with a high probability, although the distributions for the different populations are identical. This phenomenon is discussed in Section A 2.5 of the present book. Since in general nothing can be assumed about the symmetry of the multivariate distribution of data corresponding to a single subject, the interpretation of test results for dependent samples can be quite misleading. We therefore considered these two situations separately. We proposed tests for independent matched samples which are in some cases identical with those of Siegel (1956) for related samples. To cover the situation where each subject serves as his own control, we proposed tests for dependent samples. These tests were not discussed in Siegel's book.

The present book is divided into four chapters. In Chapter A, short explanations are given for the statistical terms which are used in the following chapters. The use of any important statistical term in Chapters B–D is followed by a reference for the appropriate explanation or definition in Chapter A. It is not imperative that Chapter A be read by someone with some knowledge of statistics who simply wishes to analyze his data. Such a person could use the scheme at the beginning or at the end of the book for selecting the appropriate test and could proceed directly with the test. Chapter A may then be used as a short dictionary of statistical terms. For other users, Chapter A can serve as a short introduction to the area of distribution-free statistics.

Various tests of significance are described in Chapters B, C, and D. The same format is used to describe each of these tests. In the first subsection, 'Purpose of the test', the situation in which the test should be used is explained. In the subsection entitled 'Design', the experimental design which yields the data for the test is described. In

'Assumptions', the assumptions are listed which must be fulfilled in order for the test to be valid. In 'Test problems', the null and alternative hypotheses for one or more of the possible one-sided, two-sided, or other test problems are formulated, and an interpretation is given.

The subsection 'Recommendation' is used to describe the circumstances for which the large-sample procedure could be used instead of the small-sample procedure. We have avoided the use of complicated rules and procedures for making the decision. Our use of these relatively simple rules may lead to recommendations that are too cautious for many situations. However, this can only be determined if one has a criterion for measuring the goodness of the approximation of the small sample by the large-sample procedure.

In the subsection 'Small-sample procedure', the procedure for a small sample is described and an example is given in 'Example for the small-sample procedure'. The small-sample example is followed by the description of the 'Large-sample procedure' and an 'Example for the large-sample procedure'. For those persons not familiar with statistical terminology or for those who experience difficulty in understanding the small- or large-sample procedure, it may be helpful to first study the example before trying to understand the more general directions given in the procedure sections. All examples are based on artificial data, though they were inspired by real studies described in literature. Many ideas for the experimental examples described were obtained from articles in the journal *Biological Psychiatry*. We preferred to use artificial data sets even in those cases where the original data were available. In this way we were able to construct data sets with very small sample sizes which, at the same time, showed all of the features which are normally found only in larger samples. The same artificial data is used for both the large- and small-sample procedure. In this way it is possible to compare the results obtained using the two procedures.

Many users might wonder why we give critical values of the standard normal distribution to 5 decimal places and *p*-values to 6 decimal places throughout the book. This was done for purely pedagogical reasons. In this way somebody checking the examples can be certain that he looks up the correct critical values and *p*-values.

For many examples, especially for those for the tests for three or more samples, the sample sizes are very small. Because the small-sample procedure is in many cases quite cumbersome, requiring the determination of all possible arrangements (permutations) of the data, it was necessary to have small sample sizes in order to keep the calculations within reasonable limits. As a consequence of the small sample size, the results for both the small- and large-sample procedures are usually not significant since the same data sets are used for both procedures. Our examples illustrate that it is possible to perform a statistical test, even with only one or two subjects in a group. Such small sample sizes might be particularly useful if pilot studies are run and one wishes to determine whether a larger study would be worthwhile. Studies with larger sample sizes, while requiring a large investment of time and money, have the advantage that effects will be more readily detected and that the data can be analyzed with the less tedious large-sample procedures.

In the last subsection, 'Remarks', the origin of the test and its relationship to other tests is explained. Sometimes this section also includes a justification for the choice of this particular test rather than one of the other possible tests for this category.

In Chapter B, two-sample tests of heterogeneity are described. These are classified with respect to the specific design (two independent, two dependent, or two independent matched samples) and the scale of measurement (at least nominal, ordinal, or interval scale). In addition to these nine tests, two modifications of Wilcoxon's rank-sum test (for ordered categories and for censored data) are considered as well as two fourfold-table tests for independent and dependent samples.

In Chapter C, tests of the dependence of two samples are described for nominal-, ordinal-, and interval-scale data. In addition to these three tests, two modifications of Spearman's rank correlation test (for ordered categories and for censored data) are considered as well as a fourfold-table test for nominal-scale data.

In Chapter D, tests of heterogeneity for three or more samples are described. These tests are classified with respect to the specific design and with respect to the scale of measurement in much the same way as the two-sample tests in Chapter B. In addition to these nine tests, four other tests for independent samples of ordinal-scaled measurements are considered. These include two modifications of the Kruskal-Wallis test (for ordered categories and for censored data), a modification of Spearman's rank correlation test for a test for trend, and a test for interaction in a 2×2 factorial design.

One might ask why tests for ordered categories or for trend were not also considered for tests of heterogeneity for dependent or independent matched samples, or why tests for censored data or for interaction were not also considered for interval-scale data. Although it would have been possible to include these tests, this would have increased the size of the book considerably though the likelihood that these experimental situations would ever arise is, in our opinion, quite small.

For each possible situation only one test was proposed. One criterion we used for selecting a particular test was its efficiency in comparison with its parametric and nonparametric competitors. However, a second criterion was sometimes considered to be even more important. We tried to present only procedures which could be performed with the aid of a pocket calculator and which could be easily explained. No exact small-sample tests exist for the case of tests of heterogeneity for two or more dependent samples. Therefore we proposed the use of very conservative small-sample tests based on the random selection of data from the set of observed data. For some of these situations, namely for nominal-scale data, quite efficient large-sample tests are described in the literature. However, we have proposed the use of large-sample tests which, though less efficient, can be performed without complicated matrix inversions and can be more easily explained.

Some people might feel that the book is too redundant in many respects. For example, the steps used to describe the small-sample and large-sample procedures, the designs, assumptions, and test problems are nearly identical for many tests. Furthermore, some tests which are given for different situations are either equivalent to each

other or are special cases of a particular test. Thus, for example, the rank correlation test for censored data could replace Spearman's rank correlation test, the contingency-table test for ordered categories for both variables, and the rank test for trend. The last three tests are all equivalent. Other examples of tests which can replace several other tests are Schemper's test for censored data, Fisher's contingency-table test, and Wall's test. However, we thought that it would be easier for the potential user to find and perform the appropriate procedure and to interpret the results if we organized the book with respect to the different possible situations and if we described a separate test for each situation. Otherwise a large number of possible applications and interpretations would have been necessary for some tests, leading to possible confusion.

In contrast to Siegel (1956) and other authors, we did not consider specific corrections for ties in rank tests. All rank tests in this book can be used on tied observations (using midranks) without further modifications of the test statistics.

Last but not least I should mention that this book, as is true of most other books, could not have been written without the help of some other people. I wish to thank Helmut Quentmeier for his help in the computation of the tables in Sections II and III of the Appendix and Rolf Diehl for his help in the computation of all other tables in the Appendix. I would like to thank Jennifer Nagel, who not only corrected my English, but also pointed out the many places in the text which needed rewriting because they were incomprehensible or misleading. Finally I must thank Karin Boden, Karin Boucke, Monika Gretzke and particularly Ilse Marie Mahr who typed several versions of the manuscript in a short time and with admirable precision. While I am deeply indebted to all these persons for their help, it should be clear that the responsibility for any errors in the book is mine alone.

Joachim Krauth

Contents

A Explanation of statistical terms

A 1 *Design*

A 1.1 *One-sample, two-sample, r-sample, and factorial designs*

One-sample design. For a **one-sample design**, a sample of n measurements, which may be univariate or multivariate (A 1.3), is drawn. Depending on the problem, a sample of subjects or a sample of measurements which is obtained from the subjects, is considered.

Two-sample design. An **independent variable**, i.e., a variable which can be manipulated by the researcher, may sometimes have only two levels, for example a treatment T and a control C, or a treatment A and a treatment B, or a time t_1 and a time t_2. In a **two-sample design**, a sample of measurements is drawn for each of these two levels. The two sample sizes (n_1 and n_2) may be equal or may differ in size.

r-sample design. An independent variable which can be manipulated by the researcher may attain r different levels, where r is greater than one. Examples of the r different levels are three treatment groups (T_1, T_2, and T_3) and two control groups (C_1 and C_2), i.e., $r = 5$, or four treatment groups (T_1, T_2, T_3, and T_4), i.e., $r = 4$, or three points in time (t_1, t_2, and t_3), i.e., $r = 3$. In an ***r*-sample design**, a sample of measurements is drawn for each level of the independent variable. The r sample sizes ($n_1, \ldots, n_r$) may be equal or may differ in size. When $r = 2$, the special case of the two-sample design results.

Factorial design. The number of independent variables which may be manipulated by the researcher at the same time can be designated as s. The s independent variables are also called **factors**. The first variable or factor may have r_1 levels, the second variable, r_2 levels, etc., and the last variable, r_s levels. The number of levels ($r_1, \ldots, r_s$) for

each factor should be greater than one. Consider an example of a design with $s = 4$ and $r_1 = 2$, $r_2 = 4$, $r_3 = 3$, and $r_4 = 6$. Such a design results if one considers the variables "drug" with the levels medication and placebo ($r_1 = 2$), "temperature" with the levels 15 °C, 20 °C, 25 °C and 30 °C ($r_2 = 4$), "equipment" with the levels apparatus 1, 2, and 3 ($r_3 = 3$), and "time" with the levels 10 min, 20 min, 30 min, 40 min, 50 min, and 60 min after application of the drug ($r_4 = 6$).

In a **factorial design**, a sample of measurements is drawn for each of the $r_1 \cdot r_2 \cdot \ldots \cdot r_s$ possible combinations of levels. For the example given above this would be $2 \cdot 4 \cdot 3 \cdot 6 = 144$ samples. The sample sizes should be greater than zero and they may all be equal or may differ in size. The factorial design with $s = 2$ factors, each with $r_1 = r_2 = 2$ levels, is of particular interest because the results of distribution-free tests for this design can be more easily interpreted than those for more complex designs. The factorial design with $s = 1$ factor is equivalent to the r-sample design and the design with $s = 1$ and $r_1 = 2$ is equivalent to the two-sample design.

A 1.2 *Independent, dependent, and independent matched samples*

Independent samples. Consider a two-sample design (A 1.1) with two samples of measurements. The two samples are called **independent** if all measurements are independent in the larger pooled sample. A measurement X is said to be independent of one or more other measurements ($U, V, W, \ldots$) if no predictions about the value of X can be made if the values of $U, V, W, \ldots$ are known (A 4.3). If one measurement is obtained from each subject, and if there is no way in which the subjects can influence each other or inform each other as to the nature of the experiment, we can assume independence. However, if several measurements are obtained from the same subject with respect to one variable at several points in time, or with respect to several variables at one or more points in time, independence of the measurements cannot be assumed. The same is true if the subjects influence each other. Measurements which are not independent are called **dependent**.

Similarly, for r-sample designs or in factorial designs, samples are said to be independent if all measurements in the pooled sample are independent.

Dependent samples. In general, dependent samples are samples which are not independent. Here only the following special case is considered. Assume that a sample of n subjects is drawn and that for each subject a fixed number (r) of measurements is obtained, where r is greater than one. The r measurements for a single subject may correspond to the measurement of a variable at r points in time or to the measurement of several variables at one or more points in time. If the subjects do not influence each other, the measurements for different subjects can be assumed to be independent. However, it must be assumed that the r measurements obtained from a single subject are dependent. In this case we can speak of **r dependent samples**. If $r = 2$, the design with two dependent samples results. A typical example of this type of design is the pre–post treatment design where measurements taken before and after the treatment are considered for each subject.

Instead of r dependent samples, one might speak of a one-sample design (A 1.1) with multivariate or r-variate observations (A 1.3).

Independent matched samples. Sometimes subjects are combined into n groups (**blocks**) of the same size r (with r greater than one) for the express purpose of producing high homogeneity within the blocks. To this end, subjects who are as similar as possible to each other with respect to one or more **matching variables** are combined into a block. Such matching variables might be, for example, age, intelligence or diagnosis, or for animals, weight or membership in the same litter. The r levels of the independent variable (A 1.1) are assigned randomly to the r subjects of a block. Assuming that the $r \cdot n$ subjects do not in any way influence each other, an r-sample design (A 1.1) results, with an additional structure imposed by matching. The term **randomized block design** is often used to describe a design with independent matched samples. If $r = 2$, one speaks of **independent matched pairs**.

A 1.3 *Univariate and multivariate observations*

Univariate observations. If only one measurement or score is obtained from each subject, then these observations are called **univariate observations.**

Multivariate observations. If more than one measurement is obtained from each subject, then these measurements are called **multivariate observations**. The univariate measurements which are the components of such multivariate observations are generally assumed to be dependent. In most cases, it is assumed that for each subject the multivariate observation consists of exactly r measurements, where r (the number of levels of the independent variable) is larger than or equal to two. When $r = 2$, the term **bivariate observations** is used. The design with r dependent samples (A 1.2) corresponds to a one-sample design (A 1.1) with **r-variate** observations.

One can also consider designs with independent samples (A 1.2) where the measurement for each subject is a multivariate observation. Such designs can be evaluated by multivariate statistical procedures.

A 1.4 *Randomization*

The term **randomization** in experimental designs can mean either the random selection of the total sample from the population under consideration, or the random assignment of this total sample to the different levels or combinations of levels of the independent variables, i.e., the different experimental conditions. Both kinds of randomization are achieved simultaneously if a random sample is drawn separately from the population for each experimental condition.

In most situations, the experimenter has access to only a part of the target population. This segment of the population may be, for example, the patients in the psychiatric ward of a certain hospital, the rats in a shipment obtained from a certain supplier, or the students in an introductory psychology course. Because the target population is not