INTRODUCTION TO NONPARAMETRIC STATISTICAL METHODS

C. A. HESSE, BSc, MPhil, PhD.
Senior Lecturer of Statistics,
Methodist University College Ghana.

J. B. OFOSU, BSc, PhD, FSS.
Professor of Statistics,
Methodist University College Ghana.

E. N. NORTEY, BSc, MPhil, PhD.
Senior Lecturer of Statistics,
University of Ghana.

Published and Printed by

AKRONG PUBLICATIONS LIMITED
P. O. BOX M. 31
ACCRA, GHANA

(0244 648 757, 0264 648 757)

ISBN: 978–9988–2–6059–0

Published, 2017

akrongh@yahoo.com.

PREFACE

A statistical method is called ***non-parametric*** if it makes no assumption on the population distribution or sample size. This is in contrast with most parametric methods in elementary statistics that assume that the data set used is quantitative, the population has a normal distribution and the sample size is sufficiently large. In general, conclusions drawn from non-parametric methods are not as powerful as the parametric ones. However, as non-parametric methods make fewer assumptions, they are more flexible, more robust, and applicable to non-quantitative data.

This book is designed for students to acquire basic skills needed for solving real life problems where data meet minimal assumption and secondly to beef up their reading list as well as provide them with a “one shop stop” textbook on Nonparametric.

Our Approach

This book is an introduction to basic ideas and techniques of nonparametric statistical methods and is intended to prepare students of the sciences as well as the humanities, for a better understanding of some underlying explanations of real life situations. Researchers will find the text useful since it provides a step-by-step presentation of procedures, use of more practical data sets, and new problems from real-life situations. The book continues to emphasize the importance of nonparametric methods as a significant branch of modern statistics and equips readers with the conceptual and technical skills necessary to select and apply the appropriate procedures for any given situation.

Written by leading statisticians, ***Introduction to Nonparametric Statistical Methods***, provides readers with crucial nonparametric techniques in a variety of settings, emphasizing the assumptions underlying the methods. The book provides an extensive array of examples that clearly illustrate how to use nonparametric approaches for handling one- or two-sample location and dispersion problems, dichotomous data, one-way analysis of variance, rank tests, goodness-of-fit tests and tests of randomness.

A wide range of topics is covered in this text although the treatment is limited to the elementary level. There are solved, partly solved and unsolved assignments with every section, to make the student or reader familiar with the methods introduced.

C. A. Hesse
J. B. Ofosu
E. N. Nortey
July, 2017

CONTENTS

Chapter One
Preliminaries

1.1 Introduction

The typical introductory courses in hypothesis-testing and confidence interval examine primarily parametric statistical procedures. A main feature of these statistical procedures is the assumption that we are working with random samples from normal populations. These procedures are known as ***parametric methods*** because they are based on a particular parametric family of distributions – in this case, the normal. For example, given a set of independent observations from a normal distribution, we often want to infer something about the unknown parameters. Here the *t*-test is usually used to determine whether or not the hypothesized value μ_0 for the population mean should be rejected or not. More usefully, we may construct a confidence interval for the 'true' population mean.

Parametric inference is sometimes inappropriate or even impossible. To assume that samples come from any specified family of distributions may be unreasonable. For example, we may not have examination marks for each candidate but know only the numbers of candidates who obtained the ordered grades *A*, *B*+, *B*, *B*–, *C*+, *C*, *D* and *F*. Given these grade distributions for two different courses, we may want to know if they indicate a difference in performance between the two courses. In this case it is inappropriate to use the traditional (parametric) method of analysis.

In this book we describe procedures called ***nonparametric*** and ***distribution-free methods.*** Nonparametric methods provide an alternative series of statistical methods that require no or very limited assumptions to be made about the data. These methods are most often used to analyse data which do not meet the distributional requirements of parametric methods. In particular, skewed data are frequently analysed by non-parametric methods, although data transformation can sometimes make the data suitable for parametric analyses. These procedures have considerable appeal. One of their advantages is that the data need not be quantitative but can be categorical (such as yes or no) or rank data.

Generally, if both parametric and nonparametric methods are applicable to a particular problem, we should use the more efficient parametric method.

1.2 Parametric and nonparametric methods

The word statistics has several meanings. It is used to describe a collection of data and also to designate operations that may be performed with primary data. The scientific discipline called statistical inference uses observed data – in this context called a sample – to make inference about a larger observable collection of data called a population. We associate distributions with populations. For example, if the random variable which describes a population is $N(\mu,\ \sigma^2)$, then we say that the population is $N(\mu,\ \sigma^2)$.

Parametric methods are often those for which we know that the population is normal, or we can approximate using a normal distribution after we invoke the central limit theorem. Ultimately the classification of a method as parametric depends upon the assumptions that are made about a population. A few parametric methods include the testing of a statistical hypothesis about a population mean under two different conditions:

1. when sampling is from a normally distributed population with known variance,
2. when sampling is from a normally distributed population with unknown variance.

The nonparametric methods, however, are not based on the underlying assumptions and thus do not require a population's distribution to be denoted by specific parameters.

1.3 Parametric versus nonparametric methods

The analysis of data often begins by considering the appropriateness of the normal distribution as a model for describing the distribution of the population. If this distribution is reasonable, or if the normal approximation is deemed adequate, then the analysis will be carried out using normal-theory methods. If the normal distribution is not appropriate, it is common to consider the possibility of a transformation of the data. For instance, a simple transformation of the form $Y = \log(X)$ may yield data that are normally distributed, so that normal-theory methods may be applied to the transformed data.

If neither of these approaches seems reasonable, there are two ways to proceed. It may be possible to identify the type of distribution that is appropriate – say, exponential – and then use the methods that specifically apply to that distribution. However, there may not be sufficient data to ascertain the form of the distribution, or the data may come from a distribution for which methods are not readily available. In such situations one hopes not to make untenable assumptions, and this is where nonparametric methods come into play.

Nonparametric methods require minimal assumptions about the form of the distribution of the population. For instance, it might be assumed that the data are from a population that has continuous distribution, but no other assumptions are made. Or it might be assumed that the population distribution depends on location and scale parameters, but the functional form of the distribution, whether normal or whatever, is not specified. By contrast, parametric methods require that the form of the population distribution be completely specified except for finite number of parameters. For instance, the familiar one-sample t-test for means assumes that observations are selected from a population that has a normal distribution, and the only values not known are the population mean and standard deviation. The simplicity of nonparametric methods, the widespread availability of such methods in statistical packages, and the desirable statistical properties of such methods make them attractive additions to the data analyst's tool kit.

1.4 Classes of nonparametric methods

Nonparametric methods may be classified according to their function, such as two-sample tests, tests for trends, and so on. This is generally how this book is organized. However, methods may also be classified according to the statistical ideas upon which they are based. Here, we consider the ideas that underlie the methods discussed in this book.

The typical introductory course in statistics examines primary parametric statistical procedures. Recall that these procedures include tests based on the Student's *t*-distribution, analysis of variance, correlation analysis and regression analysis. A characteristic of these procedures is the fact that the appropriateness of their use for the purpose of inference depends on certain assumptions. Inferential procedures in analysis of variance, for example, assume that samples have been drawn from normally distributed populations with equal variances.

Since populations do not always meet the assumptions underlying parametric tests, we frequently need inferential procedures whose validity do not depend on rigid assumptions. Nonparametric statistical procedures fill this need in many instances, since they are valid under very general assumptions. As we shall discuss more fully later, nonparametric procedures also satisfy other needs of the researcher.

By convention, two types of statistical procedures are treated as nonparametric: (1) truly nonparametric procedures and (2) distribution-free procedures. Strictly speaking, nonparametric procedures are not concerned with population parameters. For example, in this book we shall discuss tests for randomness where we are concerned with some characteristic other than the value of a population parameter. The validity of distribution-free procedures does not depend on the functional form of the population from which the sample has been drawn. It is customary to refer to both types of procedure as nonparametric. Kendal and Sundrum (1953) discussed the differences between the terms nonparametric and distribution-free.

1.5 When to use nonparametric procedures

The following are some situations in which the use of a nonparametric procedure is appropriate.

1. The hypothesis to be tested does not involve a population parameter.
2. The data have been measured on a scale weaker than that required for the parametric procedure that would otherwise be employed. For example, the data may consist of count data or rank data, thereby precluding the use of some otherwise appropriate parametric procedure.
3. The assumptions necessary for the valid use of a parametric procedure are not met. In many instances, the design of a research project may suggest a certain parametric procedure. Examination of the data, however, may reveal that one or more assumptions underlying the test are grossly violated. In that case, a nonparametric procedure is frequently the only alternative.
4. Results are needed in a hurry and calculations must be done by hand.

The literature in nonparametric statistics is extensive. A bibliography by Savage (1962) contained some 3 000 entries. An up-to-date bibliography would undoubtedly contain many times that number.

1.6 Advantages of nonparametric statistics

The following are some of the advantages of the available nonparametric statistical procedures.

1. **Make fewer assumptions.**
   Nonparametric Statistical Procedures are procedures that generally do not need rigid parametric assumptions with regards to the populations from which the data are taken.

2. **Wider scope.**
   Since there are fewer assumptions that are made about the sample being studied, nonparametric statistics are usually wider in scope as compared to parametric statistics that actually assume a distribution.

3. **Need not involve population parameters.**
   Parametric tests involve specific probability distributions (e.g., the normal distribution) and the tests involve estimation of the key parameters of that distribution (e.g., the mean or difference in means) from the sample data. However, nonparametric tests need not involve population parameters.

4. **The chance of their being improperly used is small.**
   Since most nonparametric procedures depend on a minimum set of assumptions, the chance of their being improperly used is small.

5. **Applicable even when data is measured on a weak measurement scale.**
   For interval or ratio data, you may use a parametric test *depending* on the shape of the distribution. Non-parametric test can be performed even when you are working with data that is nominal or ordinal.

6. **Easy to understand.**
   Researchers with minimum preparation in Mathematics and Statistics usually find nonparametric procedures easy to understand.

7. **Computations can quickly and easily be performed.**
   Nonparametric tests usually can be performed quickly and easily without automated instruments (calculators and computers). They are designed for small numbers of data, including counts, classifications and ratings.

1.7 Disadvantages of nonparametric tests

Nonparametric procedures are not without disadvantages. The following are some of the more important disadvantages.

1. **May Waste Information.**
   The researcher may waste information when parametric procedures are more appropriate to use. If the assumptions of the parametric methods can be met, it is generally more efficient to use them.

2. **Difficult to compute by hand for large samples.**
   For large sample sizes, data manipulations tend to become more laborious, unless computer software is available.

3. **Tables not widely available.**
   Often special tables of critical values are needed for the test statistic, and these values cannot always be generated by computer software. On the other hand, the critical values for the parametric tests are readily available and generally easy to incorporate in computer programs

1.8 The scope of this book

The emphasis in this book is on the application of nonparametric statistical methods. Wherever available, the examples and exercises use real data, gleaned primary from the results of research published in various journals. We hope that the use of real situations and real data will make the book more interesting to you. We have included problems from a wide variety of statistical techniques described. We have included, also, a wide variety of statistical techniques. The techniques we discuss are those most likely to prove helpful to the researcher and most likely to appear in the research literature. In this text we have covered not only hypothesis testing, but interval estimation as well.

1.9 Format and organization

In presenting these statistical procedures, we have adopted a format designed to make it easy for you to use the book. Each hypothesis-testing procedure is broken down into four components: (1) assumptions, (2) hypothesis, (3) test statistics, and (4) decision rule.

Thus, for a given test, you can quickly determine the assumptions on which the test is based, the hypotheses that are appropriate, how to compute the test statistic, and how to determine whether to reject the null hypothesis. First, we discuss these topics in general, and then we use an example to illustrate the application of the test.

Where appropriate for a given test, we discuss ties, the large-sample approximation, and the power efficiency. For each procedure, we cite references that you may consult if you are interested in learning more about the procedure or in further pursuing a related topic. Finally we provide exercises for each procedure. These exercises serve two purposes: They illustrate appropriate uses of a test, and they give you a chance to determine whether you have mastered the computational techniques, and learnt how to set the hypotheses and use the applicable decision rule.

In the remaining chapters, we cite two types of reference: those that are cited in the body of the text and refer you to the statistical literature, and those that are cited in the examples and exercises and refer you to the research literature.

References

Armitage, P. (1971). Statistical Methods in Medical Research, Oxford and Edinburgh: Blackwell Scientific Publications.

Colton, T. (1974). Statistics in Medicine, Boston: Little Brown.

Dunn, Olive J., (1964). Basic Statistics: A Primer for the Biomedical Sciences, New York: Wiley.

Kendall, M. G. and Sundrum (1953). Distribution-Free Methods and Order Properties. Rev. Int. Statist. Inst. 21, 124 – 134.

Savage, I. R. (1962). Bibliography on Nonparametric Statistics. Harvard University Press.

Remington, R. D. and Schork, M. A. (1970). Statistics with Applications to the Biological and Health Sciences, Englewood Cliffs, N.J.: Prentice-Hall.

Chapter Two

One-Sample Nonparametric Methods

2.1 Introduction

In classical parametric tests (which assume that the population from which the sample data have been drawn is normally distributed), the parameter of interest is the population mean. In this chapter, we shall be concerned with the nonparametric analog of the one-sample z and t tests. These are nonparametric procedures (which utilize data consisting of a single set of observations) that are appropriate when the location parameter is the median, rather than the mean.

Several nonparametric procedures are available for making inferences about the median. Two of the nonparametric tests which are useful in situations where the conditions for the parametric z and t tests are not met, are the one-sample sign test and the Wilcoxon signed-ranks test.

Recall that the median of a set of data is defined as the ***middle value when data are arranged in order of magnitude***. For continuous distributions, we define the median as the point $\tilde{\mu}$ for which the probability that a value selected at random from the distribution is less than $\tilde{\mu}$, and the probability that a value selected at random from the distribution is greater than $\tilde{\mu}$, are both equal to $1/2$. When the population from which the sample has been drawn is symmetric, any conclusions about the median are applicable to the mean, since in symmetrical distributions the mean and the median coincide.

In this chapter, we shall also discuss procedures for making inferences concerning the population proportion and testing for randomness and the presence of trend.

Wherever possible, we shall observe the following format in presenting the hypothesis-testing procedures.

1. **Assumptions**
   We list the assumptions necessary for the validity of the test, and describe the data on which the calculations are based.
2. **Parameter of interest**
   From the problem context, we identify the parameter of interest.

3. **Hypotheses**
   We state the null hypothesis H_0 and the alternative hypothesis H_1.

4. **Test statistic**
   We write down a formula or direction for computing the relevant test statistic. When we give a formula, we describe the methodology for evaluating it.

5. **Significance level**
   We choose a significance level α.

6. **Decision rule**
   We determine the critical region. The Appendix gives appropriate tables for the distribution of the test statistic. From these tables, we can determine the critical values of the test statistic corresponding to the chosen α.

7. **Value of the test statistic**
   We compute the value of the test statistic from the sample data.

8. **Decision**
   If the computed value of the test statistic is as extreme as or more extreme than a critical value, we reject H_0 and conclude that H_1 is true. If we cannot reject H_0, we conclude that ***there is not enough information to warrant its falsity***.

2.2 The one-sample sign test

The sign test is perhaps the oldest of all nonparametric procedures. Let $X_1, X_2, \ldots, X_n$ be an observed random sample of size *n* from a population with median $\tilde{\mu}$. The sign test utilizes only the signs of the differences between the observed values X_i and the hypothesized median $\tilde{\mu}_0$. Thus, the data is converted into a series of plus (+) and minus (–) signs.

2.2.1 Assumptions

1. The sample available for analysis is a random sample of independent measurements from a population with an unknown median $\tilde{\mu}$.
2. The variable of interest is measured on at least an ordinal scale.
3. The variable of interest is continuous.

2.2.2 Hypotheses

The hypothesis to be tested concerns the value of the population median. To test the hypothesis

$$H_0\colon \tilde{\mu} = \tilde{\mu}_0$$

where $\tilde{\mu}_0$ is a specified median value, against a corresponding one-sided or two-sided alternative, we use the ***Sign Test***. The test statistic S depends on the alternative hypothesis, H_1.

(a) **One-sided test**

For a one sided test, the alternative hypothesis is either $H_1\colon \tilde{\mu} < \tilde{\mu}_0$ or $H_1\colon \tilde{\mu} > \tilde{\mu}_0$.

(i) If we wish to test

$$H_0\colon \tilde{\mu} = \tilde{\mu}_0 \text{ against}$$
$$H_1\colon \tilde{\mu} < \tilde{\mu}_0,$$

then the ***test statistic*** is defined by

$$S = N^+,$$

where N^+ = Number of observations X_i greater than $\tilde{\mu}_0$

= Number of +signs when the differences $X_i - \tilde{\mu}_0$ are computed, $i = 1, 2, \ldots n$.

If the alternative hypothesis is true, then we should expect $X_i - \tilde{\mu}_0$ to yield significantly fewer positive (+) signs than negative (−) signs. Thus, a smaller number of (+) signs leads to the rejection of H_0. When H_0 is true, we expect the number of (−) signs to be equal to that of the (+) signs and hence

$$P(S < \tilde{\mu}_0) = P(S > \tilde{\mu}_0) = \tfrac{1}{2}.$$

Thus, when H_0 is true, S has the binomial distribution with parameters n and $\frac{1}{2}$. That is,

$$S \sim b\left(n,\ \tfrac{1}{2}\right).$$

Decision rule

The p-value of the test is defined by

$$p = P\left(S \le s_o \middle| H_0 \text{ is true}\right),$$

where s_o is the observed value of the test statistic S. We reject H_0 at significance level α if $p \le \alpha$.

(ii) For a ***one-sided test***, we test

H_0: $\tilde{\mu} = \tilde{\mu}_0$ against

H_1: $\tilde{\mu} > \tilde{\mu}_0$

The ***test statistic*** is

$$S = N^-$$

where N^- = Number of observations ***less than*** $\tilde{\mu}_0$

= Number of –signs when the differences $X_i - \tilde{\mu}_0$ are computed, $i = 1, 2, \ldots n$.

If the alternative hypothesis is true, then we should expect $X_i - \tilde{\mu}_0$ to yield less negative (−) signs than would be expected if the null hypothesis were true. Likewise, when H_0 is **true**, S has the binomial distribution with parameters n and $\frac{1}{2}$. That is,

$$S \sim b\left(n, \tfrac{1}{2}\right).$$

Decision rule

The p-value of the test is defined by

$$p = P\left(S \le s_0 \middle| H_0 \text{ is true}\right),$$

where s_o is the observed value of $S = N^-$. We reject H_0 at significance level α if $p \le \alpha$.

(b) **Two-sided test**

If we wish to test

H_0: $\tilde{\mu} = \tilde{\mu}_0$ against

H_1: $\tilde{\mu} \ne \tilde{\mu}_0$,

then the ***test statistic*** is defined by

$$S = \min\left\{N^-, N^+\right\},$$

where N^- is the number of –signs and N^+ is the number of +signs when the differences $X_i - \tilde{\mu}_0$ are computed.

We should reject the null hypothesis if we have too few negative (–) signs or too few positive (+) signs. When H_0 is **true**, S has the binomial distribution with parameters n and $\frac{1}{2}$.

Decision Rule

The *p*-value of the test is defined by

$$p = 2P\left(S \le s_0 \middle| H_0 \text{ is true}\right),$$

where s_0 is the observed value of the test statistic S. We reject H_0 at significance level α if $p \le \alpha$.

Problem with zero differences

- We assume that the variable of interest is continuous. Therefore, in theory, no zero differences should occur when we compute $x_i - \tilde{\mu}_0$.
- In practice, however, zero differences do occur. The usual procedure is to discard observations leading to zero differences and reduce *n* accordingly. In that case the hypothesis may be re-stated in probability terms. For example, a two-sided case will have its null hypothesis as

$$P\left(X < \tilde{\mu}_0\right) = P\left(X > \tilde{\mu}_0\right) = 0.5.$$

Example 2.1

Appearance transit times for 11 patients with significantly occluded right coronary arteries are given below:

Subject	1	2	3	4	5	6	7	8	9	10	11
Transit time (in sec)	1.80	3.30	5.65	2.25	2.50	3.50	2.25	3.10	2.70	2.70	3.00

Can we conclude, at the 0.05 level of significance, that the median appearance transit time in the population from which the data were drawn, is different from 3.50 seconds?

Solution

The parameter of interest is $\tilde{\mu}$, the median appearance transit time in the population. We wish to test the hypothesis

$$H_0: \tilde{\mu} = 3.50 \quad \text{against}$$

$$H_1: \tilde{\mu} \ne 3.50,$$

at the $\alpha = 0.05$ level of significance. Since this is a two-sided test, the test statistic is

$$S = \min\left\{N^-,\ N^+\right\},$$

where N^- is the number of observations less than 3.50 and N^+ is the number of observations greater than 3.50. When H_0 is true, $S \sim b(10,\ 0.5)$.

Note: We discard one observation which has the same value as the hypothesized median, leaving us with a usable sample size of 10.

Let s_0 be the observed value of the test statistic. We reject H_0 at the 0.05 level of significance when $p \le 0.05$, where

$$p = 2P\left(S \le s_0 \middle| 10, 0.5\right).$$

	1	2	3	4	5	6	7	8	9	10
X_i	1.80	3.30	5.65	2.25	2.50	2.25	3.10	2.70	2.70	3.00
Sign of $X_i - 3.50$	–	–	+	–	–	–	–	–	–	–

From the above table, $N^- = 9$ and $N^+ = 1$. The observed value of the test statistic is therefore given by

$$s_0 = \min\{9, 1\} = 1.$$

Since this is a two-sided test, the *p*-value of the test is given by

$$p = 2P\left(S \le 1 \middle| 10, 0.5\right) = 2 \times 0.0107 = 0.0214.$$

Since the *p*-value of the test, 0.0214, is less than 0.05, we reject H_0 at the 0.05 level of significance and conclude that the population median is not 3.50.

Example 2.2

The following data are IQs of arrested drug abusers who are aged 16 years or older. Is there any evidence that the median IQ of drug abusers in the population is greater than 107? Use $\alpha = 0.05$.

99	100	90	94	135	108	107	111	119	104	127	109	117	105	125

Solution

The parameter of interest is $\tilde{\mu}$, the median IQ of drug abusers in the population. We wish to test the hypothesis

$$H_0: \tilde{\mu} = 107 \quad \text{against}$$
$$H_1: \tilde{\mu} > 107,$$

at the $\alpha = 0.05$ level of significance. The test statistic is

$$S = N^-$$

where N^- is the number of observations less than 107. When H_0 is true, $S \sim b(14, 0.5)$.

Note: We discard one observation which has the same value as the hypothesized median, leaving us with a usable sample size of 14.

Let s_0 be the observed value of the test statistic. We reject H_0 at the 0.05 level of significance when $p \le \alpha = 0.05,$ where the p-value of the test is given by

$$p = P\left(S \le s_0 \middle| 14, 0.5\right).$$

The following table gives the signs of $X_i - 107.$

1	2	3	4	5	6	7	8	9	10	11	12	13	14
99	100	90	94	135	108	111	119	104	127	109	117	105	125
–	–	–	–	+	+	+	+	–	+	+	+	–	+

Here, $N^- = 6$ and $N^+ = 8.$ The observed value of the test statistic is $\min(6,\ 8) = 6.$ Thus,

$$s_0 = 6.$$

Since this is a one-sided test, the *p*-value of the test is given by

$$p = P\left(S \le 6 \middle| 14, 0.5\right) = 0.3953.$$

Since the *p*-value of the test, 0.3953, is greater than 0.05, we fail to reject H_0 at the 0.05 level of significance. Hence, there is not enough evidence to conclude that the median IQ of the subjects in the population is greater than 107.

2.2.3 Large sample approximation

If the sample size is larger than 15, we can use the normal approximation to the binomial distribution with a continuity correction. Thus, if n is large and $S \sim b\left(n,\ \frac{1}{2}\right),$ then it can be shown that S is approximately normally distribution with mean np and variance $np(1-p).$ That is, $S \sim N\left(np,\ np(1-p)\right).$ Thus, for the sign test, when $p = \frac{1}{2}$ and $n > 15$, we can use the test statistic

$$Z = \frac{S - \frac{1}{2}n}{\sqrt{n \times \frac{1}{2} \times \frac{1}{2}}} = \frac{S - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}}. \quad \text{.....(2.1)}$$

When H_0 is true and $n \ge 15,$ Z is approximately $N(0,\ 1).$ For the large sample approximation, it is common to use a ***continuity correction***, by replacing S by $S + 1/2$ in the definition of Z. Equation (2.1) then becomes

$$Z = \frac{\left(S + \frac{1}{2}\right) - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}}. \quad \text{.....(2.2)}$$

Example 2.3

The following data gives the ages, in years, of a random sample of 20 students from Besease Senior High School. It is believed that the median age of students in this school is smaller than 22 years. Based on these data, is there sufficient evidence to conclude that the median age of students from Besease Senior High School is smaller than 22 years?

9	13	16	16	16	17	18	19	19	19
19	20	20	21	21	23	24	25	25	27

Solution

The parameter of interest is $\tilde{\mu}$, the median age of students from Besease Senior High School. We are interested in testing the null hypothesis

$$H_0\!: \tilde{\mu} = 22 \quad \text{against}$$

$$H_1\!: \tilde{\mu} < 22.$$

The test statistic is

$$S = N^{+},$$

where N^{+} = number of observations X_i ***greater than*** 22

= number of +signs when the differences $X_i - 22$ are computed, i = 1, 2, ...20.

When H_0 is true, $S \sim b\left(20, \frac{1}{2}\right)$. Since n > 15, we use the normal approximation to the binomial distribution with a continuity correction. The test statistic then becomes

$$Z = \frac{(S + 0.5) - 0.5 \times 20}{0.5\sqrt{20}}.$$

When H_0 is true, Z is N(0, 1). Let z_o denote the observed value of the test statistic Z. We reject H_0 at the 0.05 level of significance when $z_o \le z_\alpha = z_{0.05} = -1.645$. The following table gives the signs of $X_i - 22$.

9	13	16	16	16	17	18	19	19	19
–	–	–	–	–	–	–	–	–	–
19	20	20	21	21	23	24	25	25	27
–	–	–	–	–	+	+	+	+	+

From the above table, $N^{+} = 5$. Thus, the observed value of the statistic S is 5. This gives,

$$z_o = \frac{(5 + 0.5) - 0.5 \times 20}{0.5\sqrt{20}} = -2.0125.$$

Since -2.0125 is less than -1.645, we reject H_0 at the 0.05 level of significance and conclude that the median age of students of Besease Senior High School is less than 22 years.

2.2.4 Confidence interval for the median based on the sign test

- The $100(1-\alpha)$ confidence interval for $\tilde{\mu}$ consists of those values of $\tilde{\mu}_0$ for which we would not reject a two-sided null hypothesis $H_0: \tilde{\mu} = \tilde{\mu}_0$ at the α level of significance.
- We designate the lower limit of our confidence interval by $\tilde{\mu}_L$ and the upper limit by $\tilde{\mu}_U$.
- We determine the largest positive or negative signs, (i.e. the value s') such that

$$P\left(S \le s' \middle| n,\ 0.5\right) = \frac{\alpha}{2}.$$

- When the data values are arranged in order of magnitude, the $(s'+1)^{\text{th}}$ observation is $\tilde{\mu}_L$. To find $\tilde{\mu}_U$, the upper limit of the confidence interval, we count the ordered sample values backwards from the largest. The $(s'+1)^{\text{th}}$ observation from the largest value locates $\tilde{\mu}_U$. i.e. $\tilde{\mu}_U = (n-s')^{\text{th}}$ value.

Example 2.4

Construct a 95% confidence interval for the median of the population from which the following sample data have been drawn, using the sign test.

0.07	0.69	1.74	1.90	1.99	2.41	3.07	3.08
3.10	3.57	3.71	4.01	8.11	8.23	9.10	10.16

Solution

- The point estimate of the population median is the sample median which is the mean of the two middle values in the ordered array. Thus,

  the sample median = $\frac{3.08 + 3.10}{2} = 3.09.$

- To find $\tilde{\mu}_L$, we consult a table of the binomial distribution and find that

  $P\left(S \le 3 \middle| 16,\ 0.5\right) = 0.0105$ and $P\left(S \le 4 \middle| 16,\ 0.5\right) = 0.0383.$

- Thus, we note that we cannot obtain an exact 95% confidence interval for the median. Since 100[1 – 2(0.0105)] = 97.9, which is larger than 95 and 100[1 – 2(0.0383)] = 92.34, which is smaller than 95.
- This method of constructing confidence intervals for the median does not usually yield intervals with exactly the usual coefficients of 0.90 , 0.95 , and 0.99.

- In practice, we choose between a wider interval and a higher confidence or the narrower interval and lower confidence.
- Suppose we choose $s' = 4$, then $s' + 1 = 5$. Therefore the 5th value in the ordered array is $\tilde{\mu}_L$ and the 12th (i.e. 16 – 4) value in the ordered array is $\tilde{\mu}_U$.
- Thus $\tilde{\mu}_L = 1.99$ and $\tilde{\mu}_U = 4.01$.
- The confidence coefficient is therefore 100[1 – 2(0.0383)] = 92.34. We say that we are 92.34% confident that the population median is between 1.99 and 4.01.

Large Sample Approximation

We find k such that

$$P\left(S \le k \middle| n,\ 0.5\right) = \tfrac{\alpha}{2} \text{ and}$$

$$P\left(\frac{S - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}} \le \frac{k - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}}\right) = \frac{\alpha}{2}$$

$$\Rightarrow \quad P\left(Z \le \frac{k - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}}\right) = \frac{\alpha}{2}$$

$$\Rightarrow \quad P\left(Z \le z_{\frac{1}{2}\alpha}\right) = \frac{\alpha}{2}$$

where Z is $N(0, 1)$ and $z_{\frac{1}{2}\alpha} = \dfrac{k - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}}$.

Making k the subject of the above equation, we obtain

$$k = \tfrac{1}{2}n + \tfrac{1}{2}z_{\frac{\alpha}{2}}\sqrt{n} = \tfrac{1}{2}\left(n + z_{\frac{\alpha}{2}}\sqrt{n}\right).$$

Approximately $k = s' + 1$. If the resulting value is not an integer, we use the closest integer.

Example 2.5

Refer to Example 2.3. Construct a 95% confidence interval for $\tilde{\mu}$.

Solution

Here, $n = 20$, $z_{\frac{\alpha}{2}} = z_{0.025} = -1.96$.

$$k = s' + 1 = \tfrac{1}{2}\left(20 - 1.96\sqrt{20}\right) = 5.6 \simeq 6. \text{ and } s' = 5.$$

Therefore the 6^{th} observation in the ordered array is $\tilde{\mu}_L$ and the $(20-5)^{\text{th}} = 15^{\text{th}}$ observation in the ordered array is $\tilde{\mu}_U$. Thus, $\tilde{\mu}_L = 16$ and $\tilde{\mu}_U = 23$. Hence the 95% confidence interval for $\tilde{\mu}$ is $16 < \tilde{\mu} < 23$.

2.3 The Wilcoxon signed-ranks test

As we have seen, the sign test utilizers only the signs of the differences between observed values and the hypothesized median. For testing $H_0: \tilde{\mu} = \tilde{\mu}_0$, there is another procedure that uses the magnitude of the differences when these are available. The Wilcoxon signed-ranks procedure makes use of additional information to rank the differences between the sample measurements and the hypothesized median. The Wilcoxon signed-ranks test uses more information than the sign test, making it a more powerful test when the sampled population is symmetric. However, the sign test is preferred when the sampled population is not symmetric.

2.3.1 Assumptions

1. The sample available for analysis is a random sample of size *n* from a population with an unknown median $\tilde{\mu}$.
2. The variable of interest is measured on a continuous scale.
3. The sampled population is symmetric.
4. The scale of measurement is at least interval.
5. The observations are independent.

2.3.2 Hypotheses

The parameter of interest is $\tilde{\mu}$, the population median. To test the hypothesis

$$H_0: \tilde{\mu} = \tilde{\mu}_0$$

where $\tilde{\mu}_0$ is the hypothesized median, against a corresponding one-sided or two-sided alternative, we can also use the Wilcoxon signed-ranks test.

2.3.3 Test statistic

To obtain the test statistic, we use the following procedure.

1. Subtract the hypothesized median $\tilde{\mu}_0$ from each observation X_i, that is, for each observation X_i, find

$$D_i = X_i - \tilde{\mu}_0, \ \forall\ i = 1,\ 2,\ \ldots,\ n.$$

2. If any observation X_i is equal to the hypothesized median, $\tilde{\mu}_0$, eliminate it from the calculations and reduce the sample size accordingly.
3. Rank the differences $|D_i|$, from the smallest to largest without regard to their signs. If two or more $|D_i|$ are tied, assign each tied value the mean of the rank positions of the tied differences.
4. Assign to each rank the sign of the difference of which it is ranked.
5. Obtain the sum of the ranks with positive signs; call it W^+. Obtain the sum of the ranks with negative signs; call it W^-.
6. Note that:

$$W^+ = \frac{n(n+1)}{2} - W^-.$$

7. For a given sample, we do not expect W^+ to be equal to W^-.

2.3.4 Carrying out the Wilcoxon signed ranks test

When the null hypothesis,

$$H_0: \tilde{\mu} = \tilde{\mu}_0,$$

is true, we do not expect a great difference between W^+ and W^-. Consequently, a sufficiently small value of W^+ or a sufficiently small value of W^- causes us to reject H_0.

(a) ***One-sided test***: To test

$$H_0: \tilde{\mu} = \tilde{\mu}_0, \text{ against}$$
$$H_1: \tilde{\mu} < \tilde{\mu}_0$$

at the α level of significance.

Test statistic

A sufficiently small value of W^+ leads to the rejection of the null hypothesis H_0. The test statistic therefore is

$$W = W^+.$$

Decision rule

We reject H_0 at significance level α if the observed W value w_o, is less than or equal to the tabulated W value for n and a preselected α.

(b) ***One-sided*** test: To test

$$H_0: \tilde{\mu} = \tilde{\mu}_0, \text{ against}$$
$$H_1: \tilde{\mu} > \tilde{\mu}_0$$

at the α level of significance.

Test statistic

For a sufficiently small W^- value, we reject H_0. The test statistic therefore is

$$W = W^-,$$

since a small value causes us to reject the null hypothesis.

Decision rule

We reject H_0 at significance level α if the observed W value, w_o, is less than or equal to the tabulated W value for n and a preselected value of α.

(c) For a ***two-sided*** test, we test

$$H_0: \tilde{\mu} = \tilde{\mu}_0, \quad \text{against}$$
$$H_1: \tilde{\mu} \neq \tilde{\mu}_0$$

at the α level of significance.

Test statistic

The test statistic is

$$W = \min\left(W^-,\ W^+\right),$$

since a small value of either W^- or W^+ causes us to reject the null hypothesis.

Decision rule

We reject H_0 at significance level α if the observed W value, w_o, is less than or equal to the tabulated W value for n and a preselected value of $\frac{\alpha}{2}$.

The distribution of *W*

1. The smallest value W can take is zero (0) and the largest value that W can take is the sum of the integers from 1 to n: that is, $\frac{n(n+1)}{2}$. W is therefore a discrete random variable whose support ranges between 0 and $\frac{n(n+1)}{2}$.
2. It can be shown that the probability mass function of the discrete random variable W is given by

$$P(W = w) = f(w) = \frac{c(w)}{2^n}, \qquad 0 < w < \frac{n(n+1)}{2},$$

where $c(w)$ = the number of possible ways to assign a +sign or a −sign to the first n integers so that the sum of the ranks with +signs (or –signs) is equal to w.

Example 2.6

The following are the systolic blood pressures (mmHg) of 13 patients undergoing a drug therapy for hypertension:

183	178	152	157	194	163	144	114	179	150	118	158	165

Can we conclude on the basis of these data that the median systolic blood pressure is less than 165 mmHg? Take $\alpha = 0.05$.

Solution

The parameter of interest is $\tilde{\mu}$, the median systolic blood pressure of the population. We wish to test the hypothesis

H_0: $\tilde{\mu} = 165$ against

H_1: $\tilde{\mu} < 165$.

at the $\alpha = 0.05$ level of significance. Using the Wilcoxon signed rank test, the test statistic is

$$W = W^+,$$

where W^+ is the sum of the ranks with positive signs.

We reject H_0 at the 0.05 level of significance if $w_o \le w_{12,\ 0.05} = 17$, where w_o is the observed value of the test statistic.

Table 2.1: Computation of test statistic

No.	X_i	D_i $X_i - 165$	Ranks (−)	Ranks (+)
1	114	-51	12	
2	118	-47	11	
3	144	-21	10	
4	150	-15	7	
5	152	-13	4.5	
6	157	-8	3	
7	158	-7	2	
8	163	-2	1	
9	178	13		4.5
10	179	14		6
11	183	18		8
12	194	29		9
			50.5	27.5

From Table 2.1, $W^- = 50.5$ and $W^+ = 27.5$. The value of the test statistic is therefore $w_o = 27.5$. Since 27.5 > 17, we fail to reject H_0. We conclude that the median systolic blood pressure of the subjects in the population is not less than 165 mmHg.

Example 2.7

Refer to Example 2.2. Use the ***Wilcoxon signed-ranks test*** to determine if there is any evidence that the median IQ of drug abusers in the population is different from 107. Use $\alpha = 0.05$.

Solution

Let $\tilde{\mu}$ denote the median IQ of drug abusers who are aged 16 years or older. We wish to test the hypothesis

$$H_0: \tilde{\mu} = 107 \text{ against}$$
$$H_1: \tilde{\mu} \neq 107.$$

at the $\alpha = 0.05$ level of significance. The test statistic is

$$W = \min\left(W^-, W^+\right).$$

where W^- and W^+ are the sums of the ranks with negative and positive signs, respectively.

We reject H_0 at the 0.05 level of significance if $w_o \leq w_{14,\ 0.025} = 21$, where w_o is the observed value of the test statistic.

From Table 2.2, $W^- = 40.5$ and $W^+ = 64.5$. The value of the test statistic is $w_o = 40.5$.

Table 2.2: Computation of test statistic

No.	IQ	D_i	Ranks (–)	Ranks (+)
1	90	-17	11	
2	94	-13	10	
3	99	-8	7	
4	100	-7	6	
5	104	-3	4	
6	105	-2	2.5	
7	108	1		1
8	109	2		2.5
9	111	4		5
10	117	10		8
11	119	12		9
12	125	18		12
13	127	20		13
14	135	28		14
			40.5	64.5

Since 40.5 > 21, we fail to reject H_0. We conclude that the median IQ of the subjects in the population may be 107.

2.3.5 Large sample approximation

Theorem 2.1

When the null hypothesis is true,

$$E(W) = \frac{n(n+1)}{4} \quad \text{and} \quad V(W) = \frac{n(n+1)(2n+1)}{24}.$$

Proof

When H_0 is true, W can be defined as $W = \sum_{i=1}^{n} W_i$ where

- $W_i = 0$ with probability $\frac{1}{2}$
- $W_i = i$ with probability $\frac{1}{2}$.

Thus,

$$E(W) = \sum_{i=1}^{n} E(W_i) = \sum_{i=1}^{n}\left[0\left(\tfrac{1}{2}\right)+i\left(\tfrac{1}{2}\right)\right] = \frac{1}{2}\sum_{i=1}^{n} i = \frac{1}{2}\times\frac{n(n+1)}{2} = \frac{n(n+1)}{4}.$$

Since $W_1, W_2, \ldots, W_n$ are independent,

$$V(W) = \sum_{i=1}^{n} V(W_i)$$

$$V(W_i) = E(W_i^2) - [E(W_i)]^2 = \left[0^2\left(\tfrac{1}{2}\right)+i^2\left(\tfrac{1}{2}\right)\right] - \left(\tfrac{i}{2}\right)^2 = \frac{1}{2}i^2 - \frac{1}{4}i^2 = \frac{1}{4}i^2.$$

$$V(W) = \sum_{i=1}^{n}\frac{1}{4}i^2 = \frac{1}{4}\sum_{i=1}^{n} i^2 = \frac{1}{4}\times\frac{n(n+1)(2n+1)}{6} = \frac{n(n+1)(2n+1)}{24}.$$

Theorem 2.2

When the null hypothesis is true, for large n:

$$Z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

follows an approximate standard normal distribution $N(0, 1)$.

Proof

If W is a random variable with mean $\frac{n(n+1)}{4}$ and variance $\frac{n(n+1)(2n+1)}{24}$, then by the central limit theorem,

$$Z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

is approximately $N(0, 1)$.

Adjustment for Ties

- We can incorporate an adjustment for ties among nonzero differences in the large sample approximation in the following way.
- Let t be the number of absolute differences tied for a particular nonzero rank. Then the correction factor is

$$\frac{\sum t^3 - \sum t}{48}.$$

- We can subtract this quantity from the expression in the denominator under the square root sign.
- Thus the adjusted statistic for a large sample approximation is

$$Z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}}.$$

- We illustrate the calculation of an adjustment for ties in the following data:

Table 2.3: Computation of correction factor

Observation	Rank	t	t^3
3	1.5	2	8
3	1.5		
4	3		
6	5	3	27
6	5		
6	5		
8	7.5	2	8
8	7.5		
9	10.5	4	64
9	10.5		
9	10.5		
9	10.5		
		11	107

$$\frac{\sum t^3 - \sum t}{48} = \frac{107 - 11}{48} = 2.$$

Example 2.8

The following data show the life span, in years, of a random sample of 21 recorded deaths in a certain country. It has been known in the past years that the median life span in the country is 50 years. Can we conclude from these data that the median life span in the country has improved? Use $\alpha = 0.05$

39	42	42	47	47	53	59	59	59	60	62
65	66	68	69	70	72	75	75	85	90	

Solution

Let $\tilde{\mu}$ denote the median life span in the population . We wish to test the hypothesis

H_0: $\tilde{\mu} = 50$ against

H_1: $\tilde{\mu} > 50$.

The test statistic is

$$Z = \frac{W^- - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}}.$$

When H_0 is true, W is $N(0, 1)$.

Reject H_0 at the 0.05 level of significance if $z \le z_{0.05} = -1.645$, where z is the computed value of Z. From Table 2.4,

$\sum t = 10$, $\sum t^3 = 70$, $W^{-1} = 23$ $W^+ = 210$

The value of the test statistic is

$$w_o = \frac{23 - \frac{21 \times 22}{4}}{\sqrt{\frac{21 \times 22 \times 43}{24} - \frac{70-10}{48}}} = -3.2175.$$

Since $-3.2175 < -1.645$, we reject H_0 at the 0.05 level of significance. We therefore conclude that, the median life span in the country has improved significantly.

Table 2.4: Computation of test statistic

	X_i	D_i	$\lvert D_i \rvert$	Rank	t	t^3
1	39	-11	11	10		
2	42	-8	8	4.5	2	8
3	42	-8	8	4.5		
4	47	-3	3	2	3	27
5	47	-3	3	2		
6	53	3	3	2		
7	59	9	9	7	3	27
8	59	9	9	7		
9	59	9	9	7		
10	60	10	10	9		
11	62	12	12	11		
12	65	15	15	12		
13	66	16	16	13		
14	68	18	18	14		
15	69	19	19	15		
16	70	20	20	16		
17	72	22	22	17		
18	75	25	25	18.5	2	8
19	75	25	25	18.5		
20	85	35	35	20		
21	90	40	40	21		
					10	70

2.3.6 Confidence Interval for the Median, based on the Wilcoxon Signed-Ranks Test

Arithmetic Procedure

Step1: Find the means, u_{ij}, of all possible pairs of observation x_i and x_j from the sample observation $x_1, x_2, \ldots, x_n$, that is

$$u_{ij} = \frac{x_i + x_j}{2}, \quad 1 \le i \le j \le n.$$

There are $\frac{n(n-1)}{2} + n$ such averages, distributed symmetrically about the median.

Step 2: Arrange the u_{ij} in an increasing order of magnitude.

Step 3: The median of the u_{ij}'s is a point estimate of the population median.

Step 4: Find, from the Wilcoxon Signed Ranks Test table, $t = w_{n,\,p}$ corresponding to the sample size n and appropriate value of p as determined by the desired confidence level. When the confidence coefficient is $(1-\alpha)$, then $p = \alpha/2$. If the exact value of p cannot be found in the Wilcoxon signed ranks test table, we choose a closer neighbouring value.

Step 5: The end points of the confidence interval are the k^{th} smallest and k^{th} largest values of u_{ij}'s where $k = t + 1$, where t is either value in the column labelled T corresponding to n and the value of p selected (see Wayne, 1978).

Example 2.9

Determine the 95% confidence interval for the population median by the Wilcoxon Signed-ranks procedure using the following data:

26	25	29	41	29	32	32	40	26	29

Solution

All the 55 possible pairs of means from the observations are given in the Table 2.5.

Table 2.5: All possible pairs of means from the observations

	25	26	28	29	29	29	32	38	40	41
25	25.0									
26	25.5	26.0								
28	26.5	27.0	28.0							
29	27.0	27.5	**28.5**	29.0						
29	27.0	27.5	28.5	29.0	29.0					
29	27.0	27.5	28.5	29.0	29.0	29.0				
32	28.5	29.0	30.0	30.5	30.5	30.5	**32.0**			
38	31.5	32.0	33.0	33.5	33.5	33.5	**35.0**	38.0		
40	32.5	33.0	34.0	34.5	34.5	34.5	36.0	39.0	40.0	
41	33.0	33.5	34.5	35.0	35.0	35.0	36.5	39.5	40.5	41.0

Thus, a point estimate of the population median $\tilde{\mu}$ is the 28^{th} observation of the ordered data in Table 2.5. This is 32. From the Wilcoxon signed ranks test table, $t = w_{10,\,0.025} = 8$. Thus,

$k = t + 1 = 9$. Therefore the 9th observation in the ordered array in Table 2.5 is the lower limit $\tilde{\mu}_L$ and the 9th observation from the largest value locates the upper limit $\tilde{\mu}_U$. Thus, $\tilde{\mu}_L = 28.5$ and $\tilde{\mu}_U = 35.0$. Therefore the 95% confidence interval for $\tilde{\mu}$ is $28.5 < \tilde{\mu} < 35.0$.

Large Sample Approximation

With samples larger than 30, we cannot use the Wilcoxon signed-ranks table to determine k. A large sample approximation of k is however given by (see Wayne, 1978)

$$k \cong \frac{n(n+1)}{4} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

Exercise 2(a)

1. The median age of the onset of diabetes is thought to be 45 years. The ages at onset of a random sample of 16 people with diabetes are:

26.2	30.5	35.5	38.0	39.8	40.3	45.0	45.6
45.9	46.8	48.9	51.4	52.4	55.6	60.9	65.4

Perform the
(a) sign test, (b) Wilcoxon signed-ranks test,
to determine if there is any evidence to conclude that the median age of the onset of diabetes differs significantly from 45 years. Take $\alpha = 0.05$.

2. Recent studies of the private practices of physicians who saw no Medicaid patients suggested that the median length of each patient visit was 22 minutes. It is believed that the median visit length in practices with a large Medicaid load is shorter than 22 minutes. A random sample of 20 visits in practices with a large Medicaid load yielded, in order, the following visit lengths:

9.4	13.4	15.6	16.2	16.4	16.8	18.1	18.7	18.9	19.1
19.3	20.1	20.4	21.6	21.9	23.4	23.5	24.8	24.9	26.8

(a) Use the large sample approximation of the sign test to determine if there is sufficient evidence to conclude, at the 1% level of significance, that the average visit length in practices with a large Medicaid load is shorter than 22 minutes?
(b) Based on the sign test, construct a 95% confidence interval for the median visit length in practices with a large Medicaid load.

3. The following are the blood glucose levels of 12 patients who attend St. Thomas Hospital:

86	100	120	90	101	98	109	108	93	107	99	110

Perform the Wilcoxon signed ranks test to determine if we can conclude on the basis of these data that the average glucose level in the population is greater than 96 mg/dl? Take $\alpha = 0.05$.

4. From a random sample of 14 students from Accra Catholic Senior High School, the body masses of 9 students were found to be less than 38 kg whilst those of 4 students exceeded 38 kg with the remaining students recording exactly 38 kg. Can we conclude, based on a sign test, that the average body mass of students from the school is less than 38 kg?

5. In a sample of 25 adolescents who served as the subjects in an immunologic study, one variable of interest was the diameter of skin test reaction to an antigen. The sample observations, in mm erythema, were as follows:

16.0	17.0	18.0	19.0	20.0	21.0	22.0	22.0	22.0	23.0	24.0	26.0	27.0
28.0	29.0	30.0	30.0	31.0	32.0	33.0	34.0	35.0	36.0	36.0	37.0	

Use the large sample approximation of the Wilcoxon signed ranks test to determine if we can conclude from these data that the population average is less than 30 mm. Take $\alpha = 0.05$.

6. Barrett (1991) reported data on eight cases of umbilical cord prolapse. The maternal ages were 25, 28, 17, 26, 27, 18, 25, and 30.
   (a) Perform the Wilcoxon signed ranks test to determine if there is enough evidence, based on the data, that the average age of the population from which the sample may be presumed to have been drawn is greater than 20 years. Take $\alpha = 0.01$.
   (b) Based on the Wilcoxon signed ranks test, construct a 99% confidence interval for the population median.

7. Out of a random sample of 100 recorded deaths in a certain country during the past year, 68 of them were more than 65 years whilst the remaining 32 were below 65 years. Perform a sign test to determine if we can we conclude that the average life span in the country is greater than 65 years. Use $\alpha = 0.05$.

8. Recent studies of the private practices of physicians who saw no Medicaid patients suggested that the median length of each patient visit was 22 minutes. It is believed that the median visit length in practices with a large Medicaid load is shorter than 22 minutes. A random sample of 20 visits in practices with a large Medicaid load yielded, in order, the following visit lengths:

9.4	13.4	15.6	16.2	16.4	16.8	18.1	18.7	18.9	19.1
19.3	20.1	20.4	21.6	21.9	23.4	23.5	24.8	24.9	26.8

Based on the large sample approximation of the sign test, is there sufficient evidence to conclude that the average visit length in practices with a large Medicaid load is shorter than 22 minutes?

9. To determine whether the median life span of certain spices of animal is greater than 5 years, a random sample of 25 observations were made and life span in years is the following:

11.3	5.8	3.1	4.1	7.3	4.4	1.4	2.5	6.6	7.6	24.9	30.1	2.9
5.5	7.2	3.2	3.9	7.2	20.1	3.1	6.1	4.9	19.4	4.2	6.3	

At 0.05 level of significant, use the large sample approximation of the sign test to determine if the average life span is greater than 5 years.

10. A physician states that the median number of times he sees each of his patients during the year is five. In order to evaluate the validity of this statement, he randomly selects ten of his patients and determines the number of office visits each of them made during the past year. He obtains the following values for the ten patients in his sample: 9, 10, 8, 4, 8, 3, 0, 10, 15, 9. Do the data support his contention that the median number of times he sees a patient is five?

11. Moore and Ogletree (1973) investigated the readiness of pupils at the beginning of the first grade. They compared scores on a readiness test of pupils who had attended a head start program for a full year with the scores of those who had not. The readiness test scores of 10 pupils who did not attend a Head Start program are as follows: 33, 19, 40, 35, 51, 41, 27, 55, 39, 21. Can we conclude, based on the Wilcoxon signed ranks test, that the median score of the population represented by this sample is less than 45.3? Take $\alpha = 0.05$.

12. Abu-Ayyash (1972) found that the median education of heads of households living in mobile homes in a certain area was 11.6 years. Suppose that a similar survey conducted in another area revealed the educational levels of heads of households as shown in the following data.

13	6	6	12	12	10	9	11	14	8	7	16	15	8	7

Based on the sign test, can we conclude that the average educational level of the population represented by this sample is less than 11.6 years? Take $\alpha = 0.05$.

13. Lenzer et al. (1973) reported the endurance score of animals during a 48-hour session of discrimination responding. The median score for an animal with electrodes implanted in

the hypothalamus was 97.5. Suppose that the experiment was duplicated in another laboratory, except that electrodes were implanted in the forebrain in 12 animals. Assume that investigators observed the endurance score shown in the following table.

93.6	89.1	97.7	84.4	97.8	94.5	88.3	97.5	83.7	94.6	85.5	82.6

Use the one-sample sign test to see whether the investigators may conclude at the 0.05 level of significance that the median endurance score of animals with electrodes implanted in the forebrain is less than 97.5.

14. Iwamoto (1971) found that the mean weight of a sample of a particular species of adult female monkey from a certain locality was 8.41 kg. Suppose that a sample of adult females of the same species from another locality yielded the weights as shown in the following table. By using the one-sample sign test, can we conclude, at the 0.05 level of significance, that the median weight of the population from which this second sample was drawn is greater than 8.41 kg?

8.30	9.50	9.60	8.75	8.40	9.10	9.25	9.80	10.05	8.15	10.00	9.60	9.80	9.20	9.30

2.4 The binomial test

Inferences concerning proportions are required in many areas. The population proportion is a parameter of frequent interest in research and decision-making activities. The politician is interested in knowing what proportion of voters will vote for him in the next election. All manufacturing firms are concerned about the proportion of defective items when a shipment is made. A market analyst may wish to know the proportion of families in a certain area who have central air conditioning. A sociologist may want to know the proportion of heads of household in a certain area who are women. Many questions of interest to the health worker relate to the population proportion. What proportion of patients who receive a particular treatment recover? What proportion of a population has a certain disease?

When it is impossible or impractical to survey the total population, researchers base decision regarding population proportions, on inferences made by analyzing samples drawn from the population. As usual, inference may take the form of interval estimation or hypothesis testing.

Sometimes, we want to draw inferences concerning the total number, the proportion or percentage of units in the population that possess some characteristic or attribute or fall into some defined class. A random sample of size n is drawn from a population. Suppose we wish to estimate the proportion, p, of units in the population that belong to some definite class in the population.

Testing hypotheses about population proportions is carried out in much the same way as for median when the assumptions necessary for the test are satisfied.

2.4.1 Assumptions

1. The data consist of a sample of the outcomes of n repetitions of some process. Each outcome consists of either a 'success' or a 'failure'. The proportion of the sample having a characteristic of interest is $\hat{p} = S/n$ an estimate of the population proportion p, where S is the number of successes (the total number of sampling units with a particular characteristic of interest).
2. The n trials are independent.
3. The probability of a success p, remains constant from trial to trial.

2.4.2 Hypotheses

One-sided and two-sided tests may be made, depending on the question being asked. In other words, we can test $H_0\colon p = p_0$ against one of the alternatives $p < p_0$, $p > p_0$ or $p \neq p_0$.

(a) **One-sided test**

$H_0\colon p = p_0$ against

$H_1\colon p < p_0$.

Test statistic

Since we are interested in the number of successes S, our test statistic is S. When H_0 is true, S has the binomial distribution with parameters n and p_0. That is $S \sim b(n,\, p_0)$.

Decision rule

Sufficiently small values of S lead to the rejection of H_0. Let s_o denote the observed value of S. We reject H_0 at the α level of significance if the p-value of the test ≤ α, where

$$p\text{-value} = P\left(S \le s_o \mid n, p_0\right).$$

(b) **One-sided test**

$H_0\colon p = p_0$ against

H_1: $p > p_0$.

Test statistic

The test statistic therefore is S. When H_0 is true, $S \sim b(n,\ p_0)$.

Decision rule

For sufficiently large values of S, we reject H_0. Thus, we reject H_0 at α level of significance if the p-value of the test $= P\left(S > s_o \middle| n, p_0\right) \leq \alpha$, where s_o is the observed value of S.

(c) **Two-sided test**

Here, we test

H_0: $p = p_0$ against

H_1: $p \neq p_0$.

Test statistic

The test statistic therefore is S. When H_0 is true, $S \sim b(n,\ p_0)$.

Decision rule

For sufficiently large or sufficiently small values of S, we reject H_0. The hypothesized proportion is p_0 whilst the observed sample proportion $\hat{p} = \frac{s_o}{n}$, where s_o is the observed value of S. The p-value of the test is defined by

$$p\text{-value} = \begin{cases} 2P\left(S \leq s_o \middle| n, p_0\right), & \text{if } \hat{p} < p_0, \\ 2P\left(S > s_o \middle| n, p_0\right), & \text{if } \hat{p} > p_0. \end{cases}$$

We reject H_0 at the α level of significance if the p-value of the test $\leq \alpha$.

Example 2.10

In a survey of injection drug users in a large city, Coates et al. (1991) found that 2 out of 12 were HIV positive. We wish to know if we can conclude, at the 10% level of significance, that fewer than 40% of the injection drug users in the sampled population are HIV positive.

Solution

The parameter of interest is p, the proportion of injection drug users in the sampled population who are HIV positive. We wish to test

$$H_0:\ p = 0.4 \text{ against } H_1:\ p < 0.4$$

at significance level $\alpha = 0.1$. The test statistic is S, the number of injection drug users in the sample who are HIV positive. When H_0 is true, S has the binomial distribution with parameters $n = 12$ and $p = 0.4$. Thus,

$$S \sim b(12,\ 0.4).$$

Let s_o denote the observed value of the test statistic. We reject H_0 at the 0.1 level of significance if the p-value $\leq$ 0.1, where the $p\text{-value} = P\left(S \leq s_o \middle| 12,\ 0.4\right)$. Given $s_o = 2$,

$$p\text{-value} = P\left(S \leq 2 \middle| 12,\ 0.4\right) = 0.0834.$$

Since the p-value, 0.0834 < 0.1, we reject H_0 at the 10% level of significance and conclude that fewer than 40% of the injection drug users in the sampled population are HIV positive.

Example 2.11

A researcher found anterior sub-capsular vacuoles in the eyes of 6 out of 15 diabetic patients. Using the binomial test, can we conclude that the population proportion with the condition of interest is greater than 0.2? Use $\alpha = 0.05$.

Solution

The parameter of interest is p, the proportion of diabetic patients in the population with anterior sub-capsular vacuoles in the eyes. We wish to test

$$H_0:\ p = 0.2 \text{ against } H_1:\ p > 0.2.$$

The test statistic is S, the number of diabetic patients in the sample with anterior sub-capsular vacuoles in the eyes. When H_0 is true,

$$S \sim b(15,\ 0.2).$$

Let s_o denote the observed value of the test statistic. We reject H_0 at the 0.05 level of significance if the p-value of the test ≤ 0.05, where $p\text{-value} = P\left(S > s_o \middle| 15,\ 0.2\right)$. Given $s_o = 6$,

$$p\text{-value} = P\left(S > 6 \middle| 15,\ 0.2\right) = 1 - P\left(S \leq 6 \middle| 15,\ 0.2\right) = 1 - 0.9819 = 0.0181.$$

Since the p-value 0.0181 < 0.05, we reject H_0 at the 0.05 level of significance and conclude that the population proportion p is greater than 0.2.

2.4.3 Large sample approximation

1. If S is a binomial random variable with parameters n and p_0, then the expectation and variance of S are given by

   $E(S) = np_0$ and $V(W) = np_0(1 - p_0)$.

2. Thus, when the null hypothesis is true, and n is large,

   $$Z = \frac{S - np_0}{\sqrt{np_0(1 - p_0)}}$$

   follows an approximate standard normal distribution, $N(0, 1)$.

3. The normal approximation to the binomial distribution is good if $np_0 > 5$ and $n(1 - p_0) > 5$.

3. Note that the sign-test discussed earlier is a special case of the binomial test, in which $p_0 = 0.5$.

Example 2.12

A commonly prescribed drug for relieving nervous tension is believed to be only 60% effective. Experimental results with a new drug administered to a random sample of 100 adults who were suffering from nervous tension show that 70 received relief. Is this sufficient evidence to conclude that the new drug is superior to the one commonly prescribed? Use $\alpha = 0.05$.

Solution

The parameter of interest is p, the proportion of adults in the population who received relief from nervous tension. We wish to test

$$H_0: \ p = 0.6 \text{ against } H_1: \ p > 0.6$$

at $\alpha = 0.05$ level of significance. The test statistic is

$$Z = \frac{S - np_0}{\sqrt{np_0(1 - p_0)}}.$$

Given $n = 100$ and $p_0 = 0.6$, both np_0 and $n(1 - p_0)$ are greater than 5 and so Z is approximately $N(0, 1)$ when H_0 is true. We reject H_0 if z, the computed Z value is greater than $z_{0.95} = 1.645$. Now, $S = 70$ and

$$z = \frac{70 - 100 \times 0.6}{\sqrt{100 \times 0.6 \times 0.4}} = 2.0412.$$

Since 2.0412 > 1.645, we reject H_0 at the 0.05 level of significance. We conclude that the new drug is superior to the one commonly prescribed.

2.4.4 Large sample confidence interval for p

If $\hat{p}$ is the proportion of observations in a random sample of size n that belongs to a class of interest, then an approximate 100(1 – α)% confidence interval of the proportion p of the population that belongs to this class is (see Ofosu & Hesse, 2011)

$$\hat{p} - z_{1-\frac{1}{2}\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{1-\frac{1}{2}\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $\hat{p} = s/n$ is the proportion of the sample with the characteristic of interest.

Example 2.13

In a certain university, the proportion of students who have diabetes mellitus is p. Of the 500 students selected at random from the university, 6 had diabetes mellitus.

(a) Find a point estimate of p. (b) Construct a 90% confidence interval for p.

Solution

(a) A point estimate of p is given by $\hat{p} = \frac{6}{500} = 0.012$.

(b) $n\hat{p} = 6$ and $n(1-\hat{p}) = 494$. Both $n\hat{p}$ and $n(1-\hat{p})$ are of sufficient magnitude to justify the use of the formula for constructing a confidence interval for p. To construct a 90% confidence interval, we put $1-\alpha = 0.90$. This gives α = 0.10. From the standard normal table, we find that $z_{1-\frac{1}{2}\alpha} = z_{0.95} = 1.645$. Hence a 90% confidence interval for p is

$$0.012 - 1.645\sqrt{\frac{0.012 \times 0.988}{500}} < p < 0.012 + 1.645\sqrt{\frac{0.012 \times 0.988}{500}},$$

which simplifies to $0.004 < p < 0.020$.

Exercise 2(b)

1. A researcher found that 66% of a sample of 14 infants had completed the hepatitis B vaccine series. Can we conclude on the basis of these data that, in the sampled population, more than 60% have completed the series? Use α = 0.01.

2. A health survey of 12 male inmates 50 years of age and older residing in a state's correctional facilities was made. They found that 22% of the respondents reported a history of venereal disease. On the basis of these findings, can we conclude that in the sampled population, more than 15% have a history of venereal disease? Use α = 0.05.

3. The fraction of defective integrated circuits produced in a photolithography process is being studied. A random sample of 300 circuits is tested, revealing 13 defectives. Use the data to test H_o: $p = 0.05$ against H_1: $p \neq 0.05$. Use $\alpha = 0.05$.

4. A commonly prescribed drug for relieving nervous tension is believed to be only 70% effective. Experimental results with a new drug administered to a random sample of 10 adults who were suffering from nervous tension show that 8 received relief. Is this sufficient evidence to conclude that the new drug is superior to the one commonly prescribed? Use $\alpha = 0.05$.

5. Suppose that, in the past, 40% of all adults favoured capital punishment. Do we have reason to believe that the proportion of adults favouring capital punishment today has increased if, in a random sample of 15 adults, 8 favour capital punishment? Use $\alpha = 0.05$.

2.5 The one-sample runs test for randomness

In many situations we want to know whether we can conclude that a set of observations constitute a random sample from an infinite population. Test for randomness is of major importance because the assumption of randomness underlies statistical inference (see Ofosu & Hesse, 2011). In addition, tests for randomness are important for time series analysis. The runs test procedure is used to examine whether or not a sequence of sample values is random.

Consider, for example, the following sequence of sample values

21 23 24 27 30 28 27 26 25 23 22 21

Each observation is denoted by a '+' sign if it is more than the previous observation and by a '– ' sign if it is less than the previous observation as shown in the following table.

21	23	24	27	30	28	27	26	25	23	22	21
	+	+	+	+	–	–	–	–	–	–	–
	1				2						

A run is a sequence of signs of the same kind bounded by signs of other kind. In this case, we doubt the sequence's randomness, since there are only two runs.

If the order of occurrence were

25	22	27	23	27	28	21	26	23	30	21	24
	–	+	–	+	+	–	+	–	+	–	+
	1	2	3	4		4	6	7	8	9	10

we would doubt the sequence's randomness because there are too many runs (10 in this instance).

Too few runs indicate that the sequence is not random (has persistency) whilst too many runs also indicate that the sequence is not random (is zigzag). Let us now consider the one sample runs test. This procedure helps us to decide whether a sequence of sample values is the result of a random process.

Assumptions

The data available for analysis consist of a sequence of sample values, recorded in the order of their occurrence.

Hypotheses

We wish to test

H_0: The sequence of sample values is random, against

H_1: The sequence of sample values is not random.

Test Statistic

The test statistic is R, the total number of runs.

Decision Rule

Since the null hypothesis does not specify the direction, a two-sided test is appropriate. The critical value, r_c, for the test is obtained from Table A.5, in the Appendix, for a given sample size n and at a desired level of significance α. If $r_c(\text{lower}) \leq r \leq r_c(\text{upper})$, accept H_0. Otherwise reject H_0.

Tied Values

If an observation is equal to its preceding observation, denote it by zero. While counting the number of runs, ignore it and reduce the value of n accordingly.

Large Sample Sizes

If $n > 25$, then the test statistic can be approximated by

$$Z = \frac{R - \mu_R}{\sqrt{\text{Var}(R)}},$$

which is $N(0, 1)$, when H_0 is true, where $\mu_R = E(R) = \frac{2n-1}{3}$ and $\text{Var}(R) = \frac{16n-29}{90}$. We reject H_0 at the level of significance α if $|z| > z_{1-\frac{\alpha}{2}}$, where z is the computed value of Z.

Example 2.15

The following are the blood glucose levels of 12 patients who attend St. Thomas Hospital: Test, at the 0.05 level of significance whether the sequence is random?

86	99	98	90	109	101	100	110	110	93	108	120

Solution

We wish to test

H_0: The sequence is random, against

H_1: The sequence is not random.

The test statistic is

$R =$ the number of runs.

We reject H_0 at the 0.05 level of significance if $r < r_c(\text{lower})$ or $r > r_c(\text{upper})$, where r is the observed value of R and r_c is the critical value. It can be seen that:

86	99	98	90	109	101	100	110	110	93	108	120
	+	–	–	+	–	–	+	0	–	+	+

Here n = 11 and the number of runs r = 7. From the table of critical values for runs up and down test, $r_c(\text{lower}) = 4$ and $r_c(\text{upper}) = 10$ (see Table A.5, in the Appendix)

Note: Since two consecutive observations are the same, that is 110, we use n = 11 instead of n = 12.

Since $4 < r < 10$, we fail to reject H_0 at the 0.05 level of significance and therefore conclude that the sequence is random.

Exercise 2(c)

1. The following data show the average daily temperatures recorded at Accra, Ghana, for 15 consecutive days during June 2017.

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Temperature	28	27	26	27	28	29	29	27	26	25	28	24	25	26	28

Test, at the 0.05 level of significance, if we can conclude that the pattern of temperature is random?

2. The following data show the inflation rate in Ghana from 2006 to 2017. Test, at the 0.05 level of significance, if we can conclude that the pattern of year inflation is random?

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
11.7	10.7	16.5	13.1	6.7	7.7	7.1	11.7	15.5	17.2	17.5	12.0

References

Abu-Ayyash, A. Y. (1972). The mobile home: A neglected phenomenon in geographic research. *Geog. Bull.*, **5**, 28 – 30.

Barrett, J. M. (1991). Funic reduction for the management of umbilical cord prolapse. *American Journal of Obstetrics and Gynaecology*. **165**, 654-657.

Coates, R., Millson, M., Myers, T. (1991). The benefits of HIV Antibody testing of saliva in field research. *Canadian Journal of Public Health*, **82**, 397-398.

Iwamoto, M. (1971). Morphological studies of Macaca Fuscata: VI, Somatometry. Primates, **12**, 151 – 174.

Lenzer, Irmingard I., and White, C. A. (1973). Statistical effects in continuous reinforcement and successive sensory discrimination situations. *Physiolog. Psychol*, **1**, 77 – 82.

Moore, R. C and Ogletree, E. J. (1973). A comparison of the readiness and intelligence of first grade children with and without a full year of Head Start training. *Education*, **93**, 266 – 270.

Ofosu, J. B., & Hesse, C. A. (2011). Elementary Statistical Methods. *EPP Books Services, Accra*.

Wayne, W. D. (1978). Applied nonparametric statistics. *Houghton Mifflin company, London*.