An Introduction to Nonparametric Statistics
## Contents

### Introduction 7

1 Background 9

1.1 Probability background ................................. 9

1.1.1 Location and Scale Families ....................... 9

1.1.2 Probability distributions for Observations ......... 9

1.1.2.1 Gaussian Distribution ........................ 10

1.1.2.2 Uniform Distribution .......................... 11

1.1.2.3 Laplace Distribution ........................... 11

1.1.2.4 Cauchy Distribution ............................ 11

1.1.2.5 Logistic Distribution ............................ 12

1.1.2.6 Exponential Distribution ....................... 12

1.1.3 Sampling distributions .............................. 12

1.1.3.1 Binomial Distribution .......................... 13

1.1.4 \( \chi^2 \)-distribution ................................ 13

1.1.5 \( T \)-distribution .................................. 14

1.1.6 \( F \)-distribution .................................. 14

1.2 Elementary Tasks in Frequentist Inference ............. 14

1.2.1 Hypothesis Testing .................................. 14

1.2.1.1 One-Sided Hypothesis Tests .................... 15

1.2.1.2 Two-sided Hypothesis Tests .................... 16

1.2.1.3 \( P \)-values .................................... 18

1.2.2 Confidence Intervals ............................... 19

1.2.2.1 Test Inversion with Pivotal Statistics .......... 19

1.2.2.2 A Problematic Example ......................... 20

1.2.2.3 \( P \)-value Inversion ............................ 20

1.3 Exercises ............................................ 21

2 One-Sample Nonparametric Inference 23

2.1 Parametric Inference .................................. 23

2.2 The Need for Distribution-Free Tests .................. 24

2.3 One-sample median methods ............................ 26

2.4 Confidence Intervals for the Median .................. 29

2.4.1 Inference for other percentiles .................... 32

2.5 Comparing Tests ..................................... 33

2.5.1 Power, Sample Size, and Effect Size ............... 34
3 Two-Sample Testing

3.1 Two-Sample Approximately Normal Inference

3.1.1 Two-Sample Approximately Normal Inference on Means

3.1.2 Two-Sample Approximately Normal Inference on Dispersions

3.2 General Two-Sample Rank Tests

3.2.1 Null Distributions of General Rank Statistics

3.2.2 Moments of Rank Statistics

3.3 A First Distribution-Free Test

3.4 The Mann-Whitney-Wilcoxon test

3.4.1 Exact and Approximate Mann-Whitney Probabilities

3.4.1.1 Mann-Whitney-Wilcoxon Statistic Moments and Approximate Normality

3.4.2 Other scoring schemes

3.4.3 Using data as scores: the Permutation Test

3.5 An Empirical Assessment of Levels and Powers of Two-Sample Tests

3.6 Adaptation to the Presence of Tied Observations

3.7 Mann-Whitney-Wilcoxon null hypotheses

3.7.1 An Example

3.8 Efficiency and Power of Two-Sample Tests

3.8.1 Efficacy of the Normal-Theory Test

3.8.1.1 Normally Distributed Observations

3.8.1.2 Logistically Distributed Observations

3.8.2 Efficacy of the Mann-Whitney-Wilcoxon test

3.8.2.1 Normally Distributed Observations

3.8.2.2 Logistically Distributed Observations

3.8.3 Summarizing Asymptotic Relative Efficiency

3.8.4 Power for Mann-Whitney-Wilcoxon Testing

3.9 Testing Equality of Dispersion

3.10 Two-Sample Estimation and Confidence Intervals

3.10.1 Inversion of the Mann-Whitney-Wilcoxon test

3.11 Tests for Broad Alternatives

3.12 Exercises
Contents

4 Methods for Three or More Groups 71
  4.1 Normal-Theory Methods 71
  4.1.1 Contrasts 72
  4.1.2 Multiple Comparisons 73
  4.2 General Rank Tests 74
  4.2.1 Moments of General Rank Sums 75
  4.2.2 Construction of a Chi-Square-Distributed Statistic 75
  4.3 The Kruskal-Wallis Test 78
  4.3.1 Kruskal-Wallis approximate critical values 78
  4.4 Other Scores for Multi-Sample Rank Based Tests 80
  4.5 Multiple Comparisons 82
  4.6 Ordered Alternatives 84
  4.7 Powers of Tests 86
  4.7.1 Power of Tests for Ordered Alternatives 86
  4.7.2 Power of Tests for Unordered Alternatives 87
  4.8 Efficiency Calculations 92
  4.9 Exercises 93

5 Group Differences with Blocking 95
  5.1 Paired Comparisons 95
  5.1.1 Estimating the population median difference 98
  5.1.2 Confidence Intervals 99
  5.1.3 Alternative Distribution for Wilcoxon Signed Rank Statistic 100
  5.2 Two-Way Non-Parametric Analysis of Variance 101
  5.2.1 Distribution of Rank Sums 101
  5.2.2 Two-Way Nonparametric Analysis of Variance in the Balanced Case 102
  5.2.3 Two-Way Nonparametric Analysis of Variance in the Unbalanced Case 105
  5.3 Multiple Comparisons and Scoring 106
  5.4 Tests for a Putative Ordering in Two-Way Layouts 106
  5.5 Exercises 108

6 Bivariate Methods 111
  6.1 Parametric Approach 111
  6.1.1 Permutation Inference 111
  6.2 Nonparametric Correlation 112
  6.2.1 Rank Correlation 113
  6.2.1.1 Moments of the Spearman correlation 115
  6.2.2 Kendall’s Tau 116
  6.3 Bivariate semi-parametric methods 118
  6.3.1 Inversion of the test of zero correlation 118
  6.3.1.1 Pearson Correlation 119
  6.3.1.2 Kendall’s τ 119
6.3.1.3 Spearman Correlation .......................... 121
6.4 Exercises ........................................... 123

7 Multivariate Analysis .................................. 125
  7.1 Multivariate Estimation ............................. 125
     7.1.1 Invariance Properties ........................ 126
  7.2 One-Sample Hypothesis Testing .................... 126
     7.2.1 The Standard Parametric Approach ............ 126
     7.2.2 Nonparametric Asymptotic Approaches ......... 127
     7.2.3 More General Permutation Solutions .......... 130
  7.3 Confidence Regions for a Vector Shift Parameter .... 131
  7.4 Two-sample Methods ................................ 131
     7.4.1 Hypothesis Testing ............................ 132
        7.4.1.1 Traditional normal-theory approach ....... 132
        7.4.1.2 Permutation Testing ...................... 133
        7.4.1.3 Approximations to the Permutation Distribu-
                        tion ............................... 135
  7.5 Exercises ........................................... 136

8 Density Estimation .................................... 139
  8.1 Histograms ......................................... 139
  8.2 Kernel density estimates ........................... 140
  8.3 Exercises .......................................... 144

9 Regression Function Estimates ........................ 145
  9.1 Standard Regression ............................... 145
  9.2 Kernel and Local Regression Smoothing .......... 146
  9.3 Isotonic regression .................................. 149
  9.4 Splines ............................................. 151
  9.5 Quantile Regression ................................ 153
     9.5.1 Fitting the Quantile Regression Model ......... 154
  9.6 Exercises .......................................... 156

10 The Bootstrap ....................................... 159
  10.1 Univariate Bootstrap Techniques ................. 159
      10.1.1 The Normal Method ........................... 160
      10.1.2 Basic Interval ................................. 160
      10.1.3 The Percentile Method ....................... 160
      10.1.4 $BC_a$ Method ................................ 161
      10.1.5 Summary So Far, and More Examples .......... 163
  10.2 Bootstrapping Multivariate Data Sets ............ 165
      10.2.1 Regression Models and the Studentized Bootstrap
                        Method .................................. 166
  10.3 Exercises .......................................... 168
## Contents

**A Analysis using the SAS System**  
A.1 Table 2.1 .................................................. 169  
A.2 Example 2 .................................................. 170  
A.3 Example 3 .................................................. 170  
A.4 Example 5 .................................................. 171  
A.5 Example 6 .................................................. 171  
A.6 Example 7 .................................................. 172  
A.7 Example 8 .................................................. 172  
A.8 Example 9 .................................................. 173  
A.9 Table 3.4 .................................................. 173  
A.10 Tables 3.4 and 3.5 ....................................... 173  
A.11 Example 11 ............................................... 174  
A.12 Example 12 ............................................... 174  
A.13 Example 13 ............................................... 174  
A.14 Examples 14 and 15 ..................................... 174  
A.15 Example 16 ............................................... 175  
A.16 Example 23 ............................................... 175  

**Bibliography** ............................................. 177
Introduction

Preface

This book is intended to accompany a one-semester MS-level course in non-parametric statistics. Prerequisites for the course are calculus through multivariate Taylor series, elementary matrix algebra including matrix inversion, and a first course in frequentist statistical methods including some basic probability. Most of the techniques described in this book apply to data with only minimal restrictions placed on their probability distributions, but performance of these techniques, and the performance of analogous parametric procedures, depends on these probability distributions. The first chapter below reviews probability distributions. It also reviews some objectives of standard frequentist analyses. Chapters covering methods that have elementary parametric counterparts begin by reviewing those counterparts. These introductions are intended to give a common terminology for later comparisons with new methods, and are not intended to reflect the richness of standard statistical analysis, or to substitute for an intentional study of these techniques.

Computational Tools and Data Sources

Analyses used to illustrate techniques developed in this book will be facilitated using R Core Team (2018). This program may be downloaded for free from https://cran.r-project.org/. This course will heavily depend on the R package MultNonParam, installed by typing inside of R:

library("MultNonParam")

Other packages will be called as needed; if your system does not have these installed, install them as needed.
Acknowledgements

In addition to works referred to in following chapters, I consulted Stigler (1986) and Hald (1998) for early bibliographical references. Bibliographic trails have been tracked through documentation of software packages R and SAS, and bibliography resources from JSTOR, Citulike.org, Project Euclid, and various publisher’s web sites, have been used to construct the bibliography. I am grateful to the Rutgers Physical Sciences librarian Melanie Miller, the Rutgers Department of Statistics Administrative Assistant, and the work study students that she supervised, and the Rutgers Interlibrary Loan staff, for assistance in locating reference material.

I am grateful to my students at both Rutgers, and at the University of Rochester, to whom I taught this material over the years. My experience teaching this material helped me to select material for this volume, and to determine its level and scope. I consulted various text books during this time, including those of Hettmansperger and McKean (2011), Hettmansperger (1984), and Higgins (2004).
Background

Statistics is the solution to an inverse problem: given the outcome from a random process, the statistician infers aspects of the underlying probabilistic structure that generated the data. This chapter reviews some elementary aspects of probability, and then review some classical tools for inference about a distribution’s location parameter.

1.1 Probability background

In this section, I first review a tool for embedding a probability distribution into a larger family that allows for the distribution to be recentered and rescaled. I then review some important elementary probability distributions.

1.1.1 Location and Scale Families

All of these distributions generating potential data will be continuous. Most of them are or can be naturally extended to a family of distributions, called a location-scale family, by allowing an unknown location constant to shift the center of the distribution, and a second unknown scale constant to shrink or expand the scale. That is, suppose that \( X \) has density \( f(x) \) and cumulative distribution function \( F(x) \). Then \( a + bX \) has density \( f((y-a)/b)/b \) and the cumulative distribution function \( F((y-a)/b) \). If \( X \) has a standard distribution \( X \) with location and scale 0 and 1, then \( Y \) has location \( a \) and scale \( b \).

1.1.2 Probability distributions for Observations

Some common probability distributions are shown in Figure 1.1, constructed using \texttt{fun.comparedensityplot()} in the file \texttt{common.R}. The continuous distributions described below are distributions that might plausibly give rise to a data set of independent observations. The goal of this course is to be able to perform statistical inference on a data set without knowing the family from which it came. The behavior of various statistical procedures, including both standard parametric analyses, and nonparametric techniques forming the
subject of this course, may depend on the distribution generating the data, and knowledge of these families will be used to explore this behavior.

1.1.2.1 Gaussian Distribution

The normal distribution, or Gaussian distribution, has density

\[ f_G(x) = \exp\left(-\frac{(x - \theta)^2}{2\sigma^2}\right) / (\sigma \sqrt{2\pi}). \]

The parameter \( \theta \) is both the expectation and the median, and \( \sigma \) is the standard deviation. The cumulative distribution of the Gaussian distribution is

\[ F_G(x) = \int_{-\infty}^{x} f_G(y) \, dy. \]

There is no closed form for this integral. This distribution is symmetric about \( \theta \); that is, \( f_G(x) = f_G(2\theta - x) \), and \( F_G(x) = 1 - F_G(2\theta - x) \). This distribution is continuous. The specific member of this family of distributions with \( \theta = 0 \) and \( \sigma = 1 \) is called standard normal or standard Gaussian. The cumulative distribution function for the standard normal is denoted by \( \Phi(x) \). The distribution in its generality will be denoted by \( \mathcal{G}(\theta, \sigma^2) \).

This Gaussian distribution may be extended to higher dimensions; a multivariate Gaussian random variable, or multivariate normal random variable, \( \mathbf{X} \), in a space of dimension \( d \), has a density of form

\[ \exp\left(-\frac{(\mathbf{x} - \mu)^\top \mathbf{Y}^{-1} (\mathbf{x} - \mu)}{2}\right) \, \det \mathbf{Y}^{-1/2} (2\pi)^{-d/2}. \]

Here \( \mu \) is the expectation \( \mathbb{E}[\mathbf{X}] \), and \( \mathbf{Y} \) is the variance-covariance matrix \( \mathbb{E}[(\mathbf{X} - \mu)^\top (\mathbf{X} - \mu)] \).
1.1.2.2 Uniform Distribution

The uniform distribution has density

\[ f_U(x) = \begin{cases} \frac{1}{\lambda} & \text{for } x \in (\theta - \lambda/2, \theta + \lambda/2) \\ 0 & \text{otherwise} \end{cases} \]

The cumulative distribution function of this distribution is

\[ F_U(x) = \begin{cases} 0 & \text{for } x \leq \theta - \lambda/2 \\ \frac{x}{\lambda} + 1/2 - \frac{\theta}{\lambda} & \text{for } x \in (\theta - \lambda/2, \theta + \lambda/2), \\ 1 & \text{for } x \geq \theta + \lambda/2 \end{cases} \]

Again, the expectation and median for this distribution are both \( \theta \), and the distribution is symmetric about \( \theta \). The standard deviation is \( \lambda/\sqrt{12} \). Again, this distribution is continuous. A common canonical member of this family is the distribution uniform on \([0, 1]\), with \( \theta = 1/2 \) and \( \lambda = 1 \). The distribution in its generality will be denoted by \( U(\theta, \lambda) \).

1.1.2.3 Laplace Distribution

The double exponential distribution or Laplace distribution has density

\[ f_{La}(x) = \frac{\exp\left(-|x - \theta|/\sqrt{2}/\sigma\right)}{\sigma\sqrt{2}}. \]

The cumulative distribution function for this distribution is

\[ F_{La}(x) = \begin{cases} \exp\left((x - \theta)/\sqrt{2}/\sigma\right)/2 & \text{for } x \leq \theta \\ 1 - \exp\left(-(x - \theta)/\sqrt{2}/\sigma\right)/2 & \text{for } x > \theta \end{cases} \]

As before, the expectation and median of this distribution are both \( \theta \). The standard deviation of this distribution is \( \sigma \). The distribution is symmetric about \( \theta \). The distribution is continuous. A canonical member of this family is the one with \( \theta = 0 \) and \( \sigma = 1 \). The distribution in its generality will be denoted by \( La(\theta, \sigma^2) \).

1.1.2.4 Cauchy Distribution

The location-scale extension of the Cauchy distribution has density

\[ f_C(x) = \sigma^{-1}/(\pi(1 + (x - \theta)^2/\sigma^2)). \]

The cumulative distribution function for this this distribution is

\[ F_C(x) = 1/2 + \arctan((x - \theta)/\sigma)/\pi. \]

This distribution is symmetric about its median \( \theta \), but, unlike the Gaussian, uniform, and Laplace examples, does not have either an expectation nor a variance; the quantity \( \sigma \) represents not a standard deviation but a more general
Background scaling parameter. Its upper and lower quartiles are ±σ, and so the interquartile range is 2σ. This distribution is continuous. The member of the family with θ = 0 and σ = 1 is the one properly called Cauchy, and members with σ = 1 but θ ≠ 0 are called non-central Cauchy.

An interesting and important property of the Cauchy relates to the distribution of sums of independent and identical copies of members of this family. If X and Y are independent Cauchy, then Z = (X + Y)/2 is Cauchy. One may see this by first noting that P[Z ≤ z] = \int_{-\infty}^{\infty} \int_{-\infty}^{2z-y} f_C(x)f_C(y) \, dx \, dy, and hence that f_Z(z) = \int_{-\infty}^{\infty} f_C(2z-y)f_C(y) \, dy. Dwass (1985) evaluates this integral using partial fractions. Alternatively, this fact might be verified using characteristic functions. The distribution in its generality will be denoted by \mathcal{C}(\theta, \sigma^2).

1.1.2.5 Logistic Distribution

The logistic distribution has density

\[ f_L(x) = (1 + \exp(-(x-\theta)/\sigma))^{-1}(1 + \exp((x-\theta)/\sigma))^{-1}\sigma^{-1}. \]

This distribution is symmetric about \theta, and has finite mean and variance \theta and \sigma^2\pi^2/3. The cumulative distribution function of this distribution is

\[ F_L(x) = \exp((x-\theta)/\sigma)/(1 + \exp((x-\theta)). \]

1.1.2.6 Exponential Distribution

The exponential distribution has density

\[ f_E(x) = \begin{cases} \exp(-x) & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}. \]

One could embed the exponential distribution into a location-scale family, but because the lower endpoint of the support of the distribution often is fixed at zero by the structure of the application in which it is used, the location-scale family structure will not be employed here. The cumulative distribution function of this distribution is

\[ F_E(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 - \exp(-x) & \text{for } x > 0 \end{cases}. \]

The expectation is 1, and the median is \log(2). The inequality of these values is an indication of the asymmetry of the distribution. The standard deviation is 1.

1.1.3 Sampling distributions

The distributions presented above in §1.1.2 represent distributions of observations that might potentially be analyzed non-parametrically. Distributions in
this subsection will be used in this volume primarily as sampling distributions, or approximate sampling distributions, of test statistics.

1.1.3.1 Binomial Distribution

A discrete distributions will be of use, not for use with data plausibly arising from it, but because it will be used to generate the first of the nonparametric tests considered below. Consider the binomial distribution, supported on \( \{0, 1, \ldots, n\} \) for some integer \( n \). This distribution has additional parameter \( \pi \in [0, 1] \). Its probability mass function is

\[
\binom{n}{x} \pi^x (1 - \pi)^{n-x},
\]

and its cumulative distribution function is

\[
F_B(x) = \sum_{y=0}^{x} \binom{n}{y} \pi^y (1 - \pi)^{n-y}.
\]

Curiously enough, the binomial cumulative distribution function can be expressed in terms of the \( F \) distribution cumulative distribution function. The expectation is \( n\pi \), and the variance is \( n\pi(1 - \pi) \). The median does not have a closed-form expression. This distribution is symmetric only if \( \pi = 1/2 \).

An extension of the binomial distribution is the multinomial distribution, the distribution of counts of objects independently classified into categories according to certain probabilities.

1.1.4 \( \chi^2 \)-distribution

When \( X_1, \ldots, X_k \) are independent random variables, each with a standard Gaussian distribution (that is, Gaussian with expectation zero and variance one), then the distribution of the sum of their squares is called the chi-square distribution and is denoted \( \chi^2_k \). Here the index \( k \) is called the degrees of freedom.

When \( Y \) has a multivariate Gaussian distribution with dimension \( k \), expectation \( 0 \) and variance matrix \( \mathbf{Y} \), and if \( \mathbf{Y} \) has an inverse, then

\[
\mathbf{Y}^{-1} \mathbf{Y} \sim \chi^2_k. \tag{1.1}
\]

One can see this by noting that \( \mathbf{Y} \) may be written as \( \mathbf{\Theta} \mathbf{\Theta}^\top \). Then \( \mathbf{X} = \mathbf{\Theta}^{-1} \mathbf{Y} \) is multivariate Gaussian with expectation \( 0 \), and variance matrix \( \mathbf{\Theta}^{-1} \mathbf{\Theta} \mathbf{\Theta}^\top \mathbf{\Theta}^{-1} = \mathbf{I} \), where \( \mathbf{I} \) is the identity matrix with \( k \) rows and columns. Then \( \mathbf{X} \) is a vector of independent standard Gaussian variables, and

\[
\mathbf{Y}^\top \mathbf{Y} = \mathbf{X}^\top \mathbf{X}.
\]

Furthermore, still assuming

\[
X_i \sim \mathcal{N}(0, 1), \text{ independent} \tag{1.2}
\]

the distribution of

\[
W = \sum_{i=1}^{k} (X_i - \delta_i)^2 \tag{1.3}
\]
is called a non-central chi-square distribution (Johnson et al., 1995). The density and distribution function of this distribution are complicated. The most important property of this distribution is that it depends on $\delta_1, \ldots, \delta_k$ only through $\xi = \sum_{i=1}^k \delta_i^2$: this quantity is known as the non-centrality parameter, and the distribution will be denoted by $\chi_2^k(\xi)$. This dependence on nonzero expectations only through the simple non-centrality parameter may be seen by calculating the moment generating function of this distribution.

When $Y$ has a multivariate Gaussian distribution with dimension $k$, expectation $0$ and variance matrix $\Upsilon$, and if $\Upsilon$ has an inverse, then $X = \Theta^{-1}(Y - \mu)$ is multivariate Gaussian with expectation $-\Theta^{-1}\mu$, and variance matrix $I$. Hence

$$ (Y - \mu)^\top \Upsilon^{-1} (Y - \mu) \sim \chi_2^k(\xi) \text{ with } \xi = \mu^\top \Upsilon^{-1} \mu. \quad (1.4) $$

### 1.1.5 $T$-distribution

When $U$ has a standard Gaussian distribution, $V$ has a $\chi_2^k$ distribution, and $U$ and $V$ are independent, then $T = U/\sqrt{V/k}$ has a $T_k$ distribution, with $k$ called the degrees of freedom.

### 1.1.6 $F$-distribution

When $U$ and $V$ are independent random variables, with $\chi_2^k$ and $\chi_2^m$ distributions respectively, then $F = (U/k)/(V/m)$ is said to have a $F$ distribution with $k$ and $m$ degrees of freedom; denote this distribution by $f_{k,m}$. If a variable $T$ has a $T_m$ distribution, then $T^2$ has an $f_{1,m}$ distribution.

### 1.2 Elementary Tasks in Frequentist Inference

This section reviews elementary frequentist tasks of hypothesis testing and producing confidence intervals.

#### 1.2.1 Hypothesis Testing

Statisticians are often asked to choose among potential hypotheses about the mechanism generating a set of data. This choice is often phrased as between a null hypothesis, generally implying the absence of a potential effect, and alternative hypothesis, generally the presence of a potential effect. These hypotheses in turn are expressed in terms of sets of probability distributions, or equivalently, in terms of restrictions on a summary (such as the median) of a probability distribution. A hypothesis test is a rule that takes a data set
and returns “Reject the null hypothesis that $\theta = \theta^0$” or “Do not reject null hypothesis”.

In many applications studied in this manuscript, hypothesis tests are implemented by constructing a statistic $T$, depending on data, and a constant $c$, such that the statistician rejects the null hypothesis (1.6) if $T \geq t^\circ$, and fails to reject it otherwise. The constant $c$ is called a critical value, and the collection of data sets in

$$\{\text{data}|T(\text{data}) \geq t^\circ\}$$

(1.5)

for which the null hypothesis is rejected is called the critical region.

1.2.1.1 One-Sided Hypothesis Tests

For example, if data points represent the changes in some physiological measure after receiving some therapy, measured on subjects acting independently, then a null hypothesis might be that each of the changes in measurements comes from a distribution with median zero, and the alternative hypothesis might be that each of the changes in measurements comes from a distribution with median greater than zero. Symbolically, if $\theta$ represents the median of the distribution of changes, then the null hypothesis is $\theta = 0$, and the alternative is

$$\theta \geq 0.$$ 

(1.6)

This formulation is plausible if larger values of $\theta$ make larger values of $T$ more likely, and smaller values less likely. Such an alternative hypothesis is typically called a one-sided hypothesis. The null hypothesis might then be thought of as the set of all possible distributions for observations with median zero, and the alternative is the set of all possible distributions for observations with positive median.

Because null hypotheses are generally smaller in dimension than alternative hypotheses, frequencies of errors are generally easier control for null hypotheses than for alternative hypotheses.

Tests are constructed so that, in cases in which the null hypothesis is actually true, it is rejected with no more than a fixed probability. This probability is called the test level or type I error rate, and is commonly denoted $\alpha$.

Hence the critical value in such cases is defined by

$$P_{\theta^0} [T \geq t^\circ] = \alpha.$$ 

(1.7)

The other type of possible error occurs when the alternative hypothesis is true, but the null hypothesis is not rejected. The probability of erroneously failing to reject the null hypothesis is called the type II error rate, and is denoted $\beta$. More commonly, the behavior of the test under an alternative hypothesis is described in terms of the probability of a correct answer, rather than of an error; this probability is called power; power is $1 - \beta$.

One might attempt to control power as well, but, unfortunately, in the
common case in which an alternative hypothesis contains probability distributions arbitrarily close to those in the null hypothesis, type II error rate will come arbitrarily close to one minus the test level, which is quite large. Furthermore, once a particular distribution in the alternative hypothesis is selected, the smallest possible type II error rate is fixed, and cannot be independently controlled. Hence generally, tests are constructed primarily to control level. Under this paradigm, then, a test is constructed by specifying $\alpha$, constructing a test to have this type I error, and determining whether this test rejects the null hypothesis or fails to reject the null hypothesis.

In the one-sided alternative hypothesis formulation $\{\theta > \theta_0\}$, the investigator is, at least in principle, interested in detecting departures from the null hypothesis that are arbitrarily close to the null hypothesis. (The same observation will hold for two-sided tests in the next subsection). For planning purposes, however, investigators often pick a particular value within the alternative hypothesis. This particular value might be the minimal value of practical interest, or a value that other investigators have estimated. They then calculate the power at this alternative, to ensure that it is large enough to meet their needs. A power that is too small indicates that there is a substantial chance that the investigator’s alternative hypothesis is correct, but that they will fail to demonstrate it. Powers near 80% are typical targets.

Consider a test with a null hypothesis of form $\theta = \theta_0$ and an alternative hypothesis of form $\theta = \theta_A$, using a statistic $T$ such that under the null hypothesis $T \sim \mathcal{N}(\theta_0, \sigma_0^2)$, and under the alternative hypothesis $T \sim \mathcal{N}(\theta_A, \sigma_A^2)$. Test with a level $\alpha$, and without loss of generality assume that $\theta_A > \theta_0$. In this case, the critical value is approximately $c = \theta_0 + \sigma_0 z_\alpha$. Here $z_\alpha$ is the number such that $\Phi(z_\alpha) = 1 - \alpha$. A common level for such a one-sided test is $\alpha = 0.025$; $z_{0.025} = 1.96$. The power for such a one-sided test is

$$1 - \Phi((c - \theta_A)/\sigma_A) = 1 - \Phi((\theta_0 + \sigma_0 z_\alpha - \theta_A)/\sigma_A). \quad (1.8)$$

One might plan an experiment by substituting null hypothesis values $\theta_0$ and $\sigma_0$ and alternative hypothesis values $\theta_A$ and $\sigma_A$ into (1.8), and verifying that this power is high enough to meet the investigator’s needs; alternatively, might require power to be $1 - \beta$, solve for the effect size necessary to give this power

$$\theta_A = \theta_0 + \sigma_A z_\beta + \sigma_0 z_\alpha, \quad (1.9)$$

and ask whether this effect size is plausible. More commonly, $\sigma_0$ and $\sigma_A$ both may be made to depend on parameter representing sample size, with both decreasing to zero as sample size increases. Then (1.8) is solved for sample size.

### 1.2.1.2 Two-sided Hypothesis Tests

In contrast to one-sided alternatives, one often considers a two-sided hypothesis of for $\theta \neq \theta_0$, and, in this case, one often uses a test that rejects the null
hypothesis for data sets in 
\[ \{T \leq t_L^\circ\} \cup \{T \geq t_U^\circ\} \] 
(1.10)

In this case, there are two critical values, chosen so that
\[ \alpha = P_{\theta_0} [T \leq t_L^\circ \text{ or } T \geq t_U^\circ] = P_{\theta_0} [T \leq t_L^\circ] + P_{\theta_0} [T \geq t_U^\circ]. \] 
(1.11)

Many of the statistics constructed in this volume require an arbitrary choice by the analyst, which, if chosen otherwise, would flip the sign of \( T \), plus perhaps add a constant, and so, in order keep the analytical results invariant to this choice, the critical values are chosen to make the two final probabilities in (1.11) equal. Then the critical values solve the equations
\[ P_{\theta_0} [T \geq t_U^\circ] = \alpha/2, \quad \text{and} \quad P_{\theta_0} [T \leq t_L^\circ] = \alpha/2. \] 
(1.12)

Comparing with (1.7), the two-sided critical value is calculated exactly as is the one-sided critical value for an alternative hypothesis in the appropriate direction, and for a test level half that of the one-sided test. Similarly, the two-sided p-value is often constructed by doubling the one-sided value. Hence a two-sided test of level 0.05 is constructed in the same way as two one-sided tests of level 0.025.

Often the critical region implicit in (1.11) can be represented by creating a new statistic that is large when \( T \) is either large or small. That is, one might set \( W = |T - E_{\theta_0} [T]| \). In this case, if \( t_L^\circ \) and \( t_U^\circ \) are symmetric about \( E_{\theta_0} [T] \), then the two-sided critical region might be expressed as
\[ \{W \geq w^\circ\} \] 
(1.13)
for \( w^\circ = t_U^\circ - E_{\theta_0} [T] \). Alternatively, one might define \( W = |T - E_{\theta_0} [T]|^2 \), and, under the same symmetry condition, use as the critical region (1.13) for \( w^\circ = (t_U^\circ - E_{\theta_0} [T])^2 \). In the absence of such a symmetry condition, \( w^\circ \) may be calculated from the distribution of \( W \) directly.

I will refer to statistics \( T \) for which (1.10) is a reasonable critical region as inherently one-sided, since the two-sided test is constructed from one-sided tests combining evidence pointing in opposite directions, and will refer to statistics \( W \) for which (1.13) is a reasonable critical value for two-sided alternative as inherently two-sided.

Power for the two-sided test is the same probability as calculated in (1.11), with the \( \theta_A \) substituted for \( \theta_0 \), and power substituted for \( \alpha \). Again, assume that large values of \( \theta \) make larger values of \( T \) more likely. Then, for alternatives \( \theta_A \) greater than \( \theta_0 \), the first probability added in
\[ 1 - \beta = P_{\theta_A} [T \leq t_L^\circ] + P_{\theta_A} [T \geq t_U^\circ] \]
is quite small, and is typically ignored for power calculations. An additional reason for ignoring \( P_{\theta_A} [T \leq t_L^\circ] \) is that rejection of the null hypothesis because
the evidence is in the opposite direction of that anticipated will result in conclusions from the experiment not comparable to those for which the power calculation is constructed. Hence power for the two-sided test is generally approximated as the power for the one-sided test with level half that of the two-sided tests, and \( \alpha \) in (1.8) is often 0.025, corresponding to half the test level.

Some tests to be constructed in this volume may be expressed as \( W = \sum_{j=1}^{k} U_j^2 \) for variables \( U_j \) which are, under the null hypothesis, approximately standard normal and independent; furthermore, in such cases, the critical region for such tests is often of the form \( \{ W \geq w \} \). In such cases, the test of level \( \alpha \) rejects the null hypothesis when \( W \geq \chi^2_{k, \alpha} \), for \( \chi^2_{k, \alpha} \) the \( 1 - \alpha \) quantile of the \( \chi^2_k \) distribution. Note that if the standard normal approximation for the distribution of \( U_j \) is only approximately correct, then the resulting test will have level \( \alpha \) approximately, but not exactly.

If, under the alternative hypothesis, the variables \( U_j \) have expectations \( \mu_j \) and standard deviations \( \xi_j \), the alternative distribution of \( W \) will be a complicated weighted sum of \( \chi^2(\mu_j^2) \) variables. To a first approximation, however, the proportional impact of the move from the null distribution to the alternative distribution is much higher on the component expectations than on the standard deviations, and one might treat these alternative standard deviations fixed at 1. With this simplification, the sampling distribution of \( W \) under the alternative is \( \chi^2_k(\sum_{i=1}^{k} \delta_i^2) \), the non-central chi-square distribution.

1.2.1.3 \( P \)-values

Alternatively, one might calculate a test statistic, and determine the test level at which one transitions from rejecting to not rejecting the null hypothesis. This quantity is called a \( p \)-value. For a one-sided test with critical region of form (1.5), the \( p \)-value is given by

\[
P_0 [T \geq t_{\text{obs}}],
\]

for \( t_{\text{obs}} \) the observed value of the test statistic. For two-sided critical values of form (1.10), with condition (1.12), the \( p \)-value is given by

\[
2 \min(P_0 [T \geq t_{\text{obs}}], P_0 [T \leq t_{\text{obs}}]).
\]

These \( p \)-values are interpreted as leading to rejection of the null hypothesis when they are as low as or lower than the test level specified in advance by investigator before data collection.

Inferential procedures that highlight \( p \)-values are indicative of the inferential approach of Fisher (1925), while those that highlight pre-specified test levels and powers are indicative of the approach of Neyman and Pearson (1933). I refer readers to a thorough survey (Lehmann, 1993), and note here only that while I find the pre-specified test level arguments compelling, problematic examples leading to undesirable interpretations of \( p \)-values are rare using the techniques developed in this volume, and, more generally, the contrasts
between techniques advocated by these schools of thought are not central to the questions investigated in this volume.

1.2.2 Confidence Intervals

A confidence interval of level $1 - \alpha$ for parameter $\theta$ is defined as a set $(L, U)$ such that $L$ and $U$ depend on data, and such that for any $\theta$,

$$P_{\theta}[L < \theta < U] \geq 1 - \alpha. \quad (1.16)$$

The $\leq$ signs in (1.16) might be replaced by $<$ as appropriate.

The most general method for constructing a confidence interval is test inversion. For every possible null value $\theta^0$, find a test of the null hypothesis $\theta = \theta^0$, vs. the two-sided alternative, of level no larger than $\alpha$. Then the confidence set is

$$\{\theta^0 | \text{The null hypothesis } \theta = \theta^0 \text{ is not rejected.}\}. \quad (1.17)$$

Often (1.17) is an interval. In cases when (1.17) represents an interval, one attempts to determine the lower and upper bounds of the interval, either analytically or numerically.

Often, such tests are phrased in terms of a test statistic $W(\theta)$, such that the test rejects the null hypothesis $\theta = \theta^0$ if and only if $W(\theta^0) \geq w^*(\theta^0)$ for some critical value $c$ that might depend on the null hypothesis.

1.2.2.1 Test Inversion with Pivotal Statistics

A random quantity, generally involving an unknown parameter $\theta$, is called pivotal when its distribution that does not depend on $\theta$. Such a statistic is called a pivot. For instance, in the case of independent and identically distributed from a $\mathcal{G}(\theta, \sigma^2)$ distribution, then $T = (\bar{X} - \theta)/(s/\sqrt{n})$ has a $t$ distribution with $n - 1$ degrees of freedom, regardless of $\theta$. In this case, one may find quantiles $t^*_{L}$ and $t^*_{U}$ such that

$$P[t^*_{L} < T(\theta, \text{data}) < t^*_{U}] < 1 - \alpha. \quad (1.18)$$

Because the distribution of $\bar{X}$ is continuous, the probability in (1.18) is unchanged whether the inequalities in the event are $<$ or $\leq$, and using $<$ allows the elimination of any boundary points. Then

$$\{\theta | t^*_{L} < T(\theta, \text{data}) < t^*_{U}\} \quad (1.19)$$

is a confidence interval, if it is really an interval. In the case when (1.19) is an interval, and when $T(\theta, \text{data})$ is continuous in $\theta$, then the interval is of the form $(L, U)$; that is, the interval does not include the endpoints.
1.2.2.2 A Problematic Example

One should use this test inversion technique with care, as the following problematic case shows. Suppose that $X$ and $Y$ are normal random variables, with expectations $\mu$ and $\nu$ respectively, and common known variances $\sigma^2$. Suppose that one desires a confidence interval for $\rho = \mu/\nu$ (Fieller, 1954).

The quantity $T = \sqrt{n}(\bar{X} - \rho\bar{Y})/(\sigma\sqrt{1 + \rho^2})$ has a standard normal distribution, independent of $\rho$, and hence is pivotal. A confidence region is $\{\rho : n(\bar{X} - \rho\bar{Y})^2/\sigma^2(1 + \rho^2)) \leq z_{\alpha/2}^2\}$. Equivalently, the region is

$$\{\rho | Q(\rho) < 0\}$$

for $Q(\rho) = (X^2 - \nu^2)(X^2 - \nu^2 - \rho X^2 + Y^2 - \nu^2)/(X^2 - \nu^2)$ (1.20)

for $\nu = \sigma z_{\alpha/2}$.

If $X^2 + Y^2 < \nu^2$, then the quadratic form in (1.20) has a negative coefficient for $\rho^2$, and the maximum value is at $\rho = XY/(X^2 - \nu^2)$. The maximum is $(v^2 (X^2 + Y^2))/\nu^2 - X^2 < 0$, and so the inequality in (1.20) holds for all $\rho$, and the confidence interval is the entire real line.

If $X^2 + Y^2 > \nu^2 > X^2$, then the quadratic form in (1.20) has a negative coefficient for $\rho^2$, and the maximum is positive. Hence values satisfying the inequality in (1.20) are very large and very small values of $\rho$; that is, the confidence interval is

$$(-\infty, -XY - \nu\sqrt{X^2 + Y^2 - \nu^2}) \cup (-XY + \nu\sqrt{X^2 + Y^2 - \nu^2}, \infty).$$

If $X^2 > \nu^2$, then the quadratic form in (1.20) has a positive coefficient for $\rho^2$, and the minimum is negative. Then the values of $\rho$ satisfying the inequality in (1.20) are those near the minimizer $XY/(X^2 - \nu^2)$. Hence the interval is

$$\left(\frac{XY - \nu\sqrt{X^2 + Y^2 - \nu^2}}{X^2 - \nu^2}, \frac{XY + \nu\sqrt{X^2 + Y^2 - \nu^2}}{X^2 - \nu^2}\right).$$

1.2.2.3 P-value Inversion

A slightly less general method of constructing confidence intervals is through tail probability inversion. This technique works in cases without a pivotal statistic. Suppose that one can find a univariate statistic $W$ whose distribution depends on the unknown parameter $\theta$, such that potential one-sided $p$-values $P_\theta[W \leq w]$ and $P_\theta[W \geq w]$ are monotonic in $\theta$ for each potential statistic value $w$. Typical applications have

$$P_\theta[W \geq w] \text{ nondecreasing in } \theta, \ P_\theta[W \leq w] \text{ non-increasing in } \theta, \forall w$$

(1.21)

Let $w$ be the observed value of $W$. Under monotonicity,

$$\{\theta | P_\theta[W \geq w] > \alpha/2, P_\theta[W \leq w] > \alpha/2\}$$
is an interval, and under (1.21), it is of form \((\theta^L, \theta^U)\) with endpoints satisfying

\[
P_{\theta^L}[W \geq w] = \alpha/2, P_{\theta^U}[W \leq w] = \alpha/2. \tag{1.22}
\]

There may be \(w\) such that the equation \(P_{\theta^L}[W \geq w] = \alpha/2\) has no solution, because \(P_{\theta}[W \geq w] > \alpha/2\) for all \(\theta\). In such cases, take \(\theta^L\) to be the lower bound on possible values for \(\theta\). For example, if \(\theta \in [0, 1]\), and \(W \sim \text{Bin}(n, \theta)\), then \(P_{\theta}[W \geq 0] = 1 > \alpha/2\) for all \(\theta\), \(P_{\theta}[W \geq w] = \alpha/2\) has no solution, and \(\theta^L = 0\). Similarly, there may be \(w\) such that the equation \(P_{\theta^U}[W \leq w] = \alpha/2\) has no solution, because \(P_{\theta}[W \leq w] > \alpha/2\) for all \(\theta\). In such cases, take \(\theta^U\) to be the upper bound on possible values for \(\theta\).

Construction of intervals for the binomial proportion represents a simple example in which \(p\)-values may be inverted (Clopper and Pearson, 1934).

### 1.3 Exercises

1. Demonstrate that the moment generating function for the statistic (1.3), under (1.2), depends on \(\delta_1, \ldots, \delta_k\) only through \(\sum_{j=1}^{k} \delta_j^2\).
One-Sample Nonparametric Inference

This chapter first reviews standard normal-theory inference on one sample location models. It then presents motivation for why a distribution-free approach to location testing is necessary, and presents non-parametric techniques for inference on quantiles. Later in this chapter, techniques for comparing the efficiencies of tests are introduced, and these are applied to various parametric and non-parametric tests. Finally, techniques for estimating a single cumulative distribution function are discussed.

2.1 Parametric Inference on Means

Suppose one want to learn about $\theta = \mathbb{E}[X_j]$, from a sample $X_1, \ldots, X_j, \ldots, X_n$ of independent and identically distributed random variables. When one knows the parametric family generating a set of independent data, this information may be used to construct testing and estimation methods tailored to the individual distribution. The variety of such techniques is so large that only those presuming approximately a Gaussian model will be reviewed in this volume. Hence in what follows, parametric analyses for comparison purposes will be taken to assume approximate Gaussian distributions.

If these variables have a finite variance $\sigma^2$, then the Central Limit Theorem (CLT) ensures that $\bar{X} = \frac{\sum_{j=1}^{n} X_j}{n}$ is approximately $\mathcal{N}(\theta, \sigma^2/n)$. To test the null hypothesis $\theta = \theta^0$ vs. the alternative $\theta > \theta^0$, reject the null hypothesis if $\bar{X} > \theta^0 + z_{\alpha} \sigma / \sqrt{n}$. To test the null hypothesis $\theta = \theta^0$ vs. the two-sided alternative $\theta \neq \theta^0$, reject the null hypothesis if $\bar{X} > \theta^0 + z_{\alpha/2} \sigma / \sqrt{n}$, or if $\bar{X} < \theta^0 - z_{\alpha/2} \sigma / \sqrt{n}$. If $\sigma$ is not known, substitute the estimate $s = \sqrt{\frac{\sum_{j=1}^{n} (X_j - \bar{X})^2}{n-1}}$, and compare this quantity to the $t$ distribution with $n-1$ degrees of freedom.
TABLE 2.1: True Size for the $T$ Test, and Sign Test, and Exact Sign Test, Nominal level 0.05

(a) Sample size 10, Two-Sided

<table>
<thead>
<tr>
<th></th>
<th>Gaussian</th>
<th>Cauchy</th>
<th>Laplace</th>
<th>Uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>0.05028</td>
<td>0.01879</td>
<td>0.04098</td>
<td>0.05382</td>
</tr>
<tr>
<td>Sign</td>
<td>0.02127</td>
<td>0.02166</td>
<td>0.02165</td>
<td>0.02060</td>
</tr>
<tr>
<td>Exact</td>
<td>0.02127</td>
<td>0.02166</td>
<td>0.02165</td>
<td>0.02060</td>
</tr>
</tbody>
</table>

(b) Sample size 17, Two-Sided

<table>
<thead>
<tr>
<th></th>
<th>Gaussian</th>
<th>Cauchy</th>
<th>Laplace</th>
<th>Uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>0.05017</td>
<td>0.02003</td>
<td>0.04593</td>
<td>0.05247</td>
</tr>
<tr>
<td>Sign</td>
<td>0.01234</td>
<td>0.01299</td>
<td>0.01274</td>
<td>0.01310</td>
</tr>
<tr>
<td>Exact</td>
<td>0.04847</td>
<td>0.04860</td>
<td>0.04871</td>
<td>0.04898</td>
</tr>
</tbody>
</table>

(c) Sample size 40, Two-Sided

<table>
<thead>
<tr>
<th></th>
<th>Gaussian</th>
<th>Cauchy</th>
<th>Laplace</th>
<th>Uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>0.04938</td>
<td>0.02023</td>
<td>0.04722</td>
<td>0.05029</td>
</tr>
<tr>
<td>Sign</td>
<td>0.03915</td>
<td>0.03952</td>
<td>0.03892</td>
<td>0.03904</td>
</tr>
<tr>
<td>Exact</td>
<td>0.03915</td>
<td>0.03952</td>
<td>0.03892</td>
<td>0.03904</td>
</tr>
</tbody>
</table>

2.2 The Need for Distribution-Free Tests

Table 2.1 contains actual test levels for some tests of location parameters for four of the families described in S1.1.2. This table was calculated using the R commands:

```r
library("VGAM") # vgam gives laplace distribution
library("BSDA") # Library gives z test
source("http://stat.rutgers.edu/home/kolassa/960-555/common.R")
# Below the shift by -.5 for the uniform centers it at 0.
size<-fun.comparepower(sampsz=c(10,17,40),nsamp=100000,
  dist=list("rnorm","rcauchy","rlaplace","runif"),
  hypoth=c(0,0,0,-.5),altd=c("two.sided","greater"))
print(size[,1,])
```

Sizes were determined via simulation; a large number of samples were drawn from each of the distributions under the null hypothesis, the specified test statistic was calculated, and the proportion of times the null hypothesis was rejected is tabulated. For now, restrict attention to the first line in each subtable, corresponding to the $T$ test. Null hypotheses in Table 2.1 are in terms
of the median of the distribution. The \( T \) test, however, is appropriate for hypotheses involving the expectation. In the normal, Laplace, and uniform cases, the median coincides with the expectation, and so standard asymptotic theory justifies the use of the \( T \) test. In the Cauchy example, as noted before, no expectation exists, and the \( T \) test is inappropriate. However, generally, data analysts do not have sufficient information to distinguish the Cauchy example from the set of distributions having enough moments to justify the \( T \) test, and so it is important to study the implications of such an inappropriate use of methodology.

For both sample sizes, observations from a normal distribution give the targeted level, as expected. Observations from the Laplace distribution give a level close to the targeted level. Observations from the Cauchy distribution give a level much smaller than the targeted level, which is paradoxical, because one might expect heavy tails to make it anti-conservative. Fig. 2.1 shows the density resulting from studentizing the average of independent Cauchy variables. This figure was generated using

```r
source("http://stat.rutgers.edu/kolassa/home/960-555/common.R")
fun.studentizedcaucyplot(10,10000)
```

FIGURE 2.1: Symmetrized Density of Studentized Cauchy, Sample Size 10

The resulting density is bimodal, with tails lighter than one would otherwise expect. This shows that larger values of the sample standard deviation in the denominator of the Studentized statistic act more strongly than larger values of components of the average in the numerator.
One-Sample Nonparametric Inference

In all cases above, the T test succeeds in providing a test level not much larger than the target nominal level. On the other hand, in some cases the true level is significantly below that expected. This effect decreases as sample level increases.

2.3 One-sample median methods

For moderate sample level, then, the standard one-sample $t$-test fails to control test level as the distribution of summands changes. Inference about a population median rather than the population mean avoids this problem.

Techniques in this section can be traced to Arbuthnott (1712), as described in the example below. Fisher (1930) treats this test as too obvious to require comment.

Recall that the median $\theta$ of random variable $X_j$ is defined so that

$$
P[X_j \geq \theta] \geq \frac{1}{2}, \quad P[X_j \leq \theta] \geq \frac{1}{2}.
$$

Consider independent identically distributed random variables $X_i$ for $i = 1, \ldots, n$. To test whether a putative median value $\theta^0$ is the true value, define new random variables

$$
Y_j = \begin{cases} 
  1 & \text{if } X_j - \theta^0 \leq 0 \\
  0 & \text{if } X_j - \theta^0 > 0
\end{cases}.
$$

(2.1)

Then under $H_0: \theta = \theta^0$, $Y_j \sim B(1/2, 1)$. This logic only works if

$$
P[X_j = \theta^0] = 0; \quad (2.2)
$$

assume this. It is usually easier to assess this continuity assumption than it is for distributional assumptions. Then the median inference problem reduces to one of binomial testing. Let $T(\theta^0) = \sum_{j=1}^{n} Y_j$ be the number of observations less than or equal to $\theta^0$. Pick $a$ and $b$ so that $\sum_{j=b}^{b-1} (1/2)^n \binom{n}{j} \geq 1 - \alpha$. One might choose $a$ and $b$ symmetrically, so that $a$ is the largest $a$ value such that

$$
\sum_{j=0}^{a-1} (1/2)^n \binom{n}{j} < \alpha/2; \quad (2.3)
$$

That is, $a$ is that potential value for $T$ such that less than $\alpha/2$ probability sits below it. The largest such $a$ has probability at least $1 - \alpha/2$ equal to or larger than it, and at least $\alpha/2$ at or below it; hence $a$ is the $\alpha/2$ quantile of the $\text{Bin}(n, 1/2)$ distribution, and satisfies $a \approx n/2 - \sqrt{n}/2z_{1-\alpha/2}$. Symmetrically, one might choose the smallest $b$ so that

$$
\sum_{j=b}^{n} (1/2)^n \binom{n}{j} < \alpha/2; \quad (2.4)
$$
TABLE 2.2: Exact levels and exact and asymptotic lower critical values for symmetric two-sided binomial tests of nominal level 0.05

<table>
<thead>
<tr>
<th>n</th>
<th>Exact</th>
<th>Asymptotic</th>
<th>Exact</th>
<th>Asymptotic</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0</td>
<td>0.0313</td>
<td>24</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0.0156</td>
<td>25</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0.0078</td>
<td>26</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.0391</td>
<td>27</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.0215</td>
<td>28</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0.0117</td>
<td>29</td>
<td>8</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>0.0386</td>
<td>30</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>0.0225</td>
<td>31</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>0.0129</td>
<td>32</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>0.0352</td>
<td>33</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>0.0213</td>
<td>34</td>
<td>10</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>0.0490</td>
<td>35</td>
<td>11</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>0.0309</td>
<td>36</td>
<td>11</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>0.0192</td>
<td>37</td>
<td>12</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>0.0414</td>
<td>38</td>
<td>12</td>
</tr>
<tr>
<td>21</td>
<td>5</td>
<td>0.0266</td>
<td>39</td>
<td>12</td>
</tr>
<tr>
<td>22</td>
<td>5</td>
<td>0.0169</td>
<td>40</td>
<td>13</td>
</tr>
<tr>
<td>23</td>
<td>6</td>
<td>0.0347</td>
<td>41</td>
<td>13</td>
</tr>
</tbody>
</table>

\(n + 1 - b\) is the \(\alpha/2\) quantile of the Bin\((n, 1/2)\) distribution, in this case \(b = n/2 + \sqrt{n}/2z_{1-\alpha/2}\). Then, reject null if \(T < a\) or \(T \geq b\). This test is called the sign test, or the binomial test (Higgins, 2004).

Again, direct attention to Table 2.1. Both variants of the sign test succeed in keeping test level no larger than the nominal value. However, the sign test variants, because of the discreteness of the binomial distribution, in some cases achieve levels much smaller than the nominal target. Subtable (a), for sample size 10, is the most extreme example of this; subtable (b), for sample size 17, represents the most moderate reduction in actual sample size, and subtable (c), for sample size 40, is intermediate. Note further that, while the asymptotic sign test, based on the normal approximation, is not identical to the exact version, for subtables (a) and (c) the levels coincide exactly, since whether \(p\)-values exceed 0.05 are the same for both. Subtable (b) exhibits a case in which for one data value, the exact and approximate sign tests disagree on whether \(p\)-values exceed 0.05.

Table 2.2 presents characteristics of the exact two-sided binomial test of the null hypothesis that the probability of success is half, with level \(\alpha = 0.05\) applied to small samples. In this case, the two-sided \(p\)-value is obtained by doubling the one-sided \(p\) value.

For small samples \((n < 6)\), the smallest one-sided \(p\)-value, \(1/2^n\), is greater than .025, and the null hypothesis is never rejected. Such small samples are not reflected in Table 2.2. This table consists of two subtables side by side, for
The power of the sign test is determined by $P_{\theta^A}[X_j \leq \theta^0]$ for values of $\theta^A \neq \theta^0$. Since $\theta^A > \theta^0$ if $P_{\theta^A}[X_j \leq \theta^0] < 1/2$, alternatives $\theta^A > \theta^0$ correspond to one sided alternatives $P[Y_j = 1] < 1/2$.

If $\theta^0$ is med $[X_j]$, the true population median of the $X_j$, and if there exists a set of form $(\theta^0 - \epsilon, \theta^0 + \epsilon)$, with $\epsilon > 0$, such that $P[X_j \in (\theta^0 - \epsilon, \theta^0 + \epsilon)] = 0$, then any other $\theta$ in this set is also a population median for $X_j$, and hence the test will have power against such alternatives no larger than the test level. Such occurrences are rare.

Table 2.3 represents powers for these various tests for various sample levels. The alternative is chosen to make the $T$ test have power approximately .80 for the normal and Laplace distributions, using (1.9). In this case both $\sigma_0$ and $\sigma_A$ for the normal and Laplace distributions are $1/\sqrt{n}$. Formula (1.9) is inappropriate for the Cauchy distribution, since in this case the $T$ statistic does not have a distribution that is approximately normal. For the Cauchy distribution, the same alternative as for the normal and Laplace distributions is used.

Results in Table 2.3 show that for a sample size for which the sign test level approximates the nominal level (viz., $n = 17$), use of the sign test for normal data results in moderate a loss in power relative to the $t$ test, while use of the sign test results in a moderate gain in power for Laplace observations, and in a substantial gain in power for Cauchy observations.

Example 1 An early (and very simple) application of this test was to test whether the proportion of boys born in a given year is the same as the proportion of girls born that year (Arbuthnott, 1712). Number of births were determined for a period of 82 years. Let $X_j$ represent the number of births of girls, minus the number of births of boys, in year $j$. The parameter $\theta$ represents the median amount by which the number of girls exceeds the number of boys; its null value is 0. Let $Y_j$ take the value 0 for years in which more girls than boys are born, and 1 otherwise. Note that in this case, (2.2) is violated, but $P[X_j = 0]$ is small, and this violation is not important. Test at level 0.05.

The critical values for this test are $a = 32$ and $b = 51$, obtained as the 0.025 and 0.975 quantiles of the binomial distribution with 82 trials and success probability .5. Reject the null hypothesis if $T < 32$ or if $T \geq 51$. (The asymmetry in the treatment of the lower and upper critical values
Confidence Intervals for the Median

TABLE 2.3: Power for the T Test, and Sign Test, and Exact Sign Test, Nominal level 0.05

(a) Sample size 10, Two-Sided

<table>
<thead>
<tr>
<th></th>
<th>Gaussian</th>
<th>Cauchy</th>
<th>Laplace</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T</strong></td>
<td>0.70593</td>
<td>0.14345</td>
<td>0.73700</td>
</tr>
<tr>
<td><strong>Sign</strong></td>
<td>0.41772</td>
<td>0.20506</td>
<td>0.57222</td>
</tr>
<tr>
<td><strong>Exact</strong></td>
<td>0.41772</td>
<td>0.20506</td>
<td>0.57222</td>
</tr>
</tbody>
</table>

(b) Sample size 17, Two-Sided

<table>
<thead>
<tr>
<th></th>
<th>Gaussian</th>
<th>Cauchy</th>
<th>Laplace</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T</strong></td>
<td>0.74886</td>
<td>0.10946</td>
<td>0.76456</td>
</tr>
<tr>
<td><strong>Sign</strong></td>
<td>0.35747</td>
<td>0.17954</td>
<td>0.58984</td>
</tr>
<tr>
<td><strong>Exact</strong></td>
<td>0.57893</td>
<td>0.35759</td>
<td>0.79011</td>
</tr>
</tbody>
</table>

(c) Sample size 40, Two-Sided

<table>
<thead>
<tr>
<th></th>
<th>Gaussian</th>
<th>Cauchy</th>
<th>Laplace</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T</strong></td>
<td>0.78152</td>
<td>0.06307</td>
<td>0.78562</td>
</tr>
<tr>
<td><strong>Sign</strong></td>
<td>0.55462</td>
<td>0.35561</td>
<td>0.84331</td>
</tr>
<tr>
<td><strong>Exact</strong></td>
<td>0.55462</td>
<td>0.35561</td>
<td>0.84331</td>
</tr>
</tbody>
</table>

is intentional, and is the result of the asymmetry in the definition of the distribution function for discrete variables.

In each of these years $X_j < 0$, and so $Y_j = 1$, and $T = 82$. Reject the null hypothesis of equal proportion of births. The original analysis of this data presented what is now consider the \( p \)-value; the one sided value of (1.14) is trivially \( P\left( T \geq 82 \right) = \left( \frac{1}{2} \right)^{82} \), which is tiny. The two-sided \( p \)-value of (1.15) is \( 2 \times \left( \frac{1}{2} \right)^{82} = \left( \frac{1}{2} \right)^{81} \), which is still tiny.

2.4 Confidence Intervals for the Median

Apply the test inversion approach of §1.2 to the sign test that rejects $H_0 : \theta = \theta^0$ if fewer than $t_l$ or at least $t_u$ data points are less than or equal to $\theta^0$. Let $X_{(1)}$ referring to the data values after ordering. When $\theta^0 \leq X_{(1)}$, then $T(\theta^0) = 0$. For $\theta^0 \in (X_{(1)}, X_{(2)})$, $T(\theta^0) = 1$. For $\theta^0 \in (X_{(2)}, X_{(3)})$, $T(\theta^0) = 2$. In each case, the \( ( \) at the first end of the interval and the \( ] \) at the end of the interval arises (2.1), because observations that are exactly equal to $\theta^0$ are coded as one. Hence the test rejects $H_0$ if $\theta^0 \leq X_{(t_l)}$ or $\theta^0 > X_{(t_u)}$, and, for
any $\theta^0$,
\[
P_{\theta^0} [X_{(t_l)} < \theta^0 \leq X_{(t_u)}] \geq 1 - \alpha. \quad (2.5)
\]
This equation leads to the confidence interval is $(X_{(t_l)}, X_{(t_u)})$. However, since the data have a continuous distribution, then $X_{(t_u)}$ also has a continuous distribution, and
\[
P_{\theta^0} [\theta^0 = X_{(t_u)}] = 0
\]
for any $\theta^0$. Hence $P_{\theta^0} [X_{(t_l)} < \theta^0 < X_{(t_u)}] \geq 1 - \alpha$, and one might exclude the upper end point, to obtain the interval $(X_{(t_l)}, X_{(t_u)})$.

**Example 2** Consider data from
\[http://lib.stat.cmu.edu/datasets/Arsenic\]
from a pilot study on the uptake of arsenic from drinking water. Column six of this file gives arsenic concentrations in toenail clippings, in parts per million. The data are in a Word file; a plain text version is given in
\[http://stat.rutgers.edu/home/kolassa/arsenic.dat\].

Sorted nail arsenic values are
\[0.073, 0.080, 0.099, 0.105, 0.118, 0.118, 0.119, 0.135, 0.141, 0.158, 0.175, 0.269, 0.275, 0.277, 0.310, 0.358, 0.433, 0.517, 0.832, 0.851, 2.252.\]

We determine $a$ in (2.3). We construct a confidence interval for the (natural) log of toenail arsenic. The sign test statistic has a $\text{Bin}(21, .5)$ distribution under the null hypothesis. We choose the largest $t_l$ such that $P_0 [T < t_l] \leq \alpha/2$. The first few terms in (2.3) are
\[4.76 \times 10^{-7}, 1.00 \times 10^{-5}, 1.00 \times 10^{-4}, 6.34 \times 10^{-4}, 2.85 \times 10^{-3}, 9.70 \times 10^{-3}, 2.59 \times 10^{-2}, 5.54 \times 10^{-2}, 9.70 \times 10^{-2}, 1.40 \times 10^{-1},\]

and cumulative probabilities are
\[4.77 \times 10^{-7}, 1.05 \times 10^{-5}, 1.11 \times 10^{-4}, 7.45 \times 10^{-4}, 3.60 \times 10^{-3}, 1.33 \times 10^{-2}, 3.91 \times 10^{-2}, 9.46 \times 10^{-2}, 1.92 \times 10^{-1}, 3.32 \times 10^{-1}.\]

The largest of these cumulative sums smaller than 0.025 is the sixth, corresponding to $T < 6$. Hence $t_l = 6$. Similarly, $t_u = 16$. Reject the null hypothesis that the mean is 0.26 if $T < 6$ or if $T \geq 16$. Note that 10 of the observations are greater than the null median 0.26. Hence $T = 10$. Do not reject the null hypothesis.

Alternatively, one might calculate a $p$-value. Using (1.15), the $p$-value is $2 \min(P_0 [T \geq 10], P_0 [T \leq 10]) = 1.$
Furthermore, the confidence interval for the median is \((X_{(6)}, X_{(16)}) = (0.118, 0.358)\).

The values \(t_l\) and \(t_u\) may be calculated in R by

\[
a < \text{qbinom}(0.025, 21, .5); \quad b < n + 1 - \text{qbinom}(0.025, 21, .5)
\]

and the ensemble of calculations might also have been performed in R using

\[
arcesn <- \text{as.data.frame(scan('arsenic.dat', what=list(age=0, sex=0, drink=0, cook=0, water=0, nails=0)))}
\]

\[
\text{library("BSDA")#Gives sign test.}
\]

\[
\text{SIGN.test(arsenic$nails, md=0.26)#Argument md gives null hyp.}
\]

Figure 2.2 exhibits construction of the confidence interval in the previous example; I apply these techniques on the log scale. The confidence interval is the set of log medians that yield a test statistic for which the null hypothesis is not rejected. Values of the statistic for which the null hypothesis is not reject are between the horizontal lines; log medians in the confidence intervals are values of the test statistic within this region. Fig. 2.2 demonstrates geometrically the construction of such an interval for a data set, in this case of size...
n = 10. First, order statistics (that is, the ordered values) are plotted on the horizontal axis, with the place in the ordered data set on the vertical axis. Next, draw horizontal lines at the values \( t_l \) and \( t_u \), given by (2.3) and (2.4) respectively. Finally, draw vertical lines through the data points that these horizontal lines hit.

For this particular example, the exact two-sided binomial test of level 0.05 rejects the null hypothesis that the event probability is half if the sum of event indicators is 0 or 1: \( t_l = 2 \). For \( Y_j \) of (2.1), the sum is 0 or 1 for all \( \theta \) to the left of the point marked \( X_{(t_l)} \). Similarly, this test rejects the null hypothesis if the sum of event indicators is 9 or 10: \( t_u = 9 \). The sum of the \( Y_j \) is 9 or 10 for \( \theta \) to the right of the point marked \( X_{(t_u)} \).

By symmetry, one might expect \( t_l = n - t_u \), but this is not the case. The asymmetry in definitions (2.3) and (2.4) arises because construction of the confidence interval requires counting not the data points, but the \( n \) asymmetry in definitions (2.3) and (2.4) arises because construction of the confidence interval requires counting not the data points, but the \( n \) spaces between them, plus the regions below the minimum and above the maximum, for a total of \( n + 1 \) ranges. Then \( t_l = n + 1 - t_u \).

This interval is not of the usual form \( \theta \pm 2\hat{\sigma} \), for \( \hat{\sigma} \) with a factor of \( 1/\sqrt{n} \). Cramér (1946, pp. 368f.) shows that if \( X_1, \ldots, X_n \) is a set of independent random variables, each having density \( f \), then \( \text{Var} [\text{sample median} \{X_1, \ldots, X_n\}] \approx 1/(4f(\theta)^2n) \). Chapter 8 investigates estimation of this density; this estimate can be used to estimate the median variance, but density estimation is harder than the earlier confidence interval rule.

### 2.4.1 Inference for other percentiles

The quantile \( \theta \) corresponding to probability \( \tau \) is defined by \( P[\theta | X_j \leq \theta] = \tau \). Suppose that \( \theta \) is quantile \( \tau \in (0, 1) \) of distribution of independent and identically distributed continuous random variables \( X_1, \ldots, X_n \). Then \( P[\theta | X_j \leq \theta] = \tau \), and one can produce a generalized sign test. Define the null and alternative hypotheses \( H_0: \theta = \theta^0 \) and \( H_A: \theta \neq \theta^0 \). As before, \( T(\theta) \) is the number of observations smaller than or equal to \( \theta \). For the true value \( \theta \) of the quantile, \( T \sim \text{Bin}(n, \tau) \). Choose \( t_l \) and \( t_u \) so that \( \sum_{j=t_l}^{n-1} \tau^j (1-\tau)^{n-j} \binom{n}{j} \geq 1 - \alpha \). Often, one chooses the largest \( t_l \) and smallest \( t_u \) so that

\[
\sum_{j=0}^{t_l-1} \tau^j (1-\tau)^{n-j} \binom{n}{j} < \alpha/2, \quad \sum_{j=t_u}^{n} \tau^j (1-\tau)^{n-j} \binom{n}{j} < \alpha/2 \tag{2.6}
\]

this \( t_l \) is \( \alpha/2 \) quantile of the \( \text{Bin}(n, \tau) \) distribution, and \( n + 1 - t_u \) is the \( \alpha/2 \) quantile of the \( \text{Bin}(n, 1-\tau) \) distribution. Hence \( t_l \approx n\tau - \sqrt{n\tau(1-\tau)z_{\alpha/2}} \) and \( t_u \approx n\tau + \sqrt{n\tau(1-\tau)z_{\alpha/2}} \). One then rejects \( H_0 \) if \( T < t_l \) or \( T \geq t_u \).

This test is then inverted to obtain \((X_{(t_l)}, X_{(t_u)})\) as the confidence interval for \( \theta \). Note that confidence level is conservative: \( P [X_{(t_l)} \leq \theta \leq X_{(t_u)}] = 1 - P [X_{(t_l)} \geq \theta] - P [\theta \geq X_{(t_u)}] \geq 1 - \alpha \). For any given \( \theta \), inequality is generally strict.
Example 3 Test the null hypothesis that the upper quartile (that is, the .75 quantile) of the arsenic nail data from Example 2 is the reference value 0.26, and give a confidence interval for this quantile. The analysis is the same as before, except that $t_l$ and $t_u$ are different. We determine $t_l$ and $t_u$ in (2.6). By direct calculation, or using the R commands
\[
a = \text{qbinom}(0.025, 21, 0.75); b = 21 + 1 - \text{qbinom}(0.025, 21, 1 - 0.75),
\]
we see $t_l = 12$ and $t_u = 20$. Since $T = 10 < t_l$, reject the null hypothesis that the upper quartile is 0.26. Furthermore, the confidence interval is the region between the twelfth and twentieth ordered values, $(X_{12}, X_{20}) = (0.269, 0.851)$. With data present in the R workspace, one calculates a confidence interval as
\[
\text{sort(arsenic$nails)[c(a,b)]}
\]
and the p-value as
\[
\text{tt} = 10^2 \times \min(c(pbinom(tt, 21, .75), pbinom(21 + 1 - tt, 21, 1 - 0.75))
\]
to give 0.0128.

Dependence of the test statistic $T(\theta)$ on $\theta$ is relatively simple. Later inversions of more complicated statistics will make use of the simplifying device of first, shifting all or part of the data by subtracting $\theta$, and then testing the null hypothesis that the location parameter for this shifted variable is zero.

2.5 Comparing Tests

For fixed level, alternative, and power, a test with a smaller sample size is better. Consider two families of one-sided tests, indexed by sample size, using statistics $T_1$ and $T_2$, both with test level $\alpha$, and determine the sample sizes required to give power $1 - \beta$, for the same alternative. Compare the tests by taking ratio of these two sample sizes. The ratio is called relative efficiency; the notation dates back at least as far as Noether (1950), citing Pitman (1948).

Let $c_{j,n}$ represent the critical value for test $j$ based on $n$ observations; that is, the test based on statistic $T_j$ using $n$ observations, rejects the null hypothesis if $T_j \geq c_{j,n}$. Hence $c_{j,n}$ satisfies $P_{\theta_0}[T_j \geq c_{j,n}] = \alpha$. Let $\varpi_{j,n}(\theta^A)$ represent the power for test $T_j$ using $n$ observations, under the alternative $\theta^A$:

$$\varpi_{j,n}(\theta^A) = P_{\theta^A}[T_j \geq c_{j,n}].$$  \hspace{1cm} (2.7)

Assume that

$\varpi_{j,n}(\theta^A)$ is continuous and increasing in $\theta^A$ for all $j, n$,

$$\lim_{\theta^A \to -\infty} \varpi_{j,n}(\theta^A) = 1, \quad \lim_{n \to \infty} \varpi_{j,n}(\theta^A) = 1 \text{ for all } \theta^A > \theta_0.$$  \hspace{1cm} (2.8)
Two tests, tests 1 and 2, involving hypotheses about a parameter \( \theta \), taking the value \( \theta^0 \) under the null hypothesis, and with a simple alternative hypothesis of form \( \{ \theta^A \} \), for some \( \theta^A > \theta^0 \), with similar level and power, will be compared. Pick a test level \( \alpha \) and a power \( 1 - \beta \), and the sample size \( n_1 \) for test 1. The power and level conditions on \( T_1 \) imply a value for \( \theta^A \) under the alternative hypothesis; that is, \( \theta^A \) solves \( P_{\theta^A} [T_1 \geq c_{1,n_1}] = 1 - \beta \); note that \( \theta^A \) is a function of \( n_1, \alpha, \) and \( \beta \). Under conditions (2.8), one can determine the minimal value of \( n_2 \) so that test 2 has power at least \( 1 - \beta \), under the alternative given by \( \theta^A \). Report \( n_1 / n_2 \) as the relative efficiency of test 2 to test 1; this depends on \( n_1, \alpha, \) and \( \beta \).

One can remove dependence on \( n_1 \) by considering the asymptotic relative efficiency \( ARE_{\alpha,\beta}[T_1, T_2] = \lim_{n_1 \to \infty} n_1 / n_2 \), when this limit exists.

This measure comparing efficiencies of two tests takes on a particularly easy form in a special, yet common, case, in which both statistics are asymptotically normal. In this case, the relative efficiency can be approximated in terms of standard deviations and derivatives of means under alternative hypotheses. General approximations for sample size, power, and effect sizes are investigated first; these are applied to relative efficiency later.

### 2.5.1 Power, Sample Size, and Effect Size

This subsection presents formulas for power, sample size, and effect size, that may be used for efficiency comparisons, but are also useful on their own.

#### 2.5.1.1 Power

Consider tests satisfying

\[
T_j \sim G(\mu_j(\theta), \varsigma_j^2(\theta)), \text{ for } \varsigma_j(\theta) > 0, \mu_j(\theta) \text{ increasing in } \theta.
\]  

The normal distribution in (2.9) does not need to hold exactly; approximate normality is sufficient. Without loss of generality, take \( \theta^0 = 0 \). In this case, one can find the critical values for the two tests, \( c_{j,n_j} \), such that \( P_0 [T_j \geq c_{j,n_j}] = \alpha \). Since \( (T_j - \mu_j(0))/\varsigma_j(0) \) is approximately standard normal under the null hypothesis,

\[
\alpha = P_0 [(T_j - \mu_j(0))/\varsigma_j(0) \geq z_\alpha] = P_0 [T_j \geq \mu_j(0) + \varsigma_j(0)z_\alpha].
\]

Hence \( c_{j,n_j} = \mu_j(0) + \varsigma_j(0)z_\alpha \). The power for test \( j \) is approximately

\[
\varpi_{j,n_j}(\theta^A) \approx P_{\theta^A} [T_j \geq \mu_j(0) + \varsigma_j(0)z_\alpha]
\]

\[
= 1 - \Phi \left( \left[ \mu_j(0) + \varsigma_j(0)z_\alpha - \mu_j(\theta^A) \right] / \varsigma_j(\theta^A) \right). \quad (2.10)
\]

Often the variance of the test statistic changes slowly as one moves away from the null hypothesis; in this case, the power for test \( j \) is approximately

\[
\varpi_{j,n_j}(\theta^A) \approx 1 - \Phi \left( \left[ \mu_j(0) - \mu_j(\theta^A) \right] / \varsigma_j(0) + z_\alpha \right). \quad (2.11)
\]
Comparing Tests

2.5.1.2 Sample and Effect Sizes

When the variance of the test statistic decreases in a regular way with sample size, one can invert the power relationship to determine the sample size needed for a given power and effect size. Consider tests satisfying, in addition to (2.9),

\[ \sigma_j^2(\theta) = \frac{1}{n_j}. \]  

(2.12)

Then

\[ \varpi_{j,n}(\theta^A) = 1 - \Phi \left( \sqrt{n_j} \left[ \mu_j(0) + \frac{\sigma_j(0)z_\alpha}{\sqrt{n_j}} - \mu_j(\theta^A) \right] / \sigma_j(\theta^A) \right). \]  

(2.13)

Suppose further that \( \mu_j(\theta), \sigma_j(\theta) \) differentiable on some set \( \theta \in (-\epsilon, \epsilon) \).

(2.14)

(These conditions are somewhat simpler than considered by Noether (1950); in particular, note that (2.9), (2.12), and (2.14) together are not enough to demonstrate the second condition of (2.8).) Without loss of generality, take \( \theta^0 = 0 \). In this case, one can find the critical values for the two tests, \( c_{j,n_j} \), such that \( P_T = \alpha \). The power expression (2.10) may be simplified by approximating the mean and variances at the alternative hypothesis by quantities at the null. For large \( n_j \), alternatives with power less than 1 will have alternative hypotheses near the null, and so \( \sigma_j(\theta^A) \approx \sigma_j(\theta^0) \). Approximation by the leading term alone cannot be applied to the mean, since it would remove all of the effect of the difference between null and alternative. Hence

\[ \varpi_{j,n}(\theta^A) \approx 1 - \Phi \left( \sqrt{n_j} \left[ \mu_j(0) - \mu_j(\theta^A) \right] / \sigma_j(0) \right) \]

\[ = \Phi \left( \sqrt{n_j} \left[ \mu_j(\theta^A) - \mu_j(0) \right] / \sigma_j(0) \right) - z_\alpha. \]  

(2.15)

This expression for approximate power may be solved for sample size, by noting that if \( \varpi_{j,n}(\theta^A) = 1 - \beta \), then \( \varpi_{j,n}(\theta^A) = \Phi(z_\beta) \), and (2.15) holds if

\[ \sqrt{n_j} \left[ \frac{\mu_j(\theta^A) - \mu_j(0)}{\sigma_j(0)} \right] = z_\alpha = z_\beta, \]  

or

\[ n_j = \sigma_j(0)^2 \left( z_\alpha + z_\beta \right)^2 / (\mu_j(\theta^A) - \mu_j(0))^2. \]  

(2.16)

Common values for \( \alpha \) and \( \beta \) are 0.025 and 0.2, giving upper Gaussian quantiles of \( z_\alpha = 1.96 \) and \( z_\beta = 0.84 \). Recall that \( z \) with a subscript strictly between 0 and 1 indicates that value for which a standard Gaussian random variable has that probability above it.

It may be of use in practice, and will be essential in the efficiency calculations below, to approximate which member of the alternative hypothesis corresponds with a test of a given power, with sample size held fixed. Solving
(2.16) exactly for $\theta^A$ is difficult, since the function $\mu$ is generally non-linear. Approximating this function using a one-term Taylor approximation,

$$
\mu_j(\theta^A) \approx \mu_j(0) + \mu'_j(0)\theta^A.
$$

(2.17)

The power for test $j$ is approximately

$$
1 - \Phi \left( \left( \sigma_j(0)z_\alpha - \sqrt{m_j}e_j\theta^A \right)/\sigma_j(0) \right) = 1 - \Phi \left( z_\alpha - \sqrt{m_j}e_j\theta^A \right)
$$

for $e_j = \mu'_j(0)/\sigma(0)$. Setting this power to $1 - \beta$, $z_\alpha - \sqrt{m_j}e_j\theta^A = z_{1-\beta}$. Solving this equation for $\theta^A$,

$$
\theta^A \approx (z_\alpha + z_{1-\beta})/\sqrt{m_j}e_j,
$$

(2.18)

verifying requirement that $\theta^A$ get close to zero. This expression can be used to approximate an effect size needed to obtain a certain power with a certain sample size and test level, and will be used in the context of asymptotic relative efficiency.

2.5.2 Efficiency Calculations

Equating the alternative hypothesis parameter values (2.18) corresponding to power $1 - \beta$, then

$$
\left( z_\alpha - z_{1-\beta} \right)/\sqrt{n_1}e_1 = \left( z_\alpha - z_{1-\beta} \right)/\sqrt{n_2}e_2,
$$

or

$$
\text{ARE}_{\alpha,\beta}[T_1, T_2] = n_2/n_1 = e_1^2/e_2^2.
$$

Note that this relative efficiency doesn't depend on $\alpha$ or $\beta$, or on $n_1$. As an example, suppose $X_1, \ldots, X_n$ are independent observations from a symmetric distribution with finite variance $\rho^2$ and mean $\theta$. Then $\theta$ is also the median of these observations. Compare tests $T_1$, the $t$-test, and $T_2$, the sign test. Then $T_1$ has a distribution depending on the distribution of $X_j$, and $T_2$ has a binomial distribution. Note that $T_1$ has approximately a standard normal distribution for large $n_1$. That is, $T_1 \sim \mathcal{N}(\theta/\rho, 1/n_1)$, and

$$
\mu'_1(0) = 1/\rho, \quad \sigma_1(0) = 1, \quad \text{and} \quad e_1 = 1/\rho.
$$

(2.19)

On the other hand, $T_2 \sim \mathcal{B}(\mu_2(\theta), \sigma_2(\theta)^2/n_2)$ for

$$
\mu_2(\theta) = F(\theta), \quad \sigma_2(\theta) = \sqrt{F(\theta)(1-F(\theta))}.
$$

(2.20)

Hence $\mu'_2(0) = f(0)$, and $\sigma_2(0) = 1/2$, and $e_2 = 2f(0)$.

The asymptotic relative efficiencies of these statistics depends on the distribution that generates the data. If data come from $\mathcal{N}(\theta, \rho^2)$, then $\mu'_1(0) = 1/(\sqrt{2}\pi\rho)$, $\sigma_2(0) = 1/2$, and $e_2 = \sqrt{2}/\pi/\rho$. Then $n_1/n_2 = (2/\sqrt{2}\pi)^2 = 2/\pi$. Hence, as expected, the $t$-test is more powerful; the sign test requires more than $50\%$ more observations to obtain the same power against the same alternative.

If the data come from a Laplace distribution, then $\rho = 1$, since the Laplace
Comparing Tests

Table 2.4: Efficacies for one-sample location tests

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\mu'(0)$</th>
<th>$\sigma(0)$</th>
<th>$e$</th>
<th>Relative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>$1/\rho$</td>
<td>$1/\sqrt{2\pi\rho}$</td>
<td>$1$</td>
<td>$1/\rho$</td>
</tr>
<tr>
<td>Laplace</td>
<td>$1/\sqrt{2/\pi\rho}$</td>
<td>$1$</td>
<td>$\sqrt{2/\pi\rho}$</td>
<td></td>
</tr>
<tr>
<td>Cauchy</td>
<td>$1/\pi$</td>
<td>$1/2$</td>
<td>$2/\pi$</td>
<td>$0$</td>
</tr>
</tbody>
</table>

distribution has variance 1. Substituting into (2.19), $\mu'_1(0) = 1$, $\sigma_1(0) = 1$, and $e_1 = 1$. Also $\mu'_2(0) = 1/\sqrt{2}$, $\sigma_2(0) = 1/2$, and $e_2 = \sqrt{2}$. Hence $n_1/n_2 = (\sqrt{2})^2 = 2$; in this case, the sign test is more powerful, requiring roughly half the sample size as does the $t$-test.

Now suppose these data come from a Cauchy distribution shifted to have point of symmetry $\theta$. In this case, the expectation of the distribution does not exist, the standard deviation $\rho$ is infinite, and asymptotic normality does not hold. In fact, the distribution of the the mean of Cauchy random variables is again a Cauchy random variable, with no change in the spread of the distribution. Plugging into the definition of efficacy, without worrying about regularity conditions, gives $\mu'_1(0) = 1$, $\sigma_1(0) = \infty$, and $e_1 = 0$. On the other hand, the quantities for the sign test are

$\mu'_2(0) = \pi^{-1}$, $\sigma_2(0) = 1/2$, and $e_2 = 2/\pi$.

Hence, for Cauchy responses, the efficiency of the sign test relative to the $t$-test is $n_1/n_2 = \infty$.

Table 2.4 summarizes these calculations.

2.5.3 Examples of Power Calculations

The normal approximation to power (2.10) was derived assuming only (2.9), approximate normality for the test statistics at the null and alternative, without yet using (2.14), and so (2.10) to approximate power.

Example 4 In this example I calculate power for a sign test applied to 49 observations from a Gaussian distribution with unit variance. Suppose $X_1, \ldots, X_{49} \sim \mathcal{N}(\theta, 1)$, with null hypothesis $\theta = 0$ and alternative hypothesis $\theta = 1/2$. The sign test statistic, divided by $n$, approximately satisfies (2.9) and (2.12), with $\mu$ and $\sigma$ given by (2.20). Then $\mu_1(0) = .5$, $\sigma_1(0) = \sqrt{5} \times .5 = .5$, $\mu_1(.5) = 0.691$, $\sigma_1(.5) = \sqrt{0.691 \times 0.309} = 0.462$, 0.462,
and power for a one-sided test of level 0.025, or a two-sided test of level 0.05, is approximated by (2.10): \(1 - \Phi(7 \times (0.5 + 0.972203937)/0.462) = 1 - \Phi(-0.772) = 0.780\). The null and alternative standard deviations are close enough to motivate the use of the simpler approximation (2.15), approximating power as \(1 - \Phi(7 \times (0.5 - 0.691)/0.5 + 1.96) = 1 - \Phi(-0.714) = 0.769\).

If, instead, a test of power 0.85 were desired for alternative expectation \(1/2\), with a one-sided test of level 0.025, \(z_\alpha = 1.96\), and \(z_\beta = 1.036\). From (2.16), one needs at least \(n = (0.462)^2(1.96 + 1.056)^2/(0.691 - 0.5)^2 = 53.22\) observations; choose 54.

Finally, one might determine how large of an effect one might detect using the original 49 observations with a test of level 0.025 and power 0.85. one could use \(e = \sqrt{2/\pi} = 0.797\), from the box in Table 2.4 specific to the sign test and the Gaussian distribution. Expression (2.18) gives this number as \((1.96 + 1.036)/(7 \times 0.797) = 0.537\).

2.6 Distribution function estimation

Suppose one wishes to estimate a common distribution function of \(X_1, \ldots, X_n\), identically distributed independent variables. For \(x\) in the range of \(X_j\), let \(\hat{F}(x)\) be the number of data points less than or equal to \(x\), divided by \(n\). Since the observations are independent, \(\hat{F}(x) \sim \text{Bin}(n, F(x))\). A confidence interval for \(F(x)\) is

\[
\hat{F}(x) \pm z_{\alpha/2} \sqrt{\hat{F}(x)(1 - \hat{F}(x))/n};
\]

exact versions also exist. The above intervals will extend outside \([0, 1]\), which is not reasonable; this can be circumvented by transforming the probability scale. Note further that the confidence bounds in Fig. 2.3 exhibit occurrences of larger estimates being associated with upper confidence bounds that are smaller (ex., in Fig. 2.3, the region between the second-to-largest and the largest observations), and for the region with the cumulative distribution function estimated at zero or one (that is, the region below the smallest observed value, and the region above the largest observed value), confidence limits lie on top of the estimates, indicating no uncertainty. Both of these phenomena are unrealistic, and neither can be repaired through rescaling. A preferred solution is to substitute the intervals of Clopper and Pearson (1934), described in §1.2.2.3, to avoid all three of these problems (viz., bounds outside \([0, 1]\), bounds ordered differently than the estimate, and bounds with zero variability). Such intervals are exhibited in Fig. 2.4.

Finally, the confidence associated with these bounds is point-wise, and not simultaneous. That is, if \((L_1, U_1)\) and \((L_2, U_2)\) and are \(1 - \alpha\) confidence bounds associated with two ordinates \(x_1\) and \(x_2\), then \(P[L_1 \leq F(x_1) \leq U_1] \geq 1 - \alpha\).
and $P[L_2 \leq F(x_2) \leq U_2] \geq 1 - \alpha$, at least approximately, but the preceding argument does not bound $P[L_1 \leq F(x_1) \leq U_1$ and $L_2 \leq F(x_2) \leq U_2]$ any higher than $1 - 2\alpha$.

**Example 5** Consider the arsenic data of Example 2. For every $x$ a real number, one count the number of data points less than this $x$. For any $x$ less than the smallest value 0.073, this estimate is $\hat{F}(x) = 0$. For $x$ greater than or equal to this smallest value and smaller than the next smallest value 0.080, the estimate is $\hat{F}(x) = 1/21$. This data set contains one duplicate value 0.118. For values below, but close to, 0.118 (for example, $x = 0.1179$), $\hat{F}(x) = 4/21$, since 21 of the observations are less than $x$. However, $\hat{F}(x) = 6/21$; this the jump here is twice what it is at other data values, since there are two observations here. This estimate is sketched in both Figs. 2.3 and 2.4, and may be constructed in R using `ecdf(arsenic$nails)`, presuming the data of example 2 is still present to R. The command `ecdfcis(data,exact=FALSE)` or `ecdfcis(data,exact=TRUE)` to add confidence bounds.

**FIGURE 2.3:** Empirical CDF and Confidence Bounds for Arsenic in Nails
2.7 Exercises

1. Calculate the asymptotic relative efficiency for the sign statistic relative to the one-sample t-test (which you should approximate using the one-sample $z$-test). Do this for observations from the
   a. uniform distribution, on $[-1/2, 1/2]$ with variance $1/12$ and mean under the null hypothesis of 0, and
   b. the logistic distribution, symmetric about 0, with variance $\pi^2/3$ and density $\exp(x)/(1 + \exp(x))^2$.

2. The data set

   http://ftp.uni-bayreuth.de/math/statlib/datasets/lupus

gives data on 87 lupus patients. The third column gives duration, and the fourth column gives transformed disease duration. Give a 90% confidence interval for the median duration, through inverting the sign test, and compare this to the normal theory interval for the mean. Keep in mind that the normal theory and sign test approach are only comparable if you can argue that the mean and the median for the distribution are plausibly the same. Comment on this.
3. The data set

http://lib.stat.cmu.edu/datasets/bodyfat

gives data on body fat in 252 men. The second column gives proportion of lean body tissue. Give a 95% confidence interval for upper quartile proportion of lean body tissue. Note that the first 116 lines and last 10 lines are data set description, and should be deleted. (Line 117 is blank, and should also be deleted).

4. Suppose 49 observations are drawn from a Cauchy distribution, displaced to have location parameter 1. What is the power of the sign test at level 0.05 to test the null hypothesis of expectation zero for these observations?
In this chapter, we address the question of two-sample testing. Data will generally consist of observations \( Y_1, \ldots, Y_{M_2} \) from a continuous distribution function \( G \), and observations \( X_1, \ldots, X_{M_1} \) from continuous distribution function \( F \). We begin a consideration of two independent samples, and model these distributions as identical, up to some known shift \( \theta \); that is,

\[
F(z) = G(z - \theta) \quad \forall z.
\]  

(3.1)

Techniques for testing null hypothesis of form \( \theta = \theta^0 \) in (3.1), vs. the alternative that (3.1) holds for some alternative \( \theta \neq \theta^0 \), and for estimating \( \theta \) assuming (3.1), are presented. Techniques for tests in which (3.1) is the null hypothesis, for an unspecified \( \theta \), are also presented.

3.1 Two-Sample Approximately Normal Inference

Two-sample normal-theory inference is primarily concerned about means; however, one might also compare other aspects of distributions. Under the assumption of approximate normality, the only additional aspect of the distributions to be compared is their dispersion.

3.1.1 Two-Sample Approximately Normal Inference on Means

If the observations \( X_1, \ldots, X_{M_1} \) and \( Y_1, \ldots, Y_{M_2} \) are approximately normally distributed, one might to use the two-sample pooled \( t \) test for inference on \( \theta \). That is, the test statistic is

\[
T(\theta) = (\bar{Y} - \bar{X} - \theta) / \sqrt{s^2 \left( \frac{1}{M_2} + \frac{1}{M_1} \right)},
\]  

(3.2)

for

\[
S^2_X = \frac{\sum_{i=1}^{M_1} (X_i - \bar{X})^2}{M_1 - 1}, \quad S^2_Y = \frac{\sum_{i=1}^{M_2} (Y_i - \bar{Y})^2}{M_2 - 1}
\]  

(3.3)
and
$$s_p^2 = \frac{(M_1 - 1)S_X^2 + (M_2 - 1)S_Y^2}{N - 2},$$
for $N = M_1 + M_2$.

Statistic (3.2) is called the two-sample pooled $t$ statistic, and the associated test is the two-sample pooled $t$ test. When $\theta$ is correctly specified,
$$T(\theta) \sim T_{N-2},$$
and so the standard test of size $\alpha$ rejects the null hypothesis $\theta = \theta^0$ in (3.1) when
$$|T(\theta)| \geq T_{N-2,\alpha/2},$$
for $T(\theta)$ of (3.2).

Estimates of $\theta$, under the assumption (3.1), may be constructed by setting $T(\hat{\theta}) = 0$; that is, $\hat{\theta} = \hat{Y} - \hat{X}$. Confidence intervals are generally constructed by inverting the pooled two-sample $t$ test (3.2) and (3.5) to obtain the interval $\{\theta | |T| \leq T_{N-2,\alpha/2}\} = \hat{Y} - \hat{X} \pm s_pT_{N-2,\alpha/2}\sqrt{1/M_1 + 1/M_2}$. The counterpart of (3.2), allowing for a variety of null hypotheses about $\theta$, is $T = (\bar{Y} - \bar{X} - \theta)/\sqrt{s_p^2(1/M_2 + 1/M_1)}$, yielding the confidence interval $\bar{Y} - \bar{X} \pm T_{N-2,\alpha/2}s_p\sqrt{(1/M_2 + 1/M_1)}$.

### 3.1.2 Two-Sample Approximately Normal Inference on Distributions

One might consider the formerly-alternative hypothesis (3.1), with $\theta$ unspecified, as a null hypothesis. Under the Gaussian hypothesis, (3.1) fails to hold only if the variances of the distributions are unequal. A common parametric approach for comparing variances of Gaussian variables, is to compare the separate variance estimates. Under the model (3.1),
$$T = S_Y^2/S_X^2 \sim F_{M_2-1,M_1-1}$$
for $S_X^2$ and $S_Y^2$ of (3.3) (Fisher, 1925, p. 808), although Fisher (1930) recommended transforming this ratio by taking logs to obtain an approximately normal test statistic. A simpler test, with $E[X]$ and $E[Y]$ known, is also available (Fisher, 1926).

Pearson (1931) notes that, under the hypothesis of equal variances, the log of differences in in estimated standard deviations is a monotonic transformation of $(1 + M_2T/M_1)^{-1}$, and that this quantity follows a Pearson I family; he further considers the test arising from comparing this statistic to this exact null sampling distribution. Fisher (1973), crediting Snedecor (1934), recommends comparing $T$ of (3.6) to the $F_{M_2-1,M_1-1}$ distribution.

When the underlying distribution of $X_1, \ldots, X_{M_1}$ and $Y_1, \ldots, Y_{M_2}$ is not exactly normal, the distributional result in (3.4) and in (3.6) are approximate rather than exact. This approximation of (3.6) was observed to be poor for
even moderate deviations from normality, in cases when (3.4) remains entirely adequate (Pearson, 1931). Testing for equality of dispersion is revisited in §3.9, and in some sense a nonparametric dispersion test is of more urgency than the test of location.

3.2 General Two-Sample Rank Tests

Nonparametric alternatives to the two-sample pooled \( t \) test, to be developed in this chapter, will reduce to rank tests of the form

\[
T_G^{(k)} = \sum_{j=1}^{N} a_j I_j^{(k)}, \tag{3.7}
\]

for \( I_j^{(k)} \) equal to the 1 if the item ranked \( j \) in the combined sample comes from the group \( k \), and 0 otherwise. The superscript in \( I_j^{(k)} \) refers to group, and does not represent power.

The statistic \( T_G^{(2)} \) is designed to take on large values when items in group two are generally larger than the remainder of the observations (that is, the items in group one), and to take small values when items in group two are generally smaller than the remainder of the observations. The statistic \( T_G^{(1)} \) is designed to take on large values when items in group one are generally larger than the remainder of the observations (that is, the items in group two), and to take small values when items in group one are generally smaller than the remainder of the observations. The statistic \( T_G^{(1)} \) provides no information not also captured in \( T_G^{(2)} \), since \( T_G^{(2)} = \sum_{j=1}^{N} a_j - T_G^{(1)} \).

3.2.1 Null Distributions of General Rank Statistics

In order to use this statistic in a hypothesis tests, one first needs to know its distribution under the null hypothesis. That is, one needs first to be able to calculate one-sided \( p \)-values of form \( P_0 \left[ T_G^{(2)} \geq t \right] \), for \( T_G^{(2)} \) of (3.7) and various values of \( t \), with the subscript 0 on the probability indicating calculation under the null hypothesis. One might calculate \( p \)-values for these test statistics exactly: List all \( \binom{N}{M_2} \) possible ways to divide the \( N \) objects into two groups, one of size \( M_1 \). Calculate the test statistic for each rearrangement. Count the number of rearrangements giving a test statistic as extreme or more extreme than the one we observe. Divide this count by \( \binom{N}{M_2} \) to get \( p \)-value. If there exist different rankings giving the same value for \( T_G^{(2)} \), this information might be used to simplify calculations, but in general one cannot depend on such a simplification. Hence these calculations are likely quite slow. A common
simplification is to round the scores so that the scores become integer multiples of a common value; these simplified scores may then be more amenable to exact analysis.

More commonly, the distribution of rank statistics is approximated by the normal distribution. Since $T_G^{(2)}$ is the sum of random variables that are neither identically distributed nor independent, a variant of the usual central limit theorem due to Erdős and Réyni (1959), specific to finite population sampling without replacement, is used. There are some conditions on the set of scores needed to ensure approximate normality.

### 3.2.2 Moments of Rank Statistics

The normal approximation to the distribution of $T_G^{(1)}$ and $T_G^{(2)}$ requires calculation of the null expectation and variance of the statistic. This subsection determines moments for a statistic of the form (3.7). Under the null hypothesis, these scores with $M_1$ assigned to the first group and $M_2$ assigned to the second group, with all rearrangements having equal probability. The scores for the $Y$ group may be thought of as sampled without replacement from a finite population. Let $\bar{a} = \sum_{k=1}^{N} a_k/N$. In this case, $E[T_G^{(k)}] = \sum_{j=1}^{N} a_j E[I_j^{(k)}]$, and $E[I_j^{(k)}] = M_k/N$. Hence

$$E_0[T_G^{(k)}] = M_k \sum_{j=1}^{N} a_j/N = M_k \bar{a}; \quad (3.8)$$

the subscript 0 on the expectation operator indicates that the expectation is taken under the null hypothesis.

Calculating the variance is harder, since the $I_i^{(k)}$ are not independent. Let $\hat{a} = \sum_{k=1}^{N} a_k^2/N$. Let $b_1 = \text{Var}[I_j^{(k)}] = M_k(N - M_k)/N^2$. When $j \neq i$, then

$$E[I_j^{(k)} I_i^{(k)}] = P[I_j^{(k)} = 1, I_i^{(k)} = 1] = P[I_j^{(k)} = 1|I_i^{(k)} = 1] P[I_i^{(k)} = 1] = \frac{M_k - 1}{N-1} \frac{M_k}{N}.$$ Let

$$b_2 = \text{Cov}[I_2^{(k)}, I_1^{(k)}] = \frac{M_k(M_k - 1)}{N(N-1)} - \frac{M_k^2}{N^2} = -\frac{M_1 M_k}{N^2(N-1)}.$$ So

$$\text{Var}[T_G^{(k)}] = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \text{Cov}[I_i^{(k)}, I_j^{(k)}]$$

$$= \sum_{i=1}^{N} a_i^2 b_1 + \sum_{i \neq j} a_i a_j b_2 = (b_1 - b_2) N \hat{a} + N^2 b_2 \hat{a}^2$$

$$= (N - M_k)M_k(\hat{a} - \hat{a}^2)/(N - 1). \quad (3.9)$$
TABLE 3.1: Reduction of Two-Sample Testing Problem to Fisher’s Exact Test, via Mood’s Test, for an even total sample size

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>X</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater than Median</td>
<td>A</td>
<td>N/2</td>
<td></td>
</tr>
<tr>
<td>Less than Median</td>
<td>B</td>
<td>N/2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n</td>
<td>M1</td>
<td>N</td>
</tr>
</tbody>
</table>

3.3 A First Distribution-Free Test

The nonparametric approach analogous to sign test is Mood’s Median Test. One first calculates combined sample median. Let A be number of observations from Y’s above the combined median. Let B be number of observations from Y’s below the combined median. Under the null hypothesis \( F(x) = G(x) \), and when \( N \) is even, \( A \) has a hypergeometric distribution, and Mood’s test reduces the two-sample equality of distribution problem to Fisher’s exact test, as exhibited in Table 3.1. Mood’s test \( T_M \) is equivalent to the score test \( T_G^{(2)} \) of (3.7) with

\[
a_j = \begin{cases} 
1 & \text{for } j \geq (N + 1)/2 \\
0 & \text{for } j = (N + 1)/2 \\
-1 & \text{for } j \leq (N + 1)/2 
\end{cases}
\]

although, as originally formulated, this test was applied only in the case of even sample sizes, and so the score 0 would not be used.

Mood’s test, and other rank tests, ignore the ordering of the observations from the first group among themselves, and similarly ignore the orderings of the second group among themselves. Represent the data set as a vector of \( N \) symbols, \( M_1 \) of them X and \( M_2 \) of them Y. The letter X in position \( j \) indicates that, after ordering combined set of \( N \) observations, the observation ranked \( j \) comes from the first group, and the letter Y in position \( j \) indicates that the observation ranked \( j \) comes from the second group. The advantage of Mood’s test lies in its simplicity, and its disadvantage is its low power. To see why its power is low, consider the test with \( M_1 = M_2 = 3 \), for a total of six observations. With \( M_1 = M_2 = 3 \), then \( X, Y, X, Y, X, Y \) indicates that the lowest observation is from the first group, the second lowest is from the second group, the third lowest is from the first group, the fourth lowest is from the second group, the fifth lowest (that is, the second highest) is from the first group, and the highest is from the second group. Mood’s test treats \( X, Y, X, Y, X, Y \) and \( X, Y, X, X, Y, Y \) as having equal evidence against \( H_0 \), but the second should be treated as having more evidence. Furthermore, Mood’s test takes a value between 0 and \( \min(M_2, \lfloor (N)/2 \rfloor) \). This high degree of discreteness in the statistic’s support undermines power.

Westenberg (1948) presented the equal-sample case of the statistic, and
TABLE 3.2: Reduction of Two-Sample Testing Problem to Fisher’s Exact Test, via Mood’s Test, for an odd total sample size

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>X</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater than Median</td>
<td>A</td>
<td>(N − 1)/2</td>
<td></td>
</tr>
<tr>
<td>Equal to Median</td>
<td>C</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Less than Median</td>
<td>B</td>
<td>(N − 1)/2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>M2</td>
<td>M1</td>
<td>N</td>
</tr>
</tbody>
</table>

Mood (1950) detailed the use in the case with an even combined sample size. Mood’s test is of sufficiently low importance in applications that the early references did not bother to present the slight complication that arises when the combined sample size is odd.

When the total sample size is odd, one might represent the median test as inference from a $2 \times 3$ contingency table with ordered categories, as in Table 3.2. Then $T_M = A - B$. Then the one-sided $p$-values may be calculated as

$$P_0 [A - B \geq t| C = 0] + P_0 [A - B \geq t| C = 1] \frac{M_2}{N}.$$ 

The probabilities $P_0 [A - B \geq t| C = c]$ are calculated from the hypergeometric distribution.

The null expectation and variance of $T_M$ are given by (3.8) and (3.9) respectively. Note $\bar{a} = 0$, and $\bar{a} = \begin{cases} (N - 1)/N & \text{if } N \text{ odd} \\ 1 & \text{if } N \text{ even}. \end{cases}$ Then (3.8) shows that

$$\text{E}_0 [T_M] = 0, \quad \text{Var}_0 [T_M] = \begin{cases} M_1 M_2/N & \text{if } N \text{ odd} \\ M_1 M_2/(N - 1) & \text{if } N \text{ even}. \end{cases} \quad (3.11)$$

**Example 6** Cox and Snell (1981, Example Q) present data on breaking loads (in oz.) of year, of two types (A and B) coming from six bobbins. Each combination of bobbin and type is represented four times, for 48 observations in a balanced design. Normal quantile plots are presented in Fig. 3.1. The deviation from a straight line for these points indicates departure from normality; a potential thresholding associated with bobbin indicates a role for bobbin as a second factor, to be addressed in Chapter 5. Boxplots by type are shown in Fig. 3.2. We use Mood’s median test for test for a difference in median, ignoring the effect of bobbin. The median strength is 15.75; 15 yarn samples of type B are above the joint median. This leads to Table 3.3, and $T_M = 15 - 9 = 6$. In this even-sample case, from (3.11), $\text{Var}_0 [T_M] = 24 \times 24/47 = 12.26$. Hence the approximately standard normal statistic is $(6 - 0)/\sqrt{12.26} = 1.71$ (and,
The Mann-Whitney-Wilcoxon test

TABLE 3.3: Mood’s Table for the Yarn Example

<table>
<thead>
<tr>
<th></th>
<th>Type A</th>
<th>Type B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater than Median</td>
<td>9</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>Less than Median</td>
<td>15</td>
<td>9</td>
<td>24</td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>24</td>
<td>48</td>
</tr>
</tbody>
</table>

corrected for continuity, is $\frac{5 - 0}{\sqrt{12.26}} = 1.43$. The correction for continuity here is 1, because possible values of $T_M$ are two units apart. The p-value $2 \times \Phi(1.43) = 0.153$. Do not reject the null hypothesis of equal yarn strength. You might also do this using package MultNonParam with

```
attach(yarn)
mood.median.test(strength[type=='A'],strength[type=='B'])
```

or

```
genscorestat((strength>median(strength))*2-1,type,correct=1)
```

Compare this with the two-sample t-test:

```
t.test(strength[type=='A'],strength[type=='B'])
```
to give a p-value 0.029.

Mood’s test, using (3.7) in conjunction with (3.10), is almost never used, and is presented here only as a point of departure for more powerful tests. Mood’s test looses power primarily because of the discreetness of the scores (3.10). In the balance of this chapter, we explore tests with less discrete scores.

3.4 The Mann-Whitney-Wilcoxon test

The Wilcoxon rank-sum test is defined as $T_G^{(2)}$ of (3.7), with $a_j = j$ (Wilcoxon, 1945). That is, the statistic is the sum of ranks of observations coming from the second group. The term “rank-sum” is used to differentiate this test from another test proposed by the same author, to be discussed in a later chapter.

An alternative formulation of this statistic can be obtained by examining the construction of the statistic. The formulation via (3.7) holds ranks fixed, and assigns each rank to a group. Suppose instead that $R_j$ is defined as the rank of observation $j$ from the second sample, among all of the observations
from the combined samples. Note that

\[ T_W = \sum_{j=1}^{M_2} R_j = \sum_{j=1}^{M_2} \#(\text{sample entries less than or equal to } Y_j) \]

\[ = \sum_{j=1}^{M_2} \#(X \text{ values less than or equal to } Y_j) \]

\[ + \sum_{j=1}^{M_2} \#(Y \text{ values less than or equal to } Y_j) \]

\[ = T_U + M_2(M_2 + 1)/2 \]
The Mann-Whitney-Wilcoxon test

The statistic $T_U$ is called the Mann-Whitney statistic (Mann and Whitney, 1947). This statistic is a $U$ statistic; that is, a statistic formed by summing over pairs of observations in a data set.

### 3.4.1 Exact and Approximate Mann-Whitney Probabilities

The distribution of this test statistic can be calculated exactly, via recursion (Festinger, 1946). Let $c_W(t, M_1, M_2)$ the number of ways that $M_1$ symbols $X$ and $M_2$ symbols $Y$ can be written in a vector to give $T_U = t$. Then

$$P_{M_1, M_2}[T_U = t] = c_W(t, M_1, M_2) / \binom{N}{M_2}.$$  \hspace{1cm} (3.13)

The collection of vectors giving statistic value $t$ can be divided according to whether last symbol is $X$ or $Y$. If the last symbol was $X$, then ignoring this final value, the vector still gives the same statistic value $t$, and there are $c_W(t, M_1 - 1, M_2)$ such vectors. If the last symbol was $Y$, then ignoring this final value, the factor gives the statistic value $t - M_1$, and there are $c_W(t - M_1, M_1, M_2 - 1)$ such vectors. So

$$c_W(t, M_1, M_2) = c_W(t, M_1 - 1, M_2) + c_W(t - M_1, M_1, M_2 - 1).$$  \hspace{1cm} (3.14)

The recursion stops once either sample size hits zero:

$$c_W(t, M_1, 0) = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{if } t \neq 0 \end{cases}, \quad \text{and } c_W(t, 0, M_2) = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{if } t \neq 0 \end{cases}.$$  \hspace{1cm} (3.15)

The maximal value for $t$ is $N(N+1)/2 - M_1(M_1+1)/2 = M_2(2M_1+M_2+1)/2$; hence the recursion can be stopped early by noting that

$$c_W(t, M_1, M_2) = 0 \text{ if } t < M_2(2M_1+M_2+1)/2 \text{ or if } t > M_2(2M_1+M_2+1)/2.$$  \hspace{1cm} (3.16)

A natural way to perform these calculations is with recursive calls to a computer routine to calculate lower-order probabilities, although the algorithm...
can be implemented without such explicit recursion (Dinneen and Blakesley, 1973).

### 3.4.1.1 Mann-Whitney-Wilcoxon Statistic Moments and Approximate Normality

Using this recursion can be slow, and the argument at the end of §3.2.1 can be used to show that the distribution of the test statistic is approximately normal. Fortunately, a central limit theorem applies to this statistic (Erdős and Réyni, 1959).

The Wilcoxon version of the Mann-Whitney-Wilcoxon statistic is given by (3.7) with \( a_j = j \). Then \( \sum_{j=1}^{N} a_j = \sum_{j=1}^{N} j = N(N+1)/2 \), and

\[
\bar{a} = (N+1)/2.
\] (3.17)

In order to complete the calculation, one needs \( g(w) = \sum_{j=1}^{w} j^2 \). One might guess it must be cubic in \( N \). Examine functions \( g(w) = aw^3 + bw^2 + cw + d \) so that \( g(0) = 0 \) and \( g(w) - g(w-1) = w^2 \). Then \( d = 0 \), and

\[
w^2 = (aw^3 + bw^2 + cw) - \left(aw^3 + 3aw^2 - 3aw + a - bw^2 + 2bw - b - cw + c\right) = 3aw^2 - 3aw + a + 2bw - b + c.
\]

Equating quadratic terms above gives \( a = 1/3 \). Setting the linear term to zero gives \( b = 1/2 \), and setting the constant term to zero gives \( c = 1/6 \). Then

\[
\sum_{j=1}^{w} j^2 = g(w) = w(2w+1)(w+1)/6,
\] (3.18)

\[
\bar{a} = (2N+1)(N+1)/6,
\]

\[
\bar{a} - \bar{a}^2 = (2N+1)(N+1)/6 - (N+1)^2/4 = (N^2 - 1)/12 \quad (3.19)
\]

and from (3.8), (3.9) and (3.19),

\[
E[T_W] = M_2(N+1)/2, \quad \text{Var}[T_W] = M_1M_2(N+1)/12. \quad (3.20)
\]

The results only minimally different from ANOVA on scores.

**Example 7** Refer again to the yarn data of example 6. Consider yard strengths for bobbin 3.

| 14.2(B) | 14.5(B) | 14.8(A) | 15.2(B) | 15.8(A) | 15.9(B) | 16.0(A) | 18.2(A) |

Sum the ranks associated with Type B, to get \( T_W = 1 + 2 + 4 + 6 = 13 \). Here \( M_1 = M_2 = 4 \), and \( N + 1 = M_1 + M_2 + 1 = 9 \). From (3.20), under the null hypothesis of equality of distributions, the expected value of the rank sum is \( 4 \times 9/2 = 18 \), and the variance is \( 4 \times 4 \times 9/12 = 12 \). Hence the statistic, after standardizing to zero mean and unit variance,
The Mann-Whitney-Wilcoxon test

is \((13 - 18)/\sqrt{12} = -1.44\). The \(p\) - value is 0.149. This may be done using R by

```r
wilcox.test(strength~type, data=yarn[yarn$bobbin==3,],
            exact=FALSE, correct=FALSE)
```

The continuity-corrected \(p\)-value uses statistic \((13 + 0.5 - 18)/\sqrt{12}\), and is 0.194, and might be done by

```r
wilcox.test(strength~type, data=yarn[yarn$bobbin==3,],
            exact=FALSE)
```

Finally, \(p\)-values might be calculated exactly using (3.14), (3.15), and (3.16), and in R by

```r
wilcox.test(strength~type, data=yarn[yarn$bobbin==3,],
            exact=TRUE)
```

Use the Wilcoxon rank-sum procedure to test whether nail arsenic varies by sex.

Moments (3.20) apply to the statistic given by scores \(a_j = j\). By contrast, the Mann-Whitney statistic \(T_U\) is constructed using (3.7) from scores \(a_j = j - (N + 1)/2\). The variance of this statistic is still given by (3.20); the expectation is \(E[T_U] = M_2(N+1)/2 - M_2(M_2 + 1)/2 = M_2M_1/2\).

### 3.4.2 Other scoring schemes

One might construct tests using other scores \(a_j\). A variety of techniques are available for use. One could use scores equal to expected value of order statistics from normal distribution; these are called Normal scores. Alternatively, one could use scores calculated from the normal quantile function \(a_j = \Phi^{-1}(j/(N+1))\) (Waerden, 1952), called Van der Waerden scores, or scores of form \(a_j = \sum_{i=2}^{N} i^{-1}\) (Savage, 1956), called Savage scores, or scores equal to expected value of order statistics from exponential distribution, called exponential scores. Van der Waerden scores are an approximation to normal scores. Calculating exact probabilities for general score tests, and the difficulties that this entails, was discussed at the end of §3.2.1.

Scores may be chosen to be optimal for certain distributions. Normal scores are optimal for normal observations. Exponential scores are optimal for exponential observations. Original ranks are optimal for logistic observations. Savage scores are optimal for Lehmann alternatives (3.26).

**Example 8** Consider the nail arsenic data of example 2. One might perform an analysis using these scoring methods. The R package `exactRankTests` produces scores:

```r
library("exactRankTests")#Gives savage and vw scores
```
arsenic$savagenails<-as.numeric(cscores(arsenic$nails,type="Savage"))
arsenic$vwnails<-as.numeric(cscores(arsenic$nails,type="Normal"))

The Savage scores are

0.603, 0.669, 0.850*, 0.669, −0.056*, −0.366, 0.902*,
0.371, −0.199, 0.794, 0.952, −1.149, −0.816*, −2.649,
−1.649, 0.180, −0.566*, 0.454, 0.069*, 0.531*, 0.280*.

Asterisks denote men. The mean of these scores is \( \bar{a} = −0.005 \), and the mean of the squares is \( \hat{a} = 0.823 \). Test for equality of arsenic in nails between sexes. Here \( M_1 = 8 \) and \( M_2 = 13 \). The expectation and variance of the test statistic are given by (3.8), as \( 13\bar{a} = −0.065 \), and \( 13 \times 8 \times (\bar{a}^2 - \bar{a}^2)/20 = 4.28 \). Sum scores for women, the second gender group; here \( k = 2 \), and \( T^{(k)}_G = −1.32 \), the sum of scores above without the asterisk. The z-statistic is \( (−1.32 − (−0.065))/\sqrt{4.28} = −0.60 \). The p-value is 0.548. Do not reject the null hypothesis. These calculations may be done using

source("common.R")#Contains genscorestat
genscorestat(arsenic$savagenails,arsenic$sex)
genscorestat(arsenic$vwnails,arsenic$sex)

The van der Waerden and Savage scores give p-values of \( 2.609 \times 10^{-6} \) and \( 6.209 \times 10^{-9} \) respectively.

### 3.4.3 Using data as scores: the Permutation Test

One might instead use the original data as scores. That is, sort the combined data set \((X_1,\ldots,X_{M_1},Y_1,\ldots,Y_{M_2})\) to obtain \((Z_{(1)},\ldots,Z_{(N)})\), with \(Z_{(i)} \leq Z_{(i+1)}\) for all \(i\); still assuming continuity, each inequality is strict. Then use \(a_j = Z_{(j)}\). Hence the test statistic is

\[
T_P = \sum_{j=1}^{N} Z_{(j)} I_j^{(2)} = \sum_{j=1}^{M_2} Y_j = M_2 \bar{Y}.
\]

The analysis is performed conditionally on \((Z_{(1)},\ldots,Z_{(N)})\); note that both the statistic, and its reference distribution, depend on these order statistics. Compare \(T_P\) with the numerator of the two-sample pooled t-test (3.2)

\[
\bar{Y} - \bar{X} = \bar{Y} - \frac{N \bar{Z} - M_2 \bar{Y}}{M_1} = \frac{N \bar{Y} - N \bar{Z}}{M_1} = N \frac{(T_P - M_2 \bar{Z})}{(M_1 M_2)},
\]
The Mann-Whitney-Wilcoxon test

where $\bar{Z} = \sum_{i=1}^{N} Z(i)/N$. The pooled variance estimate for the two-sample $t$ statistic is

$$s_p^2 = \frac{\sum_{j=1}^{M_1} (X_j - \bar{X})^2 + \sum_{j=1}^{M_2} (Y_j - \bar{Y})^2}{N - 2}$$

$$= \frac{\sum_{j=1}^{M_1} (X_j - \bar{Z})^2 - M_1(\bar{X} - \bar{Z})^2 + \sum_{j=1}^{M_2} (Y_j - \bar{Z})^2 - M_2(\bar{Y} - \bar{Z})^2}{N - 2}$$

$$= \frac{(N - 1)s_Z^2 - M_1(\bar{X} - \bar{Z})^2 - M_2(\bar{Y} - \bar{Z})^2}{N - 2}.$$

Some algebra shows this to be

$$s_p^2 = s_Z^2 \frac{N - 1}{N - 2} - \frac{(T_P - M_2\bar{Z})^2(1/M_1 + 1/M_2)}{(N - 2)}.$$

Hence, conditional on $(Z(1), \ldots, Z(N))$, the two-sample pooled $t$-statistic is

$$\sqrt{(N - 2)N} \frac{T_P - M_2\bar{Z}}{\sqrt{(s_Z^2(N - 1)M_1M_2 - (T_P - M_2\bar{Z})^2)/N}}.$$

for $s_Z$ the sample standard deviation of $(Z(1), \ldots, Z(N))$.

Hence the pooled two-sample $t$ statistic is a strictly increasing function of the score statistic $T_P$ with ordered data used as scores. However, while the pooled $t$ statistic is typically compared to a $t$ distribution, the rank statistic is compared to the distribution of values arising from random permutations of the group labels; note that this is the same mechanism that generates the distribution for the rank statistics with scores determined in advance. In the two-sample case, there are $(\binom{N}{M_1})$ ways to assign $M_1$ labels 1, and $M_1$ labels 2, to the order statistics $(Z(1), \ldots, Z(N))$. A less-efficient way to think of this process is to specify $N$ labels, the first $M_1$ of them 1 and the remaining $M_1$ of them 2, and randomly assign, or randomly permute, $(Z(1), \ldots, Z(N))$ without replacement; there are $N!$ such assignments, leading to at most $(\binom{N}{M_1})$ distinct values. The observed value of $T_P$ is then compared with the sampling distribution arising from this random permutation of values; such a test is called a permutation test. The same permutation concept coincides with the desired reference distribution for all of the rank statistics in this chapter.

Example 9 Again consider the nail arsenic data of example 2. Recall that there are 21 subjects in this data set, of whom 8 are male. The permutation test testing the null hypothesis of equality of distribution across gender may be performed in R using

```r
library(MultNonParam)
aov.P(dattab=arsenic$nails,permi=arsenic$sex)
```

to give a two-sided $p$-value of 0.482. In this case, all $(21\choose 8) = 203490$ ways
TABLE 3.4: Levels for Various Two-Sample Two-Sided Tests, Nominal level 0.05, from 100,000 Random Data Sets Each, Sample Size 10 Each

<table>
<thead>
<tr>
<th>Test</th>
<th>Gaussian</th>
<th>Laplace</th>
<th>Cauchy</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-test</td>
<td>0.04815</td>
<td>0.04414</td>
<td>0.01770</td>
</tr>
<tr>
<td>Exact Wilcoxon</td>
<td>0.04231</td>
<td>0.04413</td>
<td>0.04424</td>
</tr>
<tr>
<td>Approximate Wilcoxon</td>
<td>0.05134</td>
<td>0.05317</td>
<td>0.05318</td>
</tr>
<tr>
<td>Normal Scores</td>
<td>0.04693</td>
<td>0.04744</td>
<td>0.04871</td>
</tr>
<tr>
<td>Savage Scores</td>
<td>0.04191</td>
<td>0.04340</td>
<td>0.04319</td>
</tr>
<tr>
<td>Mood</td>
<td>0.02198</td>
<td>0.02314</td>
<td>0.02314</td>
</tr>
</tbody>
</table>

To reassign arsenic nail levels to the various groups were considered. The statistic $T_P$ of (3.21) was calculated for each assignment, this value was subtracted from the null expectation $\bar{Z}$, and the difference was squared to provide a two-sided statistic. The $p$-value reported is the proportion of these for which the squared differences among the reassignments meets or exceeds that seen in the original data.

3.5 An Empirical Assessment of Levels and Powers of Two-Sample Tests

As was done in Table 2.1, one might simulate data from a variety of distributions, and compare levels of the various two-sample tests. Results are in Table 3.4. From this table, we see that the extreme conservativeness of Mood's test justifies its exclusion from practical consideration. We see that the Wilcoxon test, calibrated exactly using its exact null distribution, falls short of the desired level; a less-conservative equal-tailed test would have a level exceeding the nominal target of 0.05. The conservativeness of the Savage Score test is somewhat surprising. The close agreement between level of the $T$-test and the nominal level with normal data is as expected, as is the poor agreement between level of the $T$-test and the nominal level with Cauchy data.

As was done in Table 2.3, one might perform a similar simulation under the alternative distribution to calculate power. In this case, alternative distributions were generated by offsetting one group by one unit. Results are in Table 3.5. Table 3.5 excludes the exact version of the Wilcoxon test and Mood's test, since for these sample sizes ($M_j = 10$ for $j = 1, 2$), they fail to achieve the desired level for any data distribution. The approximate Wilcoxon test has comparable power to that of the $T$-test under the conditions optimal for the $T$-test, and also maintains high power throughout.
3.6 Adaptation to the Presence of Tied Observations

The Mann-Whitney-Wilcoxon statistic is designed to be used for variables arising from a continuous distribution. Processes expected to produce data with distinct values, however, sometimes produced tied values, frequently because of limits on measurement precision. Sometimes an observation from from the first group is tied with one from the second group. Then the scheme for assigning scores must be modified. Tied observations are frequently assigned scores averaged over the scores that would have been assigned if the data had been distinct; for example, if $Z(1), \ldots, Z(N)$ are the ordered values from the combination of the two samples, and if $Z(j+1) = Z(j)$, then both observation $j$ and observation $j+1$ are assigned score $(a_j + a_{j+1})/2$. The variance of the test statistic must be adjusted for this change in scores.

When both tied observations come from the first group, or both from the second group, then one might assume that the tie arises because of imprecise measurement of a process that, measured more precisely, would have produced untied individuals. The test statistic is unaffected by assignment of scores to observations according to either of the potential orderings. However, the permutation distribution is affected, because many of the permutations considered will split the tied observations into different groups. Return to variance formula (3.9). The average rank $\bar{a}$ is unchanged by modification of ranks, but the average squared rank $\bar{a}^2$ changes by $(a_j^2 + a_{j+1}^2 - (a_j + a_{j+1})^2)/2 = (a_j - a_{j+1})^2/2$. Then, for each pair of ties in the data, the variance (3.9) is reduced by $M_1M_2(a_j - a_{j+1})^2/(N-1)$. This process could be continued for triplets, etc., with more complicated expressions for the correction. Lehmann (2006) derives these corrections for generic numbers of replicated values, in the simpler case in which $a_j = j$; in this case, the correction is applied to the simpler variance expression (3.20).

It is likely simpler, however, to bypass (3.20), and, instead of correcting (3.9), recalculating (3.9) using the new scores.

When the assumption of continuity of the distributions of underlying measurements does not hold, the distribution of rank statistics is no longer independent of the underlying data distribution, since the rank statistic distribu-
tion will then depend on the probability of ties. Hence no exact calculation of the form in §3.4.1 is possible.

3.7 Mann-Whitney-Wilcoxon null hypotheses

The Mann-Whitney-Wilcoxon test was constructed to test whether the distribution $F$ of the $X$ variables is the same as the distribution $G$ of the $Y$ variables. This null hypothesis implies that $P[X_k \leq Y_j] = 1/2$. Unequal pairs $F$ and $G$ violate the null hypothesis of this test. However, certain distribution pairs violating the null hypothesis fall in the alternative hypothesis, but the Mann-Whitney-Wilcoxon test has no power to distinguish these. This is true if $F$ and $G$ are unequal but symmetric about same point. In this case, the standard error of the test statistic (3.9) is no longer true, and the expectation under this alternative is the same as it is under the null. The same phenomenon arises if $\int_{-\infty}^{\infty} F(y)g(y) \, dy = 1/2$.

3.7.1 An Example

Suppose that $Y_j \sim \mathcal{E}(1)$, $X_i \sim \mathcal{G}(\theta, 1)$. We now determine the $\theta$ for which the above alternative hypothesis has power no larger than the test size. Solve $1/2 = \int_0^\infty (1 - \exp(-y)) \exp(-y - \theta)^2/2)(2\pi)^{-1/2} \, dy$ to obtain $\theta = .876$.

3.8 Efficiency and Power of Two-Sample Tests

In this section, consider models of the form (3.1), with the null hypothesis $\theta = \theta^0$. Without loss of generality, one may take $\theta^0 = 0$; otherwise, shift $Y_j$ by $\theta^0$.

Relative efficiency has already been defined for for test statistics $T$ such that $(T - \mu(\theta))/\sqrt{N} \approx \mathcal{N}(0, 1)$, for $N$ the total sample size. In order to compute asymptotic relative efficiency, we need to specify how $M_1$ and $M_2$ move together. Let $M_1 = \lambda N$, $M_2 = (1 - \lambda)N$, for $\lambda \in (0, 1)$.

3.8.1 Efficacy of the Normal-Theory Test

As in the one-sample case, we will approximate the large sample behavior of this test by the version with known variance. Here $\mu(\theta) = \theta$, and $\text{Var}[T] =$
Efficiency and Power of Two-Sample Tests

\[ \rho^2 \left( \frac{1}{M_2} + \frac{1}{M_1} \right) = \rho^2 \left( \frac{1}{N(1-\lambda)} + \frac{1}{N\lambda} \right); \text{ hence} \]
\[ \sigma(\theta) = \rho \sqrt{\frac{1}{\lambda} + \frac{1}{1-\lambda}} = \frac{\rho}{\zeta}, \]
for \( \rho^2 \) the variance of each observation, and \( \zeta = \sqrt{\lambda(1-\lambda)}. \)

### 3.8.1.1 Normally Distributed Observations

Suppose that \( Y_j \sim \mathcal{G}(0, 1) \), and \( X_i \sim \mathcal{G}(\theta, 1) \). The efficacy is \( e = \zeta. \)

### 3.8.1.2 Logistically Distributed Observations

Each observation has variance is \( \pi^2/3 \), and the efficacy is \( e = \zeta \sqrt{3/\pi} = .551\zeta. \)

### 3.8.2 Efficacy of the Mann-Whitney-Wilcoxon test

In order to apply the results for asymptotic relative of §2.5, the test statistic must be scaled so that the asymptotic variance is approximately equal to a constant divided by the square root of sample size, and must be such that the derivative of the mean function is available at zero. Using the \( U \)-statistic formulation, and rescaling so that the
\[ T = \sum_{i=1}^{M_1} \sum_{j=1}^{M_2} I(X_i < Y_j)/(M_1 M_2), \]
then
\[ \text{Var}[T] = \frac{N+1}{12 M_1 M_2} \approx \frac{1}{N(12\lambda(1-\lambda))}, \]
and so
\[ \sigma(0) = 1/\sqrt{12\zeta}. \] (3.22)

Also,
\[ \mu(\theta) = P_\theta [Y > X] = P_0 [Y + \theta > X] = P_0 [\theta > X - Y]. \] (3.23)

#### 3.8.2.1 Normally Distributed Observations

Suppose that \( Y_j \sim \mathcal{G}(0, 1) \), and \( X_i \sim \mathcal{G}(\theta, 1) \). The differences \( X_i - Y_j \sim \mathcal{G}(\theta, 2) \), and so
\[ \mu(\theta) = \Phi(\theta/\sqrt{2}). \] (3.24)

Hence \( \mu'(0) = \frac{1}{2\sqrt{\pi}} \). Also, (3.22) still holds, and
\[ e = \frac{1}{2\sqrt{\pi}} \sqrt{12\zeta} = \sqrt{3/\pi} \zeta = .977\zeta. \]

#### 3.8.2.2 Logistically Distributed Observations

In this case,
\[ \mu(\theta) = \int_{-\infty}^{\infty} \int_{x-\theta}^{\infty} \frac{\exp(x) \exp(y)}{(1+\exp(x))^2 (1+\exp(y))^2} dy dx \]
\[ = e^\theta (e^\theta - \theta - 1) (e^\theta - 1)^{-2}, \] (3.25)
TABLE 3.6: Asymptotic Relative Efficiencies for Two Sample Tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Pooled ( T )</th>
<th>MWW</th>
<th>ARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>( \mu(\theta) = \theta ) &lt;br&gt;( \sigma(\theta) = \sqrt{\text{Var}[X_i]</td>
<td>\zeta^{-1}} )</td>
<td>( \mu(\theta) = \text{P}[\theta &gt; X - Y] ) &lt;br&gt;( \sigma(\theta) = \frac{1}{\sqrt{12}} \zeta^{-1} )</td>
</tr>
<tr>
<td>Normal, unit variance</td>
<td>( \mu'(0) = 1 ) &lt;br&gt;( \sigma(0) = \zeta^{-1} ) &lt;br&gt;( e = \zeta )</td>
<td>( \mu'(0) = (2\sqrt{\pi})^{-1} ) &lt;br&gt;( \sigma(0) = \frac{1}{\sqrt{12}} \zeta^{-1} ) &lt;br&gt;( e = \sqrt{3/3\pi} \zeta )</td>
<td>( \frac{\pi^2}{9} = 1.10 )</td>
</tr>
<tr>
<td>Logistic</td>
<td>( \mu'(0) = (\sqrt{3/3\pi})^{-1} ) &lt;br&gt;( \sigma(0) = \frac{1}{6} \zeta^{-1} ) &lt;br&gt;( e = (1/\sqrt{3})\zeta )</td>
<td>( \mu'(0) = (1) \zeta^{-1} ) &lt;br&gt;( \sigma(0) = (1/\sqrt{3})\zeta )</td>
<td>( \zeta = (\lambda(1-\lambda))^{1/2} )</td>
</tr>
</tbody>
</table>

\( \zeta = (\lambda(1-\lambda))^{1/2} \)

and

\( \mu'(0) = 1/6, \ e = (1/6)\sqrt{12}\zeta = (1/\sqrt{3})\zeta = .577\zeta. \)

Efficacies for more general rank statistics may be obtained using calculations involving expectations of derivatives of underlying densities, with respect to the model parameter, evaluated at order statistics under the null hypothesis, without providing rank expectations away from the null (Dwass, 1956).

### 3.8.3 Summarizing Asymptotic Relative Efficiency

Table 3.6 contains results of calculations for asymptotic relative efficiencies of the Mann-Whitney-Wilcoxon test to the Pooled \( T \) test. For normal variables, as expected, the Pooled \( T \) test is more efficient, but only by 5%. For a distribution with moderate tails, the logistic, the Mann-Whitney-Wilcoxon test is 10% more efficient.

### 3.8.4 Power for Mann-Whitney-Wilcoxon Testing

Power may be calculated for Mann-Whitney-Wilcoxon testing, using (2.15) in conjunction with (3.22) for the null variance of the rescaled test, and (3.23), adapted to the particular distribution of interest. Application to normal and Laplace observations are given by (3.24) and (3.25) respectively. Zhong and Kolassa (2017) give second moments for this statistic under the alternative hypothesis, and allow for calculation of \( \sigma_1(\theta) \) for non-null \( \theta \). The second moment depends not only on the probability (3.23), but also on probabilities involving two independent copies of \( X \) and one copy of \( Y \), and of two independent
copies of $Y$ and one copy of $X$. This additional calculation allows the use of (2.10); calculations below involve the simpler formula.

**Example 10** Consider using observations $X_1, \ldots, X_{40}$ and $Y_1, \ldots, Y_{40}$ to test the null hypothesis of equal distributions vs. the alternative that (3.1) holds, with $\theta = 1$, and with observations having a Laplace distribution. Then, using (3.25), $\mu(1) = e(e-2)(e-1)^{-2} = 0.661$. The function $\mu(\theta)$ has a removable singularity at zero; fortunately the null probability is easily seen to be $1/2$. Then $\lambda = 1/2$, $N = 80$, $\mu(0) = 1/2$, $\mu(1) = 0.661$, $\sigma(0) = 1/\sqrt{12} \times (1/2) \times (1/2) = 1/\sqrt{3}$. The power for the one-sided level 0.025 test, from (2.15), is $\Phi(0.661 - 0.5)/(1/\sqrt{3}) - 1.96) = \Phi(0.534) = .707$.

One could also determine the total sample size needed to obtain 80% power. Using (2.16), one needs $(1/\sqrt{3})^2 (z_{0.025} + z_{0.2})^2/(0.661 - 0.5)^2 = 151.4$; choose 76 per group.

In contrast with the shift alternative (3.1), one might consider the Lehmann alternative

$$1 - F(z) = (1 - G(z))^k \forall z,$$

(3.26)

for some $k \neq 1$. Power calculations for Mann-Whitney-Wilcoxon tests for this alternative have the advantage that power does not depend on the underlying $G$ (Lehmann, 1953).

As noted above, while efficacy calculations are available for more general rank statistics, non-asymptotic expectation of the test statistic under the alternative is difficult enough that it is omitted here.

### 3.9 Testing Equality of Dispersion

One can adapt the above rank tests to test whether two populations have equal dispersion, assuming a common center. If one population is more spread out than another, then the members of one sample would tend to lie outside the points from the other sample. This motivates the Siegel-Tukey Test. Rank the points, with the minimum getting rank 1, the maximum getting rank 2, then second to the maximum getting rank 3, the second to the minimum getting rank 4, the third from the minimum getting rank 5 and continuing to alternate. Then sum the ranks associated with one of the samples. Under the null hypothesis, this statistic has the same distribution as the Wilcoxon Rank Sum test. Alternately, one might perform the Ansari-Bradley Test, by ranking from the outside in, with extremes getting equal rank, and again summing the ranks from one sample.
The Ansari-Bradley test has a disadvantage with respect to the Siegel-Tukey test, in that one can’t use off-the-shelf Wilcoxon tail calculations. On the other hand, the Ansari-Bradley test is exactly invariant to reflection.

Example 11  Consider again the yarn data of example 6. Test equality of dispersion between the two types of yarn. Ranks are given in Table 3.7.

\[
yarn$ab<-\text{pmin}(\text{rank}(\text{yarn$strength}), \text{rank}(-\text{yarn$strength}))
\]
\[
yarn$st<-\text{round}(\text{siegel.tukey.ranks(yarn$strength),2})
\]
\[
yarnranks<-\text{yarn[order(yarn$strength), c("strength","type","ab","st")]}]
\]

R functions may be used to perform the test.

\[
\text{source("siegel.tukey.r")}
\]
\[
\text{siegel.tukey(yarn$strength,yarn$type=="B",id.col=TRUE)}
\]
\[
\text{ansari.test(yarnsplit[[1]],yarnsplit[[2]])}
\]

to find the Siegel-Tukey \(p\)-value as 0.7179, and the Ansari-Bradley \(p\)-value as 0.6786. There is no evidence of inequality of dispersion.

3.10 Two-Sample Estimation and Confidence Intervals

Practitioners often ask what two samples can tell us about how a population location parameter (be it mean, median, or another quantile) differs in two populations. In the restricted case in which the populations are assumed to be the same except for the location parameter, this question does not depend on what location measure is intended.

Our treatment of two-sample confidence intervals will mirror that of the one-sample interval. That is, we construct a family of test statistics, indexed by this parameter, such that the distribution of the family member is independent of the parameter, when the statistic is evaluated at the correct parameter value. We then invert the test by determining for which parameters the null hypothesis is not rejected.

Denote the samples as \(X_1,\ldots,X_{M_1}\) and \(Y_1,\ldots,Y_{M_2}\) as before. Let \(\theta\) represent the amount by which a location parameter for the population from which the second sample exceeds that of the first sample. Under the assumption that the distributions are identical up to shift, then \(X_1,\ldots,X_{M_1}, Y_1-\theta,\ldots,Y_{M_2}-\theta\) all have the same distribution. Then let \(T_G^{(2)}(\theta)\) be the rank sum statistic (3.7) calculated from this data set (3.10).

Most commonly the scores are chosen to make \(T_G^{(2)}(\theta)\) the Wilcoxon rank sum statistic, or equivalently the Mann-Whitney statistic, but conceptually this could be done by inverting, for example, Mood’s median test or any other
TABLE 3.7: Yarn data with rankings for testing dispersion

<table>
<thead>
<tr>
<th>strength</th>
<th>type</th>
<th>ab</th>
<th>st</th>
<th>strength</th>
<th>type</th>
<th>ab</th>
<th>st</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.8</td>
<td>A</td>
<td>1.0</td>
<td>1.00</td>
<td>15.8</td>
<td>A</td>
<td>24.0</td>
<td>47.00</td>
</tr>
<tr>
<td>13.0</td>
<td>A</td>
<td>2.0</td>
<td>4.00</td>
<td>15.9</td>
<td>A</td>
<td>22.0</td>
<td>43.67</td>
</tr>
<tr>
<td>13.8</td>
<td>A</td>
<td>3.0</td>
<td>5.00</td>
<td>15.9</td>
<td>B</td>
<td>22.0</td>
<td>43.67</td>
</tr>
<tr>
<td>14.2</td>
<td>A</td>
<td>4.5</td>
<td>8.50</td>
<td>15.9</td>
<td>B</td>
<td>22.0</td>
<td>43.67</td>
</tr>
<tr>
<td>14.2</td>
<td>B</td>
<td>4.5</td>
<td>8.50</td>
<td>16.0</td>
<td>A</td>
<td>19.5</td>
<td>38.50</td>
</tr>
<tr>
<td>14.5</td>
<td>B</td>
<td>6.0</td>
<td>12.00</td>
<td>16.0</td>
<td>B</td>
<td>19.5</td>
<td>38.50</td>
</tr>
<tr>
<td>14.8</td>
<td>A</td>
<td>7.5</td>
<td>14.50</td>
<td>16.2</td>
<td>A</td>
<td>17.0</td>
<td>33.33</td>
</tr>
<tr>
<td>14.8</td>
<td>A</td>
<td>7.5</td>
<td>14.50</td>
<td>16.2</td>
<td>B</td>
<td>17.0</td>
<td>33.33</td>
</tr>
<tr>
<td>14.9</td>
<td>A</td>
<td>10.0</td>
<td>19.33</td>
<td>16.2</td>
<td>B</td>
<td>17.0</td>
<td>33.33</td>
</tr>
<tr>
<td>14.9</td>
<td>B</td>
<td>10.0</td>
<td>19.33</td>
<td>16.4</td>
<td>A</td>
<td>15.0</td>
<td>30.00</td>
</tr>
<tr>
<td>14.9</td>
<td>B</td>
<td>10.0</td>
<td>19.33</td>
<td>16.8</td>
<td>B</td>
<td>14.0</td>
<td>27.00</td>
</tr>
<tr>
<td>15.0</td>
<td>A</td>
<td>13.5</td>
<td>26.50</td>
<td>16.9</td>
<td>B</td>
<td>13.0</td>
<td>26.00</td>
</tr>
<tr>
<td>15.0</td>
<td>A</td>
<td>13.5</td>
<td>26.50</td>
<td>17.0</td>
<td>A</td>
<td>11.0</td>
<td>21.33</td>
</tr>
<tr>
<td>15.0</td>
<td>A</td>
<td>13.5</td>
<td>26.50</td>
<td>17.0</td>
<td>B</td>
<td>11.0</td>
<td>21.33</td>
</tr>
<tr>
<td>15.0</td>
<td>B</td>
<td>13.5</td>
<td>26.50</td>
<td>17.0</td>
<td>B</td>
<td>11.0</td>
<td>21.33</td>
</tr>
<tr>
<td>15.2</td>
<td>B</td>
<td>16.5</td>
<td>32.50</td>
<td>17.1</td>
<td>A</td>
<td>9.0</td>
<td>18.00</td>
</tr>
<tr>
<td>15.2</td>
<td>B</td>
<td>16.5</td>
<td>32.50</td>
<td>17.2</td>
<td>B</td>
<td>8.0</td>
<td>15.00</td>
</tr>
<tr>
<td>15.5</td>
<td>A</td>
<td>19.0</td>
<td>37.67</td>
<td>17.6</td>
<td>A</td>
<td>7.0</td>
<td>14.00</td>
</tr>
<tr>
<td>15.5</td>
<td>A</td>
<td>19.0</td>
<td>37.67</td>
<td>18.0</td>
<td>B</td>
<td>6.0</td>
<td>11.00</td>
</tr>
<tr>
<td>15.5</td>
<td>B</td>
<td>19.0</td>
<td>37.67</td>
<td>18.1</td>
<td>B</td>
<td>5.0</td>
<td>10.00</td>
</tr>
<tr>
<td>15.6</td>
<td>A</td>
<td>22.0</td>
<td>43.33</td>
<td>18.2</td>
<td>A</td>
<td>3.5</td>
<td>6.50</td>
</tr>
<tr>
<td>15.6</td>
<td>A</td>
<td>22.0</td>
<td>43.33</td>
<td>18.2</td>
<td>B</td>
<td>3.5</td>
<td>6.50</td>
</tr>
<tr>
<td>15.6</td>
<td>B</td>
<td>22.0</td>
<td>43.33</td>
<td>18.5</td>
<td>B</td>
<td>2.0</td>
<td>3.00</td>
</tr>
<tr>
<td>15.7</td>
<td>A</td>
<td>24.0</td>
<td>48.00</td>
<td>19.2</td>
<td>B</td>
<td>1.0</td>
<td>2.00</td>
</tr>
</tbody>
</table>
Two-Sample Testing

rank test. For general scores \( a_j \), \( T_G^{[2]}(\theta) = \sum_{j=1}^{N} a_j Z_j(\theta) \), where \( Z_j(\theta) \) is 1 if item ranked \( j \) among (3.10) came from \( Y \), and 0 otherwise.

One can define an estimator as that value of \( \theta \) that makes test statistic equal to its null expectation; that is, \( \hat{\theta} \) solves \( T_G^{[2]}(\hat{\theta}) = M \bar{a} \). Furthermore, one can determine the largest \( t_l \) and smallest \( t_u \) such that

\[
P_0 \left[ T_G^{[2]}(0) \leq t_l \right] \leq \alpha/2, \quad P_0 \left[ T_G^{[2]}(0) \geq t_u \right] \leq \alpha/2.
\]

(3.27)

Applying the normal approximation to \( T_G^{[2]}(\theta) \),

\[
t_l, t_u \approx M \bar{a} \pm z_{\alpha/2} \sqrt{\text{Var} \left[ T_G^{[2]}(0) \right]}. \tag{3.28}
\]

3.10.1 Inversion of the Mann-Whitney-Wilcoxon test

When \( a_j \) are ranks \( j \), \( T_G^{[2]}(\theta) \) is the Wilcoxon version of the test. The corresponding Mann-Whitney version \( T_U(\theta) = \sum_i \sum_j I(X_i < Y_j - \theta) \) gives the estimator and confidence interval end points more easily. In this case, the null expectation of the statistic is \( M_1 M_2 / 2 \), and \( T_U(\theta) = M_2 M_1 / 2 \) if and only if \( M_2 \times M_1 \) even, and exactly \( M_2 \times M_1 / 2 \) of \( V_{ij} = Y_j - X_i \) are greater than \( \theta \), or \( M_2 \times M_1 \) odd, and \( (M_2 \times M_1 - 1) / 2 \) of \( V_{ij} = Y_j - X_i \) are \( \geq \theta \), \( (M_2 \times M_1 - 1) / 2 \) are \( < \theta \), and one is \( \theta \). Hence the estimator is the median of differences of pairs \( Y_j - X_i \). This estimator is given by Hodges and Lehmann (1963), in the same paper giving the analogous estimator for the one-sample symmetric problem of §5.1.1. The confidence interval created by inverting the Mann-Whitney statistic

\[
T_U^{[2]}(\theta) = \sum_{i=1}^{M_1} \sum_{j=1}^{M_2} I(Y_i - X_j > \theta),
\]

(3.29)

is

\[
\{ \theta | t_l \leq T_U^{[2]}(\theta) < t_u \}, \tag{3.30}
\]

for the largest \( t_l \) and smallest \( t_u \) as in (3.27), made specific to the Mann-Whitney statistic:

\[
\sum_{k=0}^{t_l} P_{M_1,M_2} \left[ T_U^{[2]}(0) = k \right] \leq \frac{\alpha}{2}, \quad \sum_{k=t_u}^{M_1 M_2} P_{M_1,M_2} \left[ T_U^{[2]}(0) = k \right] \leq \frac{\alpha}{2}, \tag{3.31}
\]

with probabilities given by (3.13). This defines the Hodges-Lehmann estimator in terminology of Higgins (2004) and SAS Institute Inc. (2017). More standard definitions in one-sample case follows later. One conventionally uses the mean of middle two paired differences if \( M_2 \times M_1 \) even, even though not technically required by the theory.

When ties are present, (3.13) no longer gives the exact sampling distribution of \( T_U^{[2]}(0) \); exact intervals may be determined using the permutation
distribution of the tie-adjusted ranks, as in Section 3.4.3, or (3.28) may be employed.

In this case of confidence intervals derived by inverting the Mann-Whitney (or equivalently Wilcoxon rank sum) test, using (3.30), $\theta$ satisfies $T_U(\theta) = t_l$ if there are $t_l$ pairs $X_i, Y_j - \theta$ such that the second component is greater than the first, and the remainder of the pairs have the inequality reversed, or equivalently if there are $t_l$ pairs with differences $V_{ij} = Y_j - X_i$ such that $V_{ij} > \theta$, and remainder reversed.

Hence the confidence interval is formed by, first, finding $t_l$ and $t_u$ as in (3.31), and, second, calculating and sorting all $Y - X$ pairs, and, third, reporting differences $t_l$ and $t_u$ on this list. If (3.28), or some other approximation to the Mann-Whitney critical values, is employed, then the above technique indicates an order statistic that is generally not an integer. In this case, many procedures interpolate between adjacent order statistics.

Note parallels between this construction and that of the confidence interval for the median in Section 2.4. In this section, pairwise difference replace raw observations, and the Wilcoxon quantiles replace those of the binomial distribution, but the rest of the construction is identical.

**Example 12** Again consider the arsenic data of example 2. We estimate the median of the distribution of men’s nail arsenic levels minus women’s arsenic levels, and begin by calculating the $13 \times 8 = 104$ pairwise differences. One woman had an unusually large nail arsenic level; hence there is one particularly negative difference. Figure 3.3 displays construction of the confidence interval. Confidence intervals given correspond to those order statistics determined from the quantiles of the Mann-Whitney distribution, as in (3.31). These quantiles are $t_l = 25$ and $t_u = 80$. The confidence interval is then entries 25 and 80 among the list of ordered pairwise differences, or $(-0.278, 0.158)$. This was calculated using `source("common.R")`;
`fun.invertsigntest(split(arsenic$nails,arsenic$sex))`.

The above analysis did not reflect the fact of a small number of ties among these pairwise differences. The code

```
attach(arsenic); wilcox.test(nails~sex,conf.int=TRUE)
```

gives intervals found by using approximation (3.28) to the test critical values, and interpolating between the appropriate order statistics, to obtain an identical result to the same accuracy.
3.11 Tests for Broad Alternatives

The above development motivates different tests for equality of distribution, depending on the kind of departure we are most interested in detecting. For example, a Mann-Whitney test is appropriate for shift alternatives, and an Ansari-Bradley test is appropriate for differences in spread. In this section we describe a test sensitive to departures in various directions.

Use the empirical cumulative distribution function estimator as in section 2.6. The Kolmogorov-Smirnov test uses the largest difference between these as test statistic. The null hypothesis under the permutation distribution is formed from all permutations of data between the two samples. The $p$-value is the portion with as large or larger difference. If the number of such permutations is quite large, one might use a random sample instead. Asymptotic approximations to these distributions exist as well. Calculation of these statistics can be simplified by noting that it depends on differences in the empirical cumulative distribution functions evaluated exclusively at jumps in one or the other curve.

Alternatively, one might use the integral of difference between these, squared, as test statistic; this statistic is called the Cramér-von Mises test.
Conceptually, to implement this test, use all permutations of data between the two samples, and count the proportion with as large or larger integrated difference difference as the \( p \)-value. When the number of permutations is excessive, one might also do this with a random sample of permutations. Alternatively, one might use results from Stochastic Processes to approximate tail areas.

**Example 13** Again consider the the yarn data of example 6. Figure 3.4 shows the cumulative distribution functions of strengths for the two types of yarn. This figure might be generated using

```r
par(mfrow=c(1,1))
plot(range(yarn$strength),c(0,1),type="n",
     main="Yarn Strength",xlab="Yarn Strength",
     ylab="Probability")
yarnsplit<-split(yarn$strength,yarn$type)
lines(ecdf(yarnsplit[[1]]),col=1)
lines(ecdf(yarnsplit[[2]]),col=2)
legend(17,.2,lty=c(1,1),col=c(1,2),legend=c("A","B"))
```

The largest difference between distribution functions happens at strengths slightly larger than 16; this gives the Kolmogorov-Smirnov statistic. The Cramér-von Mises statistic is the integral of the squared distance between these curves. Statistics may be calculated using

```r
ks.test(yarnsplit[[1]],yarnsplit[[2]])
library("CvM2SL2Test")
cvmstat<-cvmts.test(yarnsplit[[1]],yarnsplit[[2]])
cvmts.pval(cvmstat,length(yarnsplit[[1]]),length(yarnsplit[[2]]))
```

Note that the library `CvM2SL2Test`, giving `cvmts.test` and `cvmts.pval`, is no longer supported, and must be installed from archives. Significance tests are based on permutation distributions; the Kolmogorov-Smirnov \( p \)-value is 0.2591, and the Cramér-von Mises \( p \)-value is 0.09474.

### 3.12 Exercises

1. The data set at

   http://ftp.uni-bayreuth.de/math/statlib/datasets/schizo

   gives data from an experiment using measurements used to detect schizophrenia, on non-schizophrenic patients. Various tests are
given. Pick those data points with CS in the second column and compare the first and second gain ratios, in the third and fourth columns, by taking their difference. Test the null hypothesis of zero median difference using the sign test.

2. The data set at

http://ftp.uni-bayreuth.de/math/statlib/datasets/schizo

contains data from an experiment using measurements used to detect schizophrenia, on non-schizophrenic patients. Results of various eye-tracking tests are given. The second column gives eye tracking target type. For all parts of this question select subjects with target type CS. Gain ratios are given in various columns; for the balance of this question, use the first of these, found in the third column.

a. The comments at the top of the data file tell you which subjects are female. Compare the first gain ratios for women with those for men. Test whether the gain ratios come from the same distribution, using the Wilcoxon rank sum test.

b. Select the female patients from the schizophrenia data set. Calcu-
late a 95% confidence interval for median of the difference between
the mean first gain ratios for schizophrenic patients (in the lower part of the file) and for the non-schizophrenic patients (in the upper part of the file).

c. Select the female patients from the schizophrenia data set. Test the null hypothesis that the dispersion of the first gain ratios is the same for schizophrenic patients (in the lower part of the file) as it is for the non-schizophrenic patients (in the upper part of the file). Use the Ansari-Bradley test.

d. Again select the female patients from the schizophrenia data set. Test the null hypothesis that the distribution of the first gain ratios is the same for schizophrenic patients (in the lower part of the file) as it is for the non-schizophrenic patients (in the upper part of the file). Use the Kolmogorov-Smirnov test.

e. Test whether the first gain ratios for women and men come from the same distribution. Use data values as scores.

3. Calculate the asymptotic relative efficiency for the Mann-Whitney-Wilcoxon test to the two-sample $T$-test. Do this for observations from the

a. Laplace distribution of §1.1.2.3.

b. Cauchy distribution of §1.1.2.4, again abusing notation to extend to the Cauchy’s property that the central limit theorem does not apply to sums of independent and identically-distributed observations from this distribution.

4. Determine a pair of distributions $F$ and $G$ satisfying both the Lehmann alternative (3.26) and the shift alternative (3.1).

5. Demonstration that the two-sample $t$-statistic for Gaussian data follows the expected distribution uses the fact that the mean differences are independent of the pooled standard deviation estimate, and that the pooled variance estimate times degrees of freedom has a $\chi^2$ distribution with the expected number of degrees of freedom. Using simulation, for tests arising from samples of size $M_1 = M_2 = 10$, evaluate both of these assumptions for data coming from a Laplace distribution, and from a Cauchy distribution.
In this chapter we develop nonparametric techniques for one way analysis of variance.

Suppose $X_{ki}$ are samples from $K$ potentially different populations. That is, for fixed $k$, $X_{k1}, \ldots, X_{kM_k}$ are independent and identically distributed, each with cumulative distribution function $F_k$. Here $k \in \{1, \ldots, K\}$ indexes group, and $M_k$ represents the number of observations in each group. We wish to determine whether all populations are the same. We test a null hypothesis

$$H_0 : F_1(x) = \cdots = F_K(x) \forall x,$$

vs the alternative hypothesis $H_A : \text{there exists } j, k, \text{ and } x \text{ such that } F_j(x) \neq F_k(x)$. Most tests considered in this chapter, however, are most powerful against alternatives of the form $H_A : F_k(x) \leq F_j(x) \forall x$ for some indices $k, j$, with strict inequality at some $k, j$, and $x$. Of particular interest, particularly for power calculations, are alternatives of the form

$$F_i(x - \theta_i) = F_j(x - \theta_j)$$

for some constants $\theta_1, \ldots, \theta_K$.

### 4.1 Normal-Theory Methods

Under the assumptions of normality and homoscedasticity (that is, equality of variances) of the $X_{ki}$ under both the null and alternative hypotheses, the null hypothesis $H_0$ is equivalent to $\mu_j = \mu_k$ for all pairs $j, k$ for $\mu_j = \mathbb{E}[X_{ji}]$.

One might test $H_0$ vs. $H_3$ via analysis of variance. Let $\bar{X}_k = \frac{\sum_{i=1}^{M_k} X_{ki}}{M_k}$, $\bar{X}_. = \sum_{k=1}^{K} \frac{\sum_{i=1}^{M_k} X_{ki}}{M_k}$, and

$$W = \frac{(\sum_{k=1}^{K} M_k (\bar{X}_k - \bar{X}.)^2)/(K - 1)}{\hat{\sigma}^2},$$

for

$$\hat{\sigma}^2 = \frac{\sum_{k=1}^{K} M_k \sum_{i=1}^{M_k} (X_{ki} - \bar{X}_k)^2}{\sum_{k=1}^{K} M_k - k}.$$

71
When the data are normally distributed, and (4.1) holds, the numerator and denominator of (4.3) have a $\chi^2$ distribution, and are independent; hence the ratio $W$ has a $F$ distribution. This approach is known as the Analysis of Variance (ANOVA).

When normality does not hold, the central limit theorem implies that the numerator is still approximately $\chi^2$, as long as the minimal $M_k$ is large, and as long as the distribution of the data is not too far from normal. However, neither the $\chi^2$ distribution for the denominator of (4.3), nor the independence of numerator and denominator, are guaranteed in this case. Fortunately, again for large sample sizes and data not too far from normal, the strong law of large numbers indicates that the denominator of (4.3) is close to the population variance of the observations, and the denominator degrees of freedom for the $F$ distribution is large enough to make the $F$ distribution close the the $\chi^2$ distribution. Hence in this large-sample close to normal case, the standard Analysis of Variance results will not mislead.

4.1.1 Contrasts

Consider alternatives to null hypothesis (4.1), in which $X_{ki}$ all have the same finite variance $\sigma^2$. Under the null hypothesis, all of the expectations are equal. Setting $\mu_k = E[X_{ki}]$, $\mu_j = \mu_k$ for all $j, k$ pairs. Suppose that one has in mind a particular alternative pattern to these means; for example, one might consider alternatives of the form $\Delta = \mu_{k+1} - \mu_k$ for all $0 < k < K$ and some $\Delta \neq 0$. One might construct a test particularly sensitive to this departure from the null hypothesis using the estimate $\hat{\Delta}$. If $\hat{\Delta}$ is approximately normal, then the associated test of the null hypothesis (4.1) vs. the ordered and equi-distant alternative is constructed as $T = (\hat{\Delta} - E_0[\hat{\Delta}]) / \sqrt{\text{Var}_0[\hat{\Delta}]}$; this statistic is compared to the standard Gaussian distribution in the usual way.

An intuitive estimator $\hat{\Delta}$ is the least squares estimators; for example, when $K = 3$ then $\hat{\Delta} = (\bar{X}_3 - \bar{X}_1)/2$, and when $K = 4$ then $\hat{\Delta} = (3\bar{X}_4 + \bar{X}_3 - \bar{X}_2 - 3\bar{X}_1)/10$. Generally, least squares estimators will have $c_k$ evenly spaced. For all $K$, the least squares estimator is a linear combination $\sum_{k=1}^{K} c_k \bar{X}_k$, for a set of constants such that

$$\sum_{k=1}^{K} c_k = 0.$$  

(4.5)

In this case, $E_0[\hat{\Delta}] = 0$ and $\text{Var}_0[\hat{\Delta}] = \sigma^2 \sum_{k=1}^{K} c_k^2 / M_k$, and one may use the test statistic $T = \sum_{k=1}^{K} c_k \bar{X}_k / (\sigma \sqrt{\sum_{k=1}^{K} c_k^2 / M_k})$. If $\sigma$ is known, then $T$ is compared to a standard Gaussian distribution. If $\sigma$ is estimated as in (4.4), then $T$ is compared to a $T_{N-K}$ distribution.

In this case, $W$ of (4.3) retains its level, but has less power against the ordered alternative.
Normal-Theory Methods

A linear combination \( \sum_{k=1}^{K} c_k \bar{X}_k \) of group means, with constants summing to zero as in (4.5), is called a contrast.

Standard parametric methods will be compared to nonparametric methods below. In order to make these comparisons using methods of efficiency, we express the pattern of numbers of observations in various groups in terms of a single sample size \( N \); presume that \( M_k = \lambda_k N \) for all \( k \in \{1, \ldots, K\} \). Under alternative (4.2), and with sample sizes large enough to take \( \xi^2 = \text{Var}[X_{ij}] \) as known, then

\[
\sum_{k=1}^{K} c_k \bar{X}_k \sim \mathcal{N}(\sum_{k=1}^{K} c_k \theta_k, \xi^2 (\sum_{k=1}^{K} c_k^2 / \lambda_k) / N). \tag{4.6}
\]

In the case when the shift parameters and the contrast coefficients are equally spaced, and for groups of equal size (that is, \( \theta_k = (k-1)\Delta \) for some \( \Delta > 0 \), \( c_k = 2k - (K+1) \), and \( \lambda = 1/K \)),

\[
\sum_{k=1}^{K} c_k \bar{X}_k \sim \mathcal{N}(K^2(K-1)\Delta/6, (\xi^2/3)K^2(K^2-1)/N). \tag{4.7}
\]

Then \( \mu'(\Delta) = (K^2 - 1)K/6 \), and \( \sigma(0) = \xi K \sqrt{(K^2-1)/3} \), and

\[
e = \frac{K(K^2 - 1)/6}{\xi K \sqrt{(K^2-1)/3}} = \frac{\sqrt{K^2 - 1}}{2\sqrt{3}\xi}. \tag{4.8}
\]

4.1.2 Multiple Comparisons

In the event that the null hypothesis of equal distributions is rejected, one naturally asks which distributions differ. Testing for group differences pairwise (perhaps using the two-sample \( t \) test) allows for \( K(K-1)/2 \) chances to find a significant result. If each test is done at nominal level, this will inflate family-wise error rate. This family-wise error rate will be bounded by the nominal level used for each separate test multiplied by number of possible comparisons performed (1/2\( K(K-1) \)), but such a procedure, called the Bonferroni procedure, will get usually result in a very conservative bound.

Alternatively, consider Fisher’s Least Significant Difference (LSD) method. First, perform the standard Analysis of Variance test (4.3). If this test rejects \( H_0 \), then test on each pairwise comparison between mean ranks, and report those pairs whose otherwise uncorrected \( p \)-values are smaller than the nominal size. Tests in this second stage may be performed using the two-sample pooled \( t \) test (3.2), except that \( \sqrt{\text{EMS}} \) of (4.4) may be substituted for \( s_p \), with the corresponding increase in degrees of freedom in (3.5).

Fisher’s LSD method fails to control Type I error rate (otherwise known as size) if \( K > 3 \). To see this, suppose \( F_1(x) = F_2(x) = \cdots = F_{K-1}(x) = F_K(x - \Delta) \) for \( \Delta \neq 0 \). Then null hypotheses \( F_i(x) = F_i(x) \) are true, for \( i, j < K \). One can make \( \Delta \) so large that the Analysis of Variance test rejects
equality of all distributions with probability close to 1. Then multiple true null hypotheses are tested, without control for multiplicity. If \( K = 3 \), there is only one such test with a true null hypothesis, and so no problem with multiple comparisons.

Contrast this with Tukey’s Honest Significant Difference (HSD) method (Tukey, 1953, 1993), using the Studentized range distribution. Suppose that \( Y_j \sim \Phi(0,1/M_j) \) for \( j \in \{1,\ldots,K\} \), \( U \sim \chi^2_m \), and that the \( Y_j \) and \( U \) are independent. Assume further that \( M_j \) all equal. Then the distribution of

\[
\sqrt{2 \max_{1 \leq i,j \leq K} \left( \frac{|X_j - X_i|}{(\sqrt{U/m})\sqrt{1/M_j + 1/M_k}} \right)}
\]

has the studentized range distribution with \( K \) and \( m \) degrees of freedom, approximately (Kramer, 1956); extensions also exist to correlated means (Kramer, 1957). Let \( q_{K,m,\alpha} \) be the \( 1-\alpha \) quantile of this distribution.

One then applies this distribution with \( Y_j = (\bar{X}_j - \mu_j)/\sigma \) and \( U/m \) the standard sample variance \( S^2 \). Here \( \sigma \) is the common standard deviation. If one then sets

\[
P_{jk} = 2q_{K,N-K,\sqrt{2}}(\bar{X}_j - \bar{X}_k)/(S\sqrt{(1/M_j + 1/M_k)})\]

for \( N = \sum_{k=1}^{K} M_k \), then for any test level \( \alpha \),

\[
P \left[ P_{jk} \leq \alpha \text{ for any } j,k \text{ such that } \mu_j \neq \mu_k \right] < \alpha, \tag{4.10}
\]

and of the collection of tests that rejects the hypothesis \( \mu_i = \mu_j \) if \( P_{jk} \leq \alpha \) provide simultaneous test level less than or equal to \( \alpha \). Furthermore, if

\[
C_{jk} = \bar{X}_k - \bar{X}_j \pm q_{K,m,\alpha}S\sqrt{1/M_j + 1/M_k}/\sqrt{2}, \tag{4.11}
\]

then

\[
P \left[ \mu_k - \mu_j \notin C_{jk} \text{ for some } j,k \right] \leq \alpha. \tag{4.12}
\]

This method had been suggested earlier (Tukey, 1949), before the studentized range distribution had been derived.

We now proceed to analogs to one-way analysis of variance that preserve nominal test size for small samples and highly non-normal data.

### 4.2 General Rank Tests

By analogy with test (4.3), and with the scoring ideas of §3.2, we create a test by first, ranking all of the observations in data set, to obtain rank \( R_{ki} \)
for replicate \( i \) in group \( k \). One then creates non-decreasing scores \( a_1, \ldots, a_N \), assigns scores \( A_{ki} = a_{R_{ki}} \) to the ranked observations, and calculates the score sums \( \sum_{i=1}^{M_k} A_{ki} \). One might express the score sums as in (3.7), as \( T_G^{(k)} = \sum_{j=1}^{N} a_j I_j^{(k)} \), for \( I_j^{(k)} \) equal to the 1 if the item ranked \( j \) in the combined sample comes from the group \( k \), and 0 otherwise. One then creates a statistic like the denominator in (4.3),

\[
W_G = \sum_{k=1}^{K} u_k \left( T_G^{(k)} - E_0 \left[ T_G^{(k)} \right] \right)^2,
\]

for null expectations \( E_0 \left[ T_G^{(k)} \right] \) as calculated in (3.8), and quantities \( u_k \) chosen to give \( W_G \) an approximate \( \chi^2_{K-1} \) null distribution. Below in (4.19) I demonstrate that

\[
A_{ki} = \frac{N - 1}{(N^2(\hat{a} - \bar{a})^2)M_k}.
\]

In the remainder of this section, we consider the joint distribution of the \( T_G^{(k)} \), calculate their moments, and derive the asymptotic distribution for the statistic.

### 4.2.1 Moments of General Rank Sums

First and univariate second moments of \( T_G^{(k)} \) are as in §3.2.2, and are given by (3.8) and (3.9), respectively. The covariance between \( T_G^{(j)} \) and \( T_G^{(k)} \), for \( j \neq k \), can be calculated by forming a general rank statistic \( T_G^{(j,k)} \) combining both groups \( j \) and \( k \), to obtain the sum of ranks for individuals in either group \( j \) or group \( k \). Note that \( T_G^{(j,k)} = T_G^{(j)} + T_G^{(k)} \), and, furthermore, \( \text{Var} \left[ T_G^{(j,k)} \right] \) may be found by applying (3.9), with the number of observations whose ranks are summed being \( M_k + M_j \). Then

\[
\frac{(N - M_k - M_j)(M_k + M_j)}{(N - 1)}(\hat{a} - \bar{a})^2 = \text{Var} \left[ T_G^{(j,k)} \right]
\]

\[
= \text{Var} \left[ T_G^{(j)} \right] + \text{Var} \left[ T_G^{(k)} \right] + 2\text{Cov} \left[ T_G^{(j)}, T_G^{(k)} \right]
\]

\[
= \frac{(N - M_j)M_j}{(N - 1)}(\hat{a} - \bar{a})^2 + \frac{(N - M_k)M_k}{(N - 1)}(\hat{a} - \bar{a})^2 + 2\text{Cov} \left[ T_G^{(j)}, T_G^{(k)} \right]
\]

and

\[
\text{Cov} \left[ T_G^{(j)}, T_G^{(k)} \right] = \frac{-M_jM_k}{(N - 1)}(\hat{a} - \bar{a})^2.
\]

### 4.2.2 Construction of a Chi-Square-Distributed Statistic

This subsection shows that the sum of squares of deviations of general rank scores, properly standardized, is well-approximated by a \( \chi^2 \) distribution.
The use of the normal approximation for the distribution of two-sample general rank statistics was justified at the end of §3.4.1. This argument justified the univariate normality of rank sums associated with one group at a time; below we will need multivariate normality of all of the sums of ranks over the various groups. Hájek (1960), while proving more general distributional finite population sampling results, notes that the results similar to those of Erdős and Réyni (1959) can be applied to all linear combinations of rank sums from separate groups, and hence the collection of rank sums is approximately multivariate normal. This result requires some condition forcing all group proportions to stay away from zero; \( \lim \inf_{N \to \infty} M_k/N > 0 \) should suffice.

One might use as a test statistic the differences between rank score sums by group and their null expected values, squared, divided by their variances, and then summed over group. One needs to adjust for correlation between rank sums by group. The result below shows that this is done by dropping \(-M_k\) from denominator of fraction.

Statistics formed by squaring observed deviations of data summaries from their expected values, and summing, arise in various contexts in statistics. Often these statistics are well-approximated by a \( \chi^2 \) distribution. For example, in standard parametric one-way analysis of variance, sums of group means of normally-distributed data, squared, have a \( \chi^2 \) distribution, after rescaling by the population variance. Tests involving the multinomial distribution also often have the \( \chi^2 \) distribution as an approximate referent. In both these cases, as with the rank test developed below, the \( \chi^2 \) distribution has as its degrees of freedom something less than the number of quantities added. In the analysis of variance case, the \( \chi^2 \) approximation to the test statistic may be demonstrated by considering the full joint distribution of all group means, and integrating out the distribution of the grand mean. In the multinomial case, the \( \chi^2 \) approximation to the test statistic may be demonstrated by treating the underlying cell counts as Poisson, and conditioning on the table total. In the present case, one might embed the fixed rank sum for all groups taken together into a larger multivariate distribution, and either conditioning or marginalizing, but this approach is unnatural. Instead, below a smaller set of summaries, formed by dropping rank sums for one of the groups, is considered. This yields a distribution with a full-rank variance matrix, and calculations follow.

The total number of observations is \( N = \sum_{k=1}^{K} M_k \). Let \( Y \) be the \( K-1 \) by 1 matrix

\[
\left( \frac{T_G^{(1)} - E[T_G^{(1)}]}{\sqrt{M_1}}, \ldots, \frac{T_G^{(K-1)} - E[T_G^{(K-1)}]}{\sqrt{M_{K-1}}} \right)^\top \omega
\]  

(note excluding the rank sum for the final group), for \( \omega = \sqrt{(N-1)/[N(\hat{a} - \bar{a}^2)]} \). The covariances between components \( j \) and \( k \) of \( Y \) are \(-\sqrt{M_j M_k}/N\), and variance of component \( j \) is \( 1 - M_j/N \). Let \( \nu \) be the \( K-1 \) by 1 matrix

\[

\]
General Rank Tests

\((\sqrt{M_1/N}, \ldots, \sqrt{M_{K-1}/N})^\top\). Then

\[ \text{Var}[Y] = I - \nu\nu^\top. \]  

This proof will proceed by analytically inverting \(\text{Var}[Y]\). As part of this calculation, evaluation of \(\nu^\top\nu\) will be needed. Note that 

\[ \nu^\top\nu = \sum_{j=1}^{K-1} \frac{M_j}{N} = 1 - M_K/N. \]

Then

\[ \text{Var}[Y](I + (N/M_K)\nu\nu^\top) = I + (1 + (N/M_K)\nu^\top\nu)\nu\nu^\top = I \]  

and so

\[ \text{Var}[Y]^{-1} = I + (N/M_K)\nu\nu^\top. \]

Hence

\[ Y^\top(I + (N/M_K)\nu\nu^\top)Y \sim \chi^2_{K-1}. \]

Also,

\[
Y^\top\left(I + \frac{N}{M_K}\nu\nu^\top\right)Y = \omega^2 \sum_{j=1}^{K-1} \frac{(T_G^{(j)} - M_j\bar{a})^2}{M_j} \\
+ \omega^2 \left(\sum_{j=1}^{K-1} \frac{(T_G^{(j)} - M_j\bar{a})^2}{M_j} \right)^2 / M_K \\
= \omega^2 \left(\sum_{j=1}^{K-1} \frac{(T_G^{(j)} - M_j\bar{a})^2}{M_j} + \frac{(T_G^{(K)} - M_K\bar{a})^2}{M_K} \right) \\
= \frac{N - 1}{(\bar{a} - \bar{a^2})N} \sum_{j=1}^{K} \frac{(T_G^{(j)} - M_j\bar{a})^2}{M_j}. \]

The above calculation required some notions from linear algebra. The calculation (4.16) requires an understanding of the definition of matrix multiplication, and the associative and distributive properties of matrices, and (4.17) requires an understanding of the definition of a matrix inverse. Observation (4.18 is deeper; it requires knowing that a symmetric non-negative definite matrix may be decomposed as \(V = DD^\top\), for a square matrix \(D\), and an understanding that variances matrices in the multivariate case transform as do scalar variances in the one-dimensional case.

One might compare this procedure to either the standard analysis of variance (ANOVA) procedure, which is heavily reliant on distribution of responses, Alternatively, one might perform the ANOVA analysis on ranks; this procedure does not depend on distribution of responses.
4.3 The Kruskal-Wallis Test

A simple case of the general multivariate rank statistic (4.19) may be constructed by choosing the scores for the rank statistics to be the identity, with the ranks themselves as the scores.

Kruskal and Wallis (1952) introduced the test that rejects the null hypothesis of equal distributions when the test statistic (4.19) exceeds the appropriate quantile from the null $\chi^2_{K-1}$ distribution. They apply this with identity ranks. Using (3.19), $\bar{a} - \bar{a}^2 = (N^2 - 1)/12$, and the statistic simplifies to

$$W_H = \frac{12/((N + 1)N)}{\sum_{k=1}^{K} (R_k - M_k(N + 1)/2)^2/M_k}$$  \hspace{1cm} (4.20)

exceeds the appropriate critical value. This test is called the Kruskal-Wallis test, and is often referred to as the $H$ test. Here, again, $R_{ki}$ is the rank of $X_{ki}$ within the combined sample, and $R_k = \sum_{i=1}^{M_k} R_{ki}$, and (3.19) gives the first multiplicative factor.

4.3.1 Kruskal-Wallis approximate critical values

Critical values for the Kruskal-Wallis test are often a $\chi^2_{K-1}$ quantile. Let $G_k(w; \xi)$ represent the cumulative distribution function for the $\chi^2$ distribution with $k$ degrees of freedom and non-centrality parameter $\xi$, evaluated at $w$. Let $G_k^{-1}(\pi, \xi)$ represent the quantile function for this distribution. Then the critical value for the level $\alpha$ test given by statistic (4.20) is $G_k^{-1}(1 - \alpha; 0)$, and the $p$-value is given by $G_k(W_H; 0)$.

Example 14 Andrews and Herzberg (1985, Example 58) present corn (maize) yields resulting from various fertilizer treatments. Test the null hypothesis that weights associated with various fertilizer combinations have the same distribution, vs. the alternative hypothesis that a measure of location varies among these groups. Treatment is a three-digit string representing three fertilizer components. The data are at http://lib.stat.cmu.edu/datasets/Andrews/T58.1. Fields in this file are separated by space. The first three fields are example, table, and observation number. The fourth and following fields are location, block, plot, treatment, ears of corn, and weight of corn. The yield for one of the original observations 36 was missing (denoted by -9999 in the file), and is omitted in this analysis. We calculate the Kruskal-Wallis test, with 12 groups, and hence 11 degrees of freedom, with 35 observations. Rank sums by treatment are in Table 4.1. Subtracting expected rank sums from observed rank sums, squaring, and dividing by the number of observations in the group gives 971.875. Hence $H = 971.875 \times 12/(35 \times 36) = 9.256.$
The Kruskal-Wallis Test

Comparing this to a $\chi^2$ distribution gives the $p$-value 0.598. Do not reject the null hypothesis of equality of distribution. This might have been done in R using:

```r
maize<-as.data.frame(scan("T58.1",what=list(exno=0,tabno=0,
    lineno=0,loc="",block="",plot=0,trt="",ears=0, wght=0)))
maize$wght[maize$wght==-9999]<-NA
maize$nitrogen<-as.numeric(substring(maize$trt,1,1))
# Location TEAN has no tied values. R does non-intuitive
# things with ranks of missing values. Remove missing values.
tean<-maize[(maize$loc=="TEAN")&(!is.na(maize$wght)),]
cat('n
Kruskal Wallis H Test for Maize Data \n')
kruskal.test(split(tean$wght,tean$trt))
#Alternative R syntax:
#kruskal.test(tean$wght,tean$trt)

This might be compared with analysis of variance:

#Note that treatment is already a factor.
anova(lm(wght~trt,data=tean))

and with analysis of variance of the ranks:
anova(lm(rank(wght,na.last=NA)~trt,data=tean))

These last two tests have $p$-values 0.705 and 0.655 respectively. Note the difference between these normal theory results and the Kruskal-Wallis test.

Figure 4.1 shows the support of the Kruskal-Wallis statistic on the set of possible group rank sums for a hypothetical very small data set; the contour of the approximate critical region for the test of level 0.05 is superimposed.
As is the case for the chi-square test for contingency table, points enter the critical region as the level increases in an irregular way, and so constructing an additive continuity correction to (4.20) is difficult. Yarnold (1972) constructs a continuity correction that is additive on the probability, rather on the statistic, scale. Furthermore even though group sizes are very small, the sample space for the group-wise rank sums is quite rich. This richness of the sample space, as manifest by the small ratio of the point separation (in this case, 1) to the marginal standard deviations (the square roots of the variance in (3.20)), implies that continuity correction will have only very limited utility (Chen and Kolassa, 2018).

FIGURE 4.1: Asymptotic Critical Region for Kruskall Wallis Test, level 0.05

4.4 Other Scores for Multi-Sample Rank Based Tests

One might generalized the Kruskal-Wallis test in many of the same ways as one generalized the Mann-Whitney-Wilcoxon test. One might using scoring ideas as before. In (4.20) replace $R_{ki}$ with the scores $a_{R_{ki}}$. Options include van der Waerden scores, Savage scores, and others as described earlier. This provides an adjustment for ties, by letting the scores for the untied entries be the original ranks, and the scores for the tied entries be the average ranks.

Figure 4.2 shows the support of the Kruskal-Wallis statistic on the set of possible normal scores sums for a hypothetical very small data set. Compare this figure to Figure 4.1, in which sample points for group-wise score sums are
Other Scores for Multi-Sample Rank Based Tests

FIGURE 4.2: Distribution of Normal Scores

far fewer, because more rearrangements of group identifiers lead to the same scores sums. Hence the normal scores distribution shows less discreteness.

Example 15 Revisiting the TEAN subset of the maize data of Example 14, one might perform the van der Waerden and Savage score tests,

```r
library("exactRankTests") # Gives savage and vw scores
source("common.R") # Gives asymptotic score test.
cat("Other scoring schemes: Normal Scores\n")
genmultscore(teen$wght, teen$strt,
  cscores(teen$wght, type="Normal"))
cat("Other scoring schemes: Savage Scores\n")
genmultscore(teen$wght, teen$strt,
  cscores(teen$wght, type="Savage"))
```

The p-values for normal and Savage scores are 0.9544 and 0.9786 respectively.

One might also apply a permutation test in this context. In this case, the original data are used in place of ranks. The reference distribution arises from the random redistribution of group labels among the observed responses. A normal approximation to this sampling distribution leads to analysis similar to that of an analysis of variance.
**Example 16** Revisiting the TEAN subset of the maize data of Example 14, the following syntax performs the exact permutation test, again using package MultNonPram.

```r
#date()
#aov.P(tean$wght[!is.na(tean$wght)],
#      tean$trt[!is.na(tean$wght)])
#date()
```

The `date` commands bracketing the call to `aov.P` allow calculation of elapsed time. However, these calculations are quite slow, and hence are commented out. The commands below approximate the p-value via simulation.

```r
obsp<-anova(lm(wght~trt,data=tean))[[4]][1]
out<-rep(NA,10000)
#Monte Carlo approximation to the permutation distribution
for(i in seq(length(out))){
  out[i]<-anova(lm(sample(wght)~trt,data=tean))[[4]][1]
}
mean(out>=obsp)
```

giving an approximation to the p-value of 0.7254, or, more carefully, $0.7254 \pm 1.96 \sqrt{0.7254 \times 0.2746/10000} = (0.717, 0.734)$.

---

### 4.5 Multiple Comparisons

The normal-theory multiple comparison techniques of §4.1.2 may be adapted to rank-based testing. The LSD method is adapted by substituting the Kruskal-Wallis test (4.20) of §4.3 for the Analysis of Variance test (4.3) in the first stage of the procedure, and substituting the Mann-Whitney-Wilcoxon test (3.12) for the two-sample $t$ test, with the same lack of Type I error control.

In the case of rank testing, when sample sizes are equal, the studentized range method may be applied to rank means for simultaneous population differentiation (Dunn, 1964); Conover and Iman (1979) credits this to Nemenyi (1963). This technique may be used to give corrected p-values and corrected simultaneous confidence intervals for rank means. Since rank mean expectations are generally not of interest, the application of the studentized range distribution to rank means is typically of direct interest only for testing. Use
Multiple Comparisons

\( T_G^{(k)} / M_k \) in place of \( \bar{X}_k \). Note that for \( j \neq k \),

\[
\text{Var} \left[ \frac{T_G^{(k)}}{M_k} - \frac{T_G^{(j)}}{M_j} \right] = \frac{\text{Var} \left[ T_G^{(k)} \right]}{M_k^2} + \frac{\text{Var} \left[ T_G^{(j)} \right]}{M_j^2} - 2 \frac{\text{Cov} \left[ T_G^{(j)}, T_G^{(k)} \right]}{M_j M_k} \\
= \left( \frac{N - M_j}{M_j} + \frac{N - M_k}{M_k} + 2 \right) \frac{\hat{a} - \bar{a}^2}{N - 1} \\
= \left( \frac{1}{M_j} + \frac{1}{M_k} \right) N(\hat{a} - \bar{a}^2) \frac{1}{N - 1}.
\]

Hence \( S \) in (4.9) may be replaced with \( \sqrt{N(\hat{a} - \bar{a}^2)/(N - 1)} \) to obtain simultaneous \( p \)-values to satisfy (4.10). Also, take the denominator degrees of freedom to be \( \infty \) as the second argument to \( q \). The same substitution may be made in (4.11) to obtain (4.12), but the parameters bounded in these intervals are differences in average rank, which are seldom of interest.

**Example 17** Consider again the yarn data of Example 6. Consider just type A, and explore pairwise bobbin differences. One might do all pairwise Mann-Whitney-Wilcoxon tests.

\[
yarna<-yarn[yarn$type=='A',] \\
cat('Multiple Comparisons for Yarn with No Correction
') \\
pairwise.wilcox.test(yarna$strength,yarna$bobbin,exact=F, p.adjust.method="none")
\]

This gives pairwise \( p \)-values

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.112</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>0.470</td>
<td>1.000</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>0.245</td>
<td>0.030</td>
<td>0.112</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>0.885</td>
<td>0.312</td>
<td>0.772</td>
<td>0.470</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>0.661</td>
<td>0.146</td>
<td>0.470</td>
<td>0.042</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Note that bobbins 2 and 4 seem to be the most different, followed by 4 and 6, but the above comparison is not adjusted for multiple comparisons.

One might perform the version of the Fisher LSD approach using the Kruskal-Wallis test for pairwise comparisons, as described by Higgins (2004):

\[
\text{library(MultNonParam)} \\
higgins.fisher.kruskal.test(yarna$strength,yarna$bobbin)
\]

In this case, the initial Kruskal-Wallis test fails to reject the null hypothesis of equality of distribution, and no further exploration is performed.
You can perform pairwise Mann-Whitney-Wilcoxon tests. These pairwise tests can be corrected for continuity using any correction method that depends only on \( p \)-values; this excludes Tukey or Scheffé.

```r
cat("Bonferroni Comparisons for Yarn Type A Data
")
pairwise.wilcox.test(yarna$strength, yarna$bobbin, exact=F, p.adjust.method="bonferroni")
```

This gives corrected \( p \)-values:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
<td>0.46</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.59</td>
<td>1.00</td>
</tr>
</tbody>
</table>

None are significant after correction for continuity. One might also use Tukey’s method for calculating \( p \)-values respecting multiple comparisons:

```r
#The following command is also from package MultNonParam
tukey.kruskal.test(yarna$strength, yarna$bobbin)
```

indicating no significant differences.

### 4.6 Ordered Alternatives

Again consider the null hypothesis \( \theta_1 = \cdots = \theta_K \). Test this hypothesis vs. the ordered alternative hypothesis

\[
H_A : F_i(x) \geq F_{i+1}(x) \forall x
\]  

(4.21)

for all indices \( i \in \{1, \ldots, K-1\} \), with strict equality at some \( x \) and some \( i \). This alternative reduces parameter space to \( 1/2^K \) of former size. One might use as the test statistic

\[
J = \sum_{i<j} U_{ij},
\]

(4.22)

where \( U_{ij} \) is Mann-Whitney-Wilcoxon statistic for testing groups \( i \) vs. \( j \). Reject the null hypothesis when \( J \) is large. This statistic may be expressed as \( \sum_{k=1}^K c_k R_k \), plus a constant, for some \( c_k \) satisfying (4.5); that is, \( J \) may be defined as a contrast of the rank means, the approach of this subsection may be viewed as the analog of the parametric approach of §4.1.1.

Critical values for \( J \) can be calibrated using asymptotic normality. Under
the null hypothesis, the expectation of $J$ is

$$E_0 [J] = \sum_{i<j} M_i M_j / 2 = N^2 / 4 - \sum_i M_i^2 / 4, \quad (4.23)$$

and the variance is

$$\text{Var}_0 [J] = \frac{1}{12} \sum_{i=2}^K \text{Var}_0 [U_i] = \frac{1}{12} \sum_{i=2}^K M_i m_{i-1} (m_i + 1); \quad (4.24)$$

here $U_i$ is the Mann-Whitney statistic for testing group $i$ vs. all preceding groups combined, and $m_i = \sum_{j=1}^i M_j$. The second equality in (4.24) follows from independence of the values $U_i$ (Terpstra, 1952). A simpler expression for this variance is (Terpstra, 1952)

$$\text{Var}_0 [J] = \frac{1}{72} \left[ N(N+1)(2N+1) - \sum_{i=1}^K M_i (M_i+1)(2M_i+1) \right]. \quad (4.25)$$

This test might be corrected for ties, and has certain other desirable properties (Terpstra, 1952).

Jonckheere (1954), apparently independently, suggested a statistic that is twice $J$, centered to have zero expectation, and calculates the variance, skewness, and kurtosis. This test is generally called the (Jonckheere-Terpstra Test).

**Example 18** Consider again the Maize data from area TEAN in Example 14. The treatment variable contains three digits; the first indicated nitrogen level, with four levels, and is extracted in the code in in Example 14. We apply the Jonckheere-Terpstra Test of library(clinfun). One can perform

```r
library(clinfun)
jonkheere.test(tean$wght,tean$nitrogen)
cat(‘\n K-W Test for Maize, to compare with JT  \n’
kruskal.test(tean$wght,tean$nitrogen)
```

to perform this test, and the comparative three degree of freedom Kruskal-Wallis test. The Jonckheere-Terpstra Test gives a p-value 0.3274, as compared with the Kruskal-Wallis p-value 0.4994.
4.7 Powers of Tests

In this section I consider power of tests calculated from linear and quadratic combinations of indicators

\[ I_{ik,jl} = \begin{cases} 
1 & \text{if } X_{ik} < X_{jl} \\
0 & \text{if } X_{ik} > X_{jl} 
\end{cases} \tag{4.26} \]

The Jonckheere-Terpstra statistic (4.22) is of this form, as is the Kruskal-Wallis statistic (4.20), since the constituent rank sums can be written in terms of pairwise variable comparisons. Powers will be expressed in terms of the expectations

\[ \kappa_{ij} = P_A [X_{i1} < X_{j1}] \text{ for } i \neq j, \text{ and } \kappa_{ii} = 1/2. \tag{4.27} \]

Under the null hypothesis of equal populations, \( \kappa_{ij} = 1/2 \) for all unequal \( i \) and \( j \).

In the case of multidimensional alternative hypotheses, effect size and efficiency calculations are more difficult. In the case with \( K \) ordered categories, there are effectively \( K - 1 \) identifiable parameters, since, because the location of the underlying common null distribution for the data is unspecified, group location parameters \( \theta_j \) can all be increased or decreased by the same constant amount while leaving the underlying model unchanged. On the other hand, the notion of relative efficiency requires calculating an alternative parameter value corresponding to, at least approximately, the pre-specified power, as specified by (2.17). This single equation can determine only a single parameter value, and so relative efficiency calculations in this section will consider alternative hypotheses of the form

\[ \theta^A = \Delta \theta^1. \tag{4.28} \]

4.7.1 Power of Tests for Ordered Alternatives

Tests for ordered alternatives are based on statistic \( J \) of (4.22) in §4.6 have an approximate normal distribution as in (2.9), and so its power is approximated by (2.11). Under both null and alternative hypotheses,

\[ E[J] = \sum_{i<j} M_i M_j \kappa_{ij}, \tag{4.29} \]

with \( \kappa_{ij} \) defined as in (4.27). Alternative values for \( \kappa_{ij} \) under shift models (4.2) are calculated as in (3.23). Without loss of generality, one may take \( \theta_1 = 0 \).

Consider parallels with the two-group setup of §3. The cumulative distribution function \( F_1 \) of (4.2) corresponds to \( F \) of (3.1), and \( F_2 \) corresponds to \( G \) of (3.1). Then \( \mu(\theta) \) of (3.23) corresponds to \( \kappa_{12} \). Calculation of \( \kappa_{kl} \), defined
Powers of Tests

in (4.27), and applied to particular pairs of distributions, such as the normal
in (3.24) and the logistic in (3.25), and other calculations from the exercises
of §3, hold in this case as well. Each of the difference probabilities
\[ \kappa_{kl} \]
for \( k \neq l \), depends on the alternative distribution only through
\( \theta_l - \theta_k \).

Power may be calculated from (2.11).

Example 19 Consider \( K = 3 \) groups of observations, normal, with unit
variance and expectations \( \theta_1 = 0 \), \( \theta_2 = 1/2 \), and \( \theta_3 = 1 \), and all groups
of size \( M_i = 20 \). Consider a one-sided level 0.025 test. Applying (3.24),
\[ \kappa_{12} = \kappa_{23} = \Phi(0.5/\sqrt{2}) = 0.638, \text{ and } \kappa_{13} = \Phi(1/\sqrt{2}) = 0.760. \]
The null and alternative expectations of \( J \) are
\[ 60^2[(1/3)(1/3) \times 0.5 + (1/3)(1/3) \times 0.5 + (1/3)(1/3) \times 0.5] = 600 \]
and
\[ 60^2[(1/3)(1/3) \times 0.638 + (1/3)(1/3) \times 0.638 + (1/3)(1/3) \times 0.760] = 814.6 \]
respectively, from (4.29). The null variance of \( J \), from (4.25), is
\[ \varsigma^2(0) = \frac{1}{72} \left[ N(N+1)(2N+1) - \sum_{i=1}^{K} M_i(M_i+1)(2M_i+1) \right] \]
\[ = (60 \times 61 \times 1211 - 3 \times 20 \times 21 \times 41)/72 = 5433.3 \]
Applying (2.11), power is
\[ 1 - \Phi((600 - 814.6)/\sqrt{5433.3} + 1.96) = 0.829. \]
I used (2.11), rather than (2.13), since the variance of the distribution of
the statistic was most naturally given above without division by sample
size, and rather than (2.10), because calculating the statistic variance
under the alternative is tedious.

This may be computing using the R package MultNonParam using
terpstrapower(rep(20,3),(0:2)/2,"normal")
This approximate power may compared to a value determined by
Monte Carlo; this value is 0.857.

4.7.2 Power of Tests for Unordered Alternatives

Power for unordered alternatives are not most directly calculated as an ex-
tension of (2.15). In this unordered case, as noted above, the approximate
null distribution for \( W_H \) is \( \chi^2_{K-1} \). One might attempt to act by analogy with
(2.10), and calculate power using alternative hypothesis expectation and vari-
ance matrix. This is possible, but difficult, since \( W_H \) (and the other approxi-
mately \( \chi^2_{K-1} \) statistics in this chapter) are standardized using the null hy-
pothesis variance structure. Rescaling this under the alternative hypothesis
gives a statistic that may be represented approximately as the sum of squares
of independent normal random variables; however, not only do these variables
have non-zero means, which is easily addressed using a non-central \( \chi^2 \) argument as in (1.3), but also leads unequal variances for the variables comprising the summands; an analog to (1.3) would have unequal weights attached to the squared summands.

A much easier path to power involves analogy with (2.15): Approximate the alternative distribution using the null variance structure. This results in the non-central chi-square approximation using (1.4).

The variance matrix for rank sums comprising \( W_H \) is singular (that is, it does not have an inverse), and the argument justifying (1.4) relied on the presence of an inverse. The argument of §4.2.2 calculated the appropriate quadratic form, dropping one of the categories to obtain an invertible variance matrix, and then showed that this quadratic form is the same as that generating \( t \). The same argument shows that the appropriate non-centrality parameter is

\[
\xi = \left( 12 / [(N + 1)N] \right) \sum_{k=1}^{K} \left( \sum_{l=1, l \neq k}^{K} M_l \kappa_{kl} \right)^2 / M_k.
\]

Because the Mann-Whitney and Wilcoxon statistics differ only by an additive constant, the Kruskal-Wallis test may be re-expressed as

\[
\left\{ \sum_{k=1}^{K} \left( T_k - M_k(N - M_k)\kappa^\circ \right)^2 / M_k \right\} / \psi^2(N + 1)N, \tag{4.30}
\]

Here \( \kappa^\circ = 1/2 \); this is the null value of \( \kappa_{kl} \), and the null hypothesis specifies that this does not depend on \( k \) or \( l \). Also,

\[
T_k = \sum_{j=1}^{M_k} \sum_{l=1, l \neq k}^{K} \sum_{i=1}^{M_l} I(X_{kj} > X_{li}), \tag{4.31}
\]

the Mann-Whitney statistic for testing whether group \( k \) differs from all of the other groups, with all of the other groups collapsed. The non-centrality parameter can be re-expressed as

\[
\xi = \left\{ \sum_{k=1}^{K} \left( \sum_{l=1, l \neq k}^{K} M_l \kappa_{kl} \right)^2 / M_k \right\} / \psi^2(N + 1)N, \tag{4.32}
\]

where \( \mathbb{E}_A[T_k] = M_k \sum_{l=1, l \neq k}^{K} M_l \kappa_{kl} \). The non-centrality parameter is

\[
\xi = \frac{1}{\psi^2(N + 1)N} \sum_{k=1}^{K} M_k \left( \sum_{l=1, l \neq k}^{K} M_l \left( \kappa_{kl} - \kappa^\circ \right) \right)^2
\]

\[
= \frac{1}{\psi^2(N + 1)N} \sum_{k=1}^{K} M_k \left( \sum_{l=1}^{K} M_l \left( \kappa_{kl} - \kappa^\circ \right) \right)^2. \tag{4.32}
\]
The restriction on the range of the inner summation \( l \neq k \) may be dropped in (4.32), because the additional term is zero.

Let \( G_k \) and \( G_k^{-1} \) be the chi-square cumulative distribution function and quantile function, as in §4.3.1. The power for a the Kruskal-Wallis test with \( K \) groups, under alternative (4.2), is

\[
1 - G_{K-1}(G_{K-1}^{-1}(1 - \alpha, 0); \xi),
\]

(4.33)

with \( \xi \) given by (4.32).

Example 20  Continue example 19. Again, consider \( K = 3 \) groups of observations, normal, with unit variance and expectations \( \theta_1 = 0, \theta_2 = 1/2, \) and \( \theta_3 = 1, \) and all groups of size \( M_i = 20. \) The three inner sums in (4.32) are

\[
20 \times (0.5 - 0.5) + 20 \times (0.363 - 0.5) + 20 \times (0.760 - 0.5) = 7.968, \quad 20 \times (0.362 - 0.5) + 20 \times (0.5 - 0.5) + 20 \times (0.638 - 0.5) = 0, \quad \text{and} \quad 20 \times (0.240 - 0.5) + 20 \times (0.362 - 0.5) + 20 \times (0.5 - 0.5) = -7.968.
\]

Squaring, multiplying each of these by \( M_k = 20, \) and adding gives 2539.6. Multiplying by 12/(60 \times 61) gives 8.327. The critical value for a test of level 0.05 is given by the \( \chi^2 \) distribution with 2 degrees of freedom, and is 5.99. The tail probability associated with the non-central \( \chi^2 \) distribution with non-centrality parameter 8.327 and two degrees of freedom, beyond 5.99, is 0.736; this is the power for the test. As expected, this power is less than that given in example (19) for the Jonckheere-Terpstra test. This might be compared with a Monte Carlo approximation of 0.770.

This may be computing using the R package MultNonParam using kwpower(rep(20,3),(0:2)/2,"normal").

Approximation (4.32) may be approximated to give a simpler relation between the non-centrality parameter and sample size, allowing for the calculation of the sample size producing a desired power, denoted in this subsection as \( 1 - \beta. \) In (4.33), sample size enters only through the non-centrality parameter. As in standard one-dimensional sample size calculations, re-express the relation between power and non-centrality as

\[
G_{K-1}^{-1}(\beta; \xi) = G_{K-1}^{-1}(1 - \alpha, 0).
\]

(4.34)

From (4.32),

\[
\xi \approx \left\{ \sum_{k=1}^{K} \lambda_k \left( \sum_{l=1}^{K} \lambda_l (\kappa_{kl} - \kappa^0) \right)^2 \right\} N/\psi^2,
\]

(4.35)

for \( \psi = 1/\sqrt{12}, \) and

\[
N \approx \xi \psi^2 / \left\{ \sum_{k=1}^{K} \lambda_k \left( \sum_{l=1}^{K} \lambda_l (\kappa_{kl} - \kappa^0) \right)^2 \right\},
\]

(4.36)

for \( \lambda_k = \lim_{N \to \infty} M_k/N. \) Assume that \( \lambda_k > 0 \) for all \( k. \) Hence, to determine
the sample size needed for a level $\alpha$ test to obtain power $1-\beta$, for an alternative with group differences $\kappa_{kl}$, and with $K$ groups in proportions $\lambda_k$, first solve (4.34) for $\xi$, and then apply (4.36). An old-style approach to solving (4.34) involves examining tables of the sort in Haynam et al. (1982) and Haynam et al. (1982).

**Example 21** Again, consider $K = 3$ groups of observations, normal, with unit variance and expectations $\theta_1 = 0, \theta_2 = 1/2$, and $\theta_3 = 1$ Calculate the sample size needed for the level $\alpha$ Kruskal-Wallis test involving equal-sized groups to obtain power 0.8. Quantities $\kappa_{kl}$ were calculated in example 19 to be 0.362, 0.5, and 0.638. The three inner sums in (4.36) are $(0.5-0.5)/3+0.363-0.5)/3+(0.760-0.5)/3=0.133$, $(0.362-0.5)/3+(0.5-0.5)/3+0.638-0.5)/3=0$, and $(0.240-0.5)/3+(0.362-0.5)/3+(0.5-0.5)/3=-0.133$. Squaring, multiplying each of these by $\lambda_k=1/3$, adding, and multiplying by 12 gives 0.1408. Solving the equation (4.34) gives $\xi = 9.63$, and the sample size is 9.63/0.1408 $\approx$ 69, indicating 29 subjects per group.

This may be computing using the R package MultNonParam using kwsamplesize((0:2)/2,"normal").

As in the one-dimensional case, one can solve for effect size by approximating the probabilities as linear in the shift parameters. Express

$$\kappa_{kl} - \kappa^2 \approx \kappa'(\theta_k^A - \theta_l^A),$$

and explore the multiplier $\Delta$ from (4.28) giving the alternative hypothesis in a given direction. Then

$$N \approx \frac{\xi \psi^2}{\left(\kappa' \sum_{k=1}^K \sum_{j=1}^K \lambda_k \lambda_j \theta_k^A - \theta_l^A \right)^2} = \frac{\xi \psi^2}{\Delta^2 (\kappa')^2 \zeta^2},$$

for $\zeta = \sqrt{\sum_{k=1}^K \lambda_k \left(\theta_k^l \right)^2 - \left(\sum_{k=1}^K \lambda_k \theta_k^l \right)^2}$; $\zeta$ plays a role analogous to its role in §3.8, except that here it incorporates the vector giving the direction of departure from the null hypothesis.

**Example 22** Under the same conditions in Example 21, one might use equation (4.38) rather than equation (4.36). Take $\theta^1 = \theta^A = (0, 1/2, 1)$ and $\Delta = 1$. The derivative $\kappa'$ is tabulated, for the normal and logistic distributions, in Table 3.6 as $\mu'(0)$. In this normal case, $\kappa' = \mu'(0) = (2/\sqrt{\pi})^{-1} = 0.282$. Also, $\zeta^2 = (0^2/3 + (1/2)^2/3 + 1^2/3 - (0/3 + (1/2)/3 + 1/3)^2) = 5/12 - 1/4 = 1/6$. The non-centrality parameter, solving (4.34), is $\xi = 9.63$. The approximate sample size is 9.63/(12 $\times$ 12 $\times$ 0.2822/6) $= 61$, or approximately 21 observations per group; compare this with an
approximation of 23 per group using a normal approximation, without expanding the $\kappa_{ij}$.

This may be computing using the R package MultNonParam using \texttt{kwsamplesize((0:2)/2,"normal",taylor=TRUE)}.

Relation (4.38) may be solved for the effect size $\Delta$, to yield

$$\Delta = \psi \sqrt{\xi} / \left( \kappa' \{ N \zeta^2 \}^{1/2} \right).$$  \hfill (4.39)

The sum with respect to $j$ disappears from $\zeta^2$, since the sum of the proportions $\lambda_j$ is 1. Then, in order to determine the effect size necessary to give a level $\alpha$ test power $1 - \beta$ to detect an alternative hypothesis $\Delta \theta^\dagger$, determine $\xi$ from (4.33), and then $\Delta$ from (4.39).

\begin{example}
Again use the same conditions as in Example 21. Example 22 demonstrates that $\xi = 9.63$, $\kappa' = 0.282$, and the quantity in square brackets in (4.39) is $1/6$. $\Delta = (\sqrt{9.63/(60 \times 12 \times (1/6))})/0.282 = 1.004$. As $\Delta$ is almost exactly 1, the alternative parameter vector in the direction of $(0,1/2,1)$ and corresponding to a level 0.05 test of power 0.80 with 60 observations is $(0,1/2,1)$.

This may be computing using the R package MultNonParam using \texttt{kweffectsize(60,(0:2)/2,"normal")}.
\end{example}

Fig. 4.3 reflects the accuracy of the various approximations in this section. It reflects performance of approximations to the power of the Kruskal-Wallis test. Tests with $K = 3$ groups, with the same number of observations per group, with tests of level 0.05, were considered. Group sizes between 5 and 30 were considered (corresponding to total sample sizes between 15 and 90). For each sample size, (4.39) was used to generate an alternative hypothesis with power approximately 0.80, in the direction of equally-spaced alternatives. The dashed line represents a Monte Carlo approximation to the power, based on 50,000 observations. The dotted line represents the standard non-central chi-square approximation (4.33). The solid line represents this same approximation, except also incorporating the linear approximation (4.37) representing the exceedance probabilities as linear in the alternative parameters, and hence the non-centrality parameter as quadratic in the alternative parameters.

The solid line is almost exactly horizontal at the target power. The discrepancy from horizontal arises from the error in approximating (4.32) by (4.35). For small sample sizes (that is, group sizes of less than 25, or group sizes of less than 75), (4.33) is not sufficiently adequate; for larger sample sizes it should be accurate enough.

All curves in Fig. 4.3 use the approximate critical value as given in §4.3.1.
FIGURE 4.3: Approximate Powers for the Kruskall-Wallis Test

4.8 Efficiency Calculations

Power calculations for one-dimensional alternative hypotheses made use of (2.15), applying a normal approximation with exact values for means under the null and alternatives, and approximating the variance under the alternative by the variance under the null. Efficiency calculations of §2.5 approximated means at the alternative linearly using the derivative of the mean function at the null.

Consider first the one-sided Jonckheere-Terpstra test of level \( \alpha \). Let \( T_1 = J/N^2 \). Denote the critical value by \( t_1^\circ \), satisfying \( P_{\theta^0} [T_1 \geq t_1^\circ] = 1 - \alpha \).

As in (4.28), reduce the alternative hypothesis to a single dimension by letting the alternative parameter vector be a fixed vector times a multiplier \( \Delta \). The power function \( \varpi_{1,n} (\Delta) = P_{\theta^A} [T_1 \geq t_1^\circ] \) satisfies (2.9), (2.12), and (2.14), and hence the efficiency tools for one-dimensional hypotheses developed in §2.5.2 may be used. Expressing \( \mu_1 (\Delta) \) as a Taylor series with constant and linear terms,

\[
\mu_1 (\Delta) \approx \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \lambda_i \lambda_j [\kappa^\circ + \kappa' (\theta_j^A - \theta_i^A)]
\]

for \( \lambda_i = M_i/N \), where again \( \kappa^\circ \) is the common value of \( \kappa_{jk} \) under the null hypothesis, and \( \kappa' \) is the derivative of the probability in (3.23), calculated for
Exercises

various examples in §3.8.2. Hence

$$\mu'_1(0) = \sum_{i=1}^{K-1} \sum_{j=i+1}^K \lambda_i \lambda_j \kappa'(\theta_j^\dagger - \theta_i^\dagger).$$

Furthermore, from (4.25), Var $[T_1] \approx \frac{1}{36} \left[ 1 - \sum_{k=1}^{K} \lambda_k^2 \right]/N$. Consider the simple case in which $\lambda_k = 1/K$ for all $k$, and in which $\theta_j^\dagger - \theta_i^\dagger = (j - i)$. Then

$$e_j = \frac{\kappa'(K^2 - 1)/6K}{\sqrt{\frac{1}{36} [1 - 1/K^2]}} = \frac{\kappa'(K^2 - 1)}{\sqrt{K^2 - 1}} = \kappa' \sqrt{K^2 - 1}.$$

Recall that the efficacy of the normal-theory contrast, from (4.8), is $\frac{\sqrt{K^2 - 1}}{2\sqrt{\xi}}$, and the asymptotic relative efficiency is $\xi \kappa'/\psi$. This is the same as the asymptotic relative efficiency in the two-sample case in Table 3.6.

While techniques exist to consider transform non-central chi-square statistics to approximate normality (Sankaran, 1963), the most direct approach to comparing efficiency of various tests based on approximately $\chi^2$ statistics is to compare approximate sample sizes for a fixed level and power, as in (4.38). Prentice (1979) does this using ratios of the non-centrality parameter. For test $j$, let $N_j$, be the sample size needed to provide power $1 - \beta$ for the test of level $\theta^A$ against alternative $\theta^j$.

Calculations (4.32), and (4.35) through (4.39) were motivated specifically for the Kruskal-Wallis test, using Mann-Whitney sums (4.31), but hold more broadly for any group summary statistics replacing (4.31), so long as $E_0[T_k] \approx M_k \sum_{i=1, i \neq k}^K M_i \kappa^\circ$ for some differentiable $\kappa^\circ$, and as long as (4.30), with the new definition of $T_k$, is approximately $\chi^2_{K-1}$. In particular, a rescaled version of the $F$-test statistic (4.3) $(K - 1)W = \frac{\sum_{k=1}^K M_k (\bar{X}_k - \bar{X})^2}{\bar{\sigma}^2}$ is such a test, with $\psi$ the variance of the $X_{kj}$, $T_k = \bar{X}_k - \bar{X}$, $\kappa^\circ = 0$, and $\kappa' = 1$. The value of $\zeta^2$ remains unchanged for all such tests, and, for two tests having values of $\kappa'$ at the null hypothesis distinguished by indices, and similarly having the norming factors $\psi$ distinguished by indices, the ratios of sample sizes needed for approximately the same power for the same alternative, is approximately

$$N_2/N_1 \approx \left[ (\kappa'_1/\psi_1)/(\kappa'_2/\psi_2) \right]^2,$$

which is the same as as in two-sample case.

\textbf{4.9 Exercises}

1. The data set
http://ftp.uni-bayreuth.de/math/statlib/datasets/federalistpapers.txt

gives data from an analysis of a series of documents. The first column gives document number, the second gives the name of a text file, the third gives a group to which the text is assigned, the fourth represents a measure of the use of first person in the text, the fifth presents a measure of inner thinking, the sixth presents a measure of positivity, and the seventh presents a measure of negativity. There are other columns that you can ignore. (The version on line, above, has odd line breaks. A fixed version can be found at stat.rutgers.edu/home/kolassa/960-555/federalistpapers.txt).

a. Test the null hypothesis that negativity is equally distributed across the groups using a Kruskal-Wallis test.

b. Test at $\alpha = .05$ the pairwise comparisons for negativity between groups using the Bonferroni adjustment, and repeat for Tukey’s HSD.

2. The data set

http://ftp.uni-bayreuth.de/math/statlib/datasets/Plasma_Retinol

gives data relating various quantities, including smoking status (1 never, 2 former, 3 current) in column 3 and beta plasma in column 13. Perform a nonparametric test to investigate an ordered effect of smoking status on beta plasma.
Group Differences with Blocking

This chapter concerns questions raised in the previous chapter, in the presence of blocking. The simplest example of blocking is that of paired observations; paired observations are considered first.

5.1 Paired Comparisons

Suppose pairs of values \((X_i, Y_i)\) are observed on \(n\) subjects, indexed by \(i\), and suppose further that \((X_i, Y_i)\) is independent of \((X_j, Y_j)\) for \(i, j \in \{1, \ldots, n\}\), \(i \neq j\), and that all of the vectors \((X_i, Y_i)\) have the same distribution. Test the null hypothesis that the marginal distribution of \(\{X_i\}\) is the same as that of \(\{Y_i\}\), vs. the alternative hypothesis that the distributions are different. One might test these hypotheses by calculating the difference \(Z_i = X_i - Y_i\) between values, and apply the one-sample technique of Chapter 2, the sign test; in this way, we have reduced the problem to one already solved.

After applying the differencing operation, one might expect the differences to be more symmetric than either of the two original variables, and one might exploit this symmetry. To derive a test under the assumption of symmetry, let \(R_j\) be the rank of \(|Z_j|\) among all absolute values. Let

\[
S_j = \begin{cases} 
1 & \text{if item } j \text{ is positive} \\
0 & \text{if item } j \text{ is negative}
\end{cases}
\]

Define the Wilcoxon signed-rank test statistic to be

\[
T = \sum R_j S_j \quad (5.1)
\]

Under the null hypothesis that the distribution of the \(X_i\) is the same as the distribution of the \(Y_i\), and again, assuming symmetry of the differences, then \((S_j, \cdot)\) and \((R_j, \cdot)\) are independent random vectors, because \(S_j\) and \(|X_j|\) are pairwise independent under \(H_0\).

Components of the random vector \((R_1, \ldots, R_n)\) are dependent, and hence calculation of the variance from \(T\) via (5.1) requires calculation of the sum of identically distributed but not independent random variables. An alternative
formulation, as the sum of independent but not identically distributed random variables, will prove more tractable. Let

\[ V_j = \begin{cases} 
1 & \text{if the item whose absolute value is ranked } j \text{ is positive} \\
0 & \text{if the item whose absolute value is ranked } j \text{ is negative}
\end{cases} \]

Then \( T = \sum jV_j \), and

\[ E_0[T] = \sum jE_0[V_j] = n(n + 1)/4 \]  \hspace{1cm} (5.2)

and

\[ \text{Var}_0[T] = \sum j^2\text{Var}_0[V_j] = \sum j^2/4 = n(2n + 1)(n + 1)/24. \]  \hspace{1cm} (5.3)

One can also calculate exact probabilities for \( T \) recursively, as you could for the two sample statistic. There are \( 2^n \) ways to assign signs to ranks 1,...,\( n \). Let \( f(t,n) \) be the number of such assignments yielding \( T = t \) with \( n \) observations. Again, as in §3.4.1, summing the counts for shorter random vectors with alternative final values,

\[ f(t,n) = \begin{cases} 
0 & \text{for } t < 0 \text{ or } t > n(n + 1)/2 \\
1 & \text{if } n = 1 \text{ and } t \in \{0,1\} \\
f(t,n - 1) + f(t - n,n - 1) & \text{otherwise}
\end{cases} \]  \hspace{1cm} (5.4)

**Example 24** Consider data calculated on the size of brains of twins (Tramo et al., 1998). This data set from

http://lib.stat.cmu.edu/datasets/IQ_Brain_Size

contains data on 10 sets of twins. Each child is represented by a separate line in the data file, for a total of 20 lines. We investigate whether brain volume (in field 9) is influenced by birth order (in field 4). Brain volumes for the first and second child, and their difference, are given in Table 5.1. The rank sum statistic is 3+4+10+8+7 = 32, the null expected rank sum is 10×11/2 = 27.5, the null variance is 10×21×11/24 = 96.25, and so the two-sided \( p \)-value is \( 2 \times \Phi(-(32 - 27.5)/\sqrt{96.25}) = 2 \times \Phi(-0.408) = .683 \). There is no evidence that twin brain volume differs by birth order. This calculation might also have been done in R using

```r
twinbrain<-as.data.frame(scan("twinbrain.dat", what=list(CCMIDSA=0,FIQ=0,HC=0,ORDER=0,PAIR=0,SEX=0, TOTSA=0, TOTVOL=0,WEIGHT=0)))
fir<-twinbrain[twinbrain$ORDER==1,] fir$v1<-fir$TOTVOL sec<-twinbrain[twinbrain$ORDER==2,] sec$v2<-sec$TOTVOL both<-merge(fir,sec,by="PAIR")[,c("v1","v2")]
```
TABLE 5.1: Twin Brain Volume

<table>
<thead>
<tr>
<th>Pair</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>1005</td>
<td>1035</td>
<td>1281</td>
<td>1051</td>
<td>1034</td>
<td>1079</td>
<td>1104</td>
<td>1439</td>
<td>1029</td>
<td>1160</td>
</tr>
<tr>
<td>Second</td>
<td>963</td>
<td>1027</td>
<td>1272</td>
<td>1079</td>
<td>1070</td>
<td>1173</td>
<td>1067</td>
<td>1347</td>
<td>1100</td>
<td>1204</td>
</tr>
<tr>
<td>Diff.</td>
<td>-42</td>
<td>-8</td>
<td>-9</td>
<td>28</td>
<td>36</td>
<td>94</td>
<td>-37</td>
<td>-92</td>
<td>71</td>
<td>44</td>
</tr>
<tr>
<td>Rank</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>10</td>
<td>5</td>
<td>9</td>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>

both$diff<-both$v2-both$v1
wilcox.test(both$diff)
giving an exact p-value of 0.695. Compare these to the results of the sign and t-tests:

library("BSDA")#Need for sign test.
SIGN.test(both$diff)
t.test(both$diff)
giving p-values of 1 and 0.647 respectively.

As with the extension from two sample testing to multi-sample testing referred to in §3.4.2, one can extend the other rank-based modifications of §3.4.2 and §3.6 to the blocking context as well. Ties may handled by using mean ranks, and testing against a specific distribution for the alternative may be tuned using the appropriate scores. The more general score statistic is $T = \sum_j a_j S_j$; in this case, $E[T] = \sum_j a_j/2$, and $\text{Var}[T] = \sum_j a_j^2/4$.

Example 25 Perform the asymptotic score tests on the brain volume differences of Example 24.

library("MultNonParam")
cat("Asymptotic test of savage scores\n") symscorestat(both$diff,both$savagescores)#From MultNonParam cat("Asymptotic test using normal scores\n") both$normalscores<-qqnorm(seq(length(both$diff)),plot.it=F)$x symscorestat(both$diff,both$normalscores) cat("Asymptotic test using savage scores scores\n") both$savagescores<-cumsum(1/rev(seq(length(both$diff)))) symscorestat(both$diff,both$savagescores)
giving p-values of 0.863 and 0.730 respectively.

Permutation testing can also be done, using raw data values as scores.
We used the logic $X_j$ and $Y_j$ have same marginal distribution implies $Y_j - X_j$ has a symmetric distribution. You can find joint distributions for which this is not true, but these examples tend to be contrived.
5.1.1 Estimating the population median difference

Estimation based on inversion of the Wilcoxon signed-rank statistic mirrors that based on inversion of the Mann-Whitney-Wilcoxon statistic of §3.10. Note that the median is also the point of symmetry. Let $T(\theta)$ be Wilcoxon statistic calculated from the data set $Z_j(\theta) = Z_j - \theta = Y_j - X_j - \theta$, after ranking the $Z_j(\theta)$ by their absolute values, and summing the ranks with positive values of $Z_j(\theta)$. Estimate $\theta$ as the quantity $\hat{\theta}$ equating $T$ to the median of its sampling distribution; that is, $\hat{\theta}$ satisfies $T(\hat{\theta}) = n(n + 1)/4$.

After sorting the values $Z_j$, denote value at position $j$ in the ordered list by the order statistic $Z(j)$. One can express $\hat{\theta}$ in terms of $Z(j)$, by considering the behavior of $T(\theta)$ as $\theta$ varies. For $\theta > Z(n)$, $T(\theta) = 0$. When $\theta$ decreases past $Z(n)$, the $Z_i - \theta$ with the lowest absolute value switches from negative to positive, and $T(\theta)$ moves up to 1. When $\theta$ decreases past $(Z(n) + Z(n-1))/2$, then $Z(n) - \theta$ goes from having the smallest absolute value to having the second lowest absolute value, and $T$ goes to 2. The next jump in $T$ occurs if one more observation becomes positive (at $Z(n-1)$) or if the absolute value of the lowest shifted observation passes the absolute value of the next observation (at $Z(n) + Z(n-2)/2$).

Generally, the jumps happen at averages of two observations, including averages of an observation with itself. These averages are called Walsh averages. These play the same role as differences in two-sample case. First, note that $T$ is the number of positive Walsh averages. This can be seen by letting $Z[i]$ be observation ordered by absolute value, and letting $W_{ij} = (Z[i] + Z[j])/2$ for $i \leq j$. Suppose that $Z[j] > 0$. Then $Z[i] + Z[j] > 0$ for $i < j$, and all $W_{ij} > 0$ for $i < j$, and $W_{jj} > 0$. On the other hand, if $Z[j] < 0$, Then $Z[i] + Z[j] < 0$ for $i < j$, and Then all $W_{ij} < 0$ for $i < j$, and $W_{jj} < 0$.

So $\hat{\theta}$ has half of the Walsh averages below it, and half above, and $\hat{\theta}$ is median of Walsh averages. This estimator is given by Hodges and Lehmann (1963), in the same paper giving the analogous estimator for the one-sample symmetric problem of §3.10, and is called the Hodges–Lehmann estimator.

Example 26 Walsh averages may be extracted in R using

```r
aves<outer(both$diff,both$diff,"+")/2
sort(aves[upper.tri(aves,diag=TRUE)])
```

to obtain the $10 \times (10 + 1)/2 = 55$ pairwise averages

-92.0 -67.0 -64.5 -50.5 -50.0 -42.0 -39.5 -37.0 -32.0 -28.0
-25.5 -25.0 -24.0 -23.0 -22.5 -10.5 -9.0 -8.5 -8.0 -7.0
-4.5 -3.0 -0.5 1.0 1.0 3.5 9.5 10.0 13.5 14.0
14.5 17.0 17.5 18.0 26.0 28.0 28.5 31.0 31.5 32.0
36.0 36.0 40.0 42.5 43.0 44.0 49.5 53.5 57.5 61.0
65.0 69.0 71.0 82.5 94.0
Paired Comparisons

Their median is observation 28, which is 10.0. Estimate the median difference as 10.0.

5.1.2 Confidence Intervals

Confidence intervals may also be constructed using the device of (1.17), similarly as with the two-sample location shift example of §3.10. Let \( \tilde{W}_j \) be the ordered Walsh averages. Find the largest \( t^o \) such that \( P[T \leq t^o - 1] < \alpha/2 \); then \( t^o \) is the \( \alpha/2 \) quantile of the distribution of the Signed Rank statistic. Using (5.2) and (5.3), one might approximate the critical value using a normal approximation \( t^o \approx n(n + 1)/2 - z_{\alpha/2}/\sqrt{n(2(n + 1)(n + 1))/24}; \) note that this approximation uses the distribution of the signed-rank statistic, which does not depend on the distribution of the underlying data as long as symmetry holds. Recall that \( z_\beta \) is the \( 1 - \beta \) quantile of the standard normal distribution, for any \( \beta \in (0, 1) \); in particular, \( z_{\alpha/2} \) is positive for any \( \alpha \in (0, 1) \), and \( z_{0.05/2} = 1.96 \). By symmetry, \( P[T \geq n(n - 1)/2 - t^o + 1] \leq \alpha/2 \).

As noted above, \( T(\theta) \) jumps by one each time \( \theta \) passes a Walsh average. Hence the confidence interval is \( (\tilde{W}_o, \tilde{W}_{n(n+1)/2-t^o+1}) \). See Figure 5.1.
Example 27 Refer again to the brain volume data of Example 24. Find a 95% confidence interval for the difference in brain volume. The 0.025 quantile of the signed rank statistic with 10 observations is $t^0 = 9$; this can be calculated from R using `qsigrank(0.025, 10)`. Hence confidence interval endpoints are observations 9 and $55 + 1 - 9 = 47$. As tabulated in Example 26, Walsh average 9 and 47 are -32.0 and 49.5 respectively. Hence the confidence interval is (-32.0, 49.5). This might have been calculated directly in R using `wilcox.test(both$diff, conf.int=TRUE)`.

Similar techniques to those of this section were used in §2.4 to give confidence intervals for the median, and in §3.10 to give confidence intervals for median differences. In each case, a statistic dependent on the unknown parameter was constructed, and estimates and confidence intervals were constructed by finding values of the parameter equating the statistic to appropriate quantiles. However, the treatment of this section and that of §2.4 and §3.10 differ, in that the resulting statistic is a decreasing function of $\theta$ in this section, while previously it was an increasing function. The increasing parameterization of §2.4 was necessary to allow the same construction to be used in §2.4.1, when quantiles other than the median were estimated; these quantiles estimates are an increasing function of each observation separately. A parallel construction might have been used in the current section, by inverting the statistic formed by summing ranks of negative observations; as this definition runs contrary to the common definition of the Wilcoxon Signed Rank statistic, it was avoided.

5.1.3 Alternative Distribution for Wilcoxon Signed Rank Statistic

Consider alternative hypotheses $H_A$, specifying that the median of the distribution of differences is $\theta$. Then $E_\theta[T] = \sum_{i=1}^n \sum_{j=1}^i P_\theta[(Z_i + Z_j - 2\theta) \leq 0] = n(n + 1)P[Z_1 + Z_2 \geq 2\theta]/2$. To apply asymptotic relative efficiency calculations, scale the test statistic to have have an expectation that varies with the parameter value specifying the null hypothesis, and approximately independent of sample size, as in (2.9), by switching to $S = 2T/(n(n+1))$. In this case, $\mu(\theta) = P_0[Z_1 + Z_2 \geq 2\theta]$, and $\sigma^2(0) = 4n(n+1)(2n+1)/(n^2(n+1)^2)$ for distributions symmetric about 0. This will be used for asymptotic relative efficiency calculations in the exercises.
5.2 Two-Way Non-Parametric Analysis of Variance

Suppose one observes random variable $X_{kli}$, with $k \in \{1, \cdots, K\}$, $l \in \{1, \cdots, L\}$, and $i \in \{1, \cdots, M_{kl}\}$. Suppose the distribution of $X_{kli}$ are independent, with a distribution that is allowed to depend on $k$ and $l$; that is, $X_{kli} \sim F_{kl}$. We wish to test the null hypothesis that treatment has no effect, while allowing different blocks to behave differently, vs. the alternative that the distribution depends on treatment as well. That is, the null hypothesis is that $F_{kl}$ does not depend on $k$, although it may depend on $l$, and the alternative hypothesis is that for some $l$, some $j$ and $k$, and some $x$, $F_{kl}(x) \neq F_{jl}(x)$.

However, the test to be constructed will not have reasonable power unless this difference is more systematic; that is, consider alternatives for which for each pair of treatments $j$ and $k$, either $F_{jl}(x) \leq F_{jl}(x)$ for all $x$ and all $l$, or $F_{jl}(x) \geq F_{jl}(x)$ for all $x$ and all $l$. The direction of the effect must be constant across blocks. Heuristically, treat $k$ as indexing treatment and $l$ as indexing a blocking factor.

In order to build a statistic respecting treatment order across blocks, one might rank the observations within blocks; that is, let $R_{kli}$ represent the rank of $X_{kli}$ within $X_{kl1}, \ldots, X_{klM_{kl}}$, and sum them separately by block and treatment. From these rank sums one might calculate separate Kruskal-Wallis statistics for each block, and, since statistics from separate blocks are independent, one might add these Kruskal-Wallis statistics to get $\chi^2_L(K-1)$ degree of freedom statistic under the null hypothesis. Such a statistic, however, is equally sensitive to deviations from the null hypothesis in opposite directions in different blocks, and so has low power against all alternatives. Interesting alternative hypotheses involve treatment distributions in different blocks ordered in the same way.

5.2.1 Distribution of Rank Sums

In order to use these ranks to detect treatment differences that are in a consistent direction across blocks, one should consolidate rank sums over blocks before comparing across treatments. Let $R_{k..} = \sum_{l=1}^L \sum_{i=1}^{M_{kl}} R_{kli}$ and $\bar{R}_{k..} = R_{k..} / \sum_{l=1}^L M_{kl}$. Consolidation via averaging, rather than via summing, implies a choice in weighting the contributions of the various blocks that is more clearly a default choice in the case with equal numbers of observations in each block, but less clear for unbalanced data sets.

Moments of these rank sums may be calculated as an extension to (3.20). Under the null hypothesis of no treatment effect, the expectation of one rank is

$$E[R_{kli}] = (\sum_{j=1}^K M_{jl} + 1)/2,$$
and hence

\[ E[R_{k..}] = \sum_{l=1}^{L} \left[ M_{kl} \left( \sum_{j=1}^{K} M_{jl} + 1 \right)/2 \right], \quad (5.5) \]

and

The development of this test in the two-way layout remarkably preceded analogous testing in one-way layouts.

Variances of rank sums depend on the covariance structure of the ranks. Ranks that make up each sum within block are independent, but sums of ranks across blocks are dependent. Within a treatment-block cell, \( \text{Var} [R_{kl.}] \) is same as for the Mann-Whitney-Wilcoxon statistic:

\[ \text{Var} [R_{kl.}] = M_{kl} \left( \sum_{j \neq k}^{K} M_{jl} \right) \left( \sum_{j=1}^{K} M_{jl} + 1 \right)/12. \]

By independence across blocks,

\[ \text{Var} [R_{k..}] = \sum_{l=1}^{L} \left\{ M_{kl} \left( \sum_{j \neq k}^{K} M_{jl} \right) \left( \sum_{j=1}^{K} M_{jl} + 1 \right)/12 \right\}. \quad (5.6) \]

Covariances may be calculated by comparing the variance of the sum to the sum of the variances, as in (4.14), to obtain

\[ \text{Cov} [R_{kl.}, R_{ml.}] = -M_{kl} M_{ml} \left( \sum_{j=1}^{K} M_{jl} + 1 \right)/12. \]

Since blocks are ranked independently, variances and covariances for rank sums add across blocks:

\[ \text{Cov} [R_{k..}, R_{m..}] = -\sum_{l=1}^{L} M_{kl} M_{ml} \left( \sum_{j=1}^{K} M_{jl} + 1 \right)/12. \quad (5.7) \]

### 5.2.2 Two-Way Nonparametric Analysis of Variance in the Balanced Case

As with the Kruskal-Wallis test, a statistic is constructed by setting \( Y = (R_{1..} - E_{0}[R_{1..}], \ldots, R_{K-1..} - E_{0}[R_{K-1..}])^T \), using (5.5), calculating \( \text{Var}_0 [Y] \) from (5.6) and (5.7), and using

\[ W_F = Y^T \text{Var}_0^{-1} Y \quad (5.8) \]

as the test statistic. The inverse \( \text{Var}_0[Y]^{-1} \) is tractable with balanced replicates.
Friedman (1937) addresses the case in which \( M_{kl} = 1 \) for all \( k, l \). The generalization to the balanced case with replicates is trivial, and this section assumes \( M_{kl} = M \) for all \( k, l \). Here

\[
E[\bar{R}_{k..}] = (KM + 1)/2.
\]

Note

\[
\text{Var}[\bar{R}_{k..}] = (K - 1)(KM + 1)/(12L).
\]

Covariances between sets of rank means are then needed:

\[
\text{Cov}[\bar{R}_{k..}, \bar{R}_{m..}] = -(KM + 1)/(12L).
\]

In this balanced case, the test statistic (5.8) can be shown to equal sum of squares of ranks away from their average value per treatment group:

\[
W_F = 12L \sum_{k=1}^{K} [\bar{R}_{k..} - (MK + 1)/2]^2/[K(KM + 1)].
\] (5.9)

This demonstration is as for the Kruskal-Wallis test in §4.2.2. The \( \chi^2 \) limiting distribution of \( T \) is also demonstrated by observing the parallelism between this covariance structure and the covariance structure of rank sums in §4.2.2, since the covariance structure of \( Y = (\bar{R}_1, \ldots, \bar{R}_{K-1})/\sqrt{K} \) has the same form as (4.15), for some vector \( \nu \) and constant \( \omega \). A relationship like that of §4.2.2 can then be used to demonstrate that \( T \) is approximately \( \chi^2_{K-1} \).

Again, some condition on sample sizes is required; for example, in the case with the same number of replicates per treatment-block combination, having the number of blocks going to infinity suffices, and in the balanced case with a fixed number of blocks, having the number of observations in each block-treatment combination go to infinity suffices.

**Example 28** Friedman (1937) provides data on variability of expenses in various categories for families of different income levels. This data is provided in Table 5.2. This may be read into R using

\[
\text{expensesd<-as.data.frame(scan("friedman.dat", what=list(cat="",g1=0,g2=0,g3=0,g4=0,g5=0,g6=0,g7=0)))}
\]

Here \( L = 14 \), \( M = 1 \), and \( K = 7 \). Ranking within groups (for example, in R, using

\[
\text{expenserank<-t(apply(as.matrix(expensesd[,1]),1,rank)) rownames(expenserank)<-expensesd[[1]]}
\]

gives the table of ranks

<table>
<thead>
<tr>
<th>g1</th>
<th>g2</th>
<th>g3</th>
<th>g4</th>
<th>g5</th>
<th>g6</th>
<th>g7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housing</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>
Group Differences with Blocking

TABLE 5.2: Expense variability by income group, from Friedman (1937)

<table>
<thead>
<tr>
<th>Category</th>
<th>Gr. 1</th>
<th>Gr. 2</th>
<th>Gr. 3</th>
<th>Gr. 4</th>
<th>Gr. 5</th>
<th>Gr. 6</th>
<th>Gr. 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housing</td>
<td>103.30</td>
<td>68.42</td>
<td>89.53</td>
<td>77.94</td>
<td>100.00</td>
<td>108.20</td>
<td>184.90</td>
</tr>
<tr>
<td>Operations</td>
<td>42.19</td>
<td>44.31</td>
<td>60.91</td>
<td>73.90</td>
<td>43.87</td>
<td>61.74</td>
<td>102.30</td>
</tr>
<tr>
<td>Food</td>
<td>71.27</td>
<td>81.88</td>
<td>100.71</td>
<td>86.52</td>
<td>100.30</td>
<td>90.75</td>
<td>100.60</td>
</tr>
<tr>
<td>Clothing</td>
<td>37.59</td>
<td>60.05</td>
<td>56.97</td>
<td>60.79</td>
<td>71.82</td>
<td>83.04</td>
<td>117.10</td>
</tr>
<tr>
<td>Furnishings</td>
<td>58.31</td>
<td>52.73</td>
<td>96.04</td>
<td>60.42</td>
<td>104.33</td>
<td>89.78</td>
<td>85.77</td>
</tr>
<tr>
<td>Transportation</td>
<td>46.27</td>
<td>82.18</td>
<td>129.80</td>
<td>181.00</td>
<td>172.33</td>
<td>164.80</td>
<td>246.80</td>
</tr>
<tr>
<td>Recreation</td>
<td>19.00</td>
<td>23.07</td>
<td>38.70</td>
<td>45.81</td>
<td>10.63</td>
<td>15.84</td>
<td>12.50</td>
</tr>
<tr>
<td>Personal</td>
<td>8.31</td>
<td>8.43</td>
<td>9.16</td>
<td>14.28</td>
<td>50.03</td>
<td>55.18</td>
<td>55.18</td>
</tr>
<tr>
<td>Medical</td>
<td>20.15</td>
<td>33.48</td>
<td>60.08</td>
<td>69.35</td>
<td>114.34</td>
<td>45.28</td>
<td>101.60</td>
</tr>
<tr>
<td>Education</td>
<td>3.16</td>
<td>4.12</td>
<td>12.73</td>
<td>18.95</td>
<td>8.89</td>
<td>41.52</td>
<td>66.33</td>
</tr>
<tr>
<td>Community</td>
<td>4.12</td>
<td>18.87</td>
<td>8.54</td>
<td>12.92</td>
<td>25.30</td>
<td>19.85</td>
<td>16.76</td>
</tr>
<tr>
<td>Gifts</td>
<td>5.29</td>
<td>10.91</td>
<td>11.22</td>
<td>25.26</td>
<td>42.25</td>
<td>48.80</td>
<td>69.38</td>
</tr>
<tr>
<td>Other</td>
<td>6.00</td>
<td>5.57</td>
<td>22.23</td>
<td>2.45</td>
<td>1.00</td>
<td>4.00</td>
<td>4.00</td>
</tr>
</tbody>
</table>

Operations 1 3 4 6 2 5 7
Food 1 2 7 3 5 4 6
Clothing 1 3 2 4 5 6 7
Furnishings 2 1 6 3 7 5 4
Transportation 1 2 3 6 5 4 7
Recreation 1 2 3 4 7 5 6
Personal 1 2 3 6 4 7 5
Medical 1 2 4 5 7 3 6
Education 1 2 4 5 3 6 7
Community 1 5 2 3 7 6 4
Vocation 1 5 2 4 3 6 7
Gifts 1 2 3 4 5 6 7
Other 5 4 7 2 6 1 3

Rank sums are 23, 36, 53, 57, 70, 70, 83, and group means (for example, via apply(expenserank,2,mean) gives 1.643, 2.571, 3.786, 4.071, 5.000, 5.000, 5.929.
Subtracting the overall mean, 4, squaring, and adding, gives 13.3673.
Multiplying by $12L/(K(K+1)) = 12 \times 14/(7 \times 8) = 3$ gives the statistic value $W_F = 40.10$. Comparing to the $\chi^2_6$ distribution gives a p-value of $4.35 \times 10^{-7}$. This might also have been done entirely in R using

friedman.test(as.matrix(expensesd[,1:]))

Note that when $K = 2$ and $M_{kl}$ are all 1, Friedman’s test is equivalent to the sign test applied to differences within block.
5.2.3 Two-Way Nonparametric Analysis of Variance in the Unbalanced Case

This analysis fails in the unbalanced case, because no expression for the variance as simple as (4.15) holds, and so the variance matrix cannot be inverted in closed form. Benard and Elteren (1953) present a numerical method that uses the rule of Cramer to produce the matrix product involved in $T$. Prentice (1979) performs these calculations in somewhat more generality, and the general unbalanced two-way test is generally called the Prentice test. Skillings and Mack (1981) addresses this question using matrix inversion explicitly, and performs the inversion numerically.

Example 29 Consider data on weight gain among chickens, given by Cox and Snell (1981, Example K). Weight gain after 16 weeks, protein level, protein source, and fish soluble level. The dependence of weight gain on protein level might be graphically depicted using box plots (Figure 5.2):

```r
temp1<-temp<-as.data.frame(scan("chicken.dat",what=
    list(source="", lev=0,fish=0,weight=0,w2=0)))
temp$weight<-temp1$weight;temp1$h<-0;temp1$w2<-NULL;
temp$w2<-NULL;temp1$w2<-NULL
chicken<-rbind(temp,temp1)
attach(chicken)
boxplot(split(weight,source),horizontal=TRUE,ylab="g.",
    main="Weight gain for Chickens", xlab="Protein Source")
```

Test for dependence of weight gain on protein level, blocking on source. Perform these calculations in R using

```r
library("muStat")#Need for Prentice test.
prentice.test(weight,source,blocks=lev)
```

to obtain the p-value 0.1336

FIGURE 5.2: Weight gain for Chickens
5.3 Multiple Comparisons and Scoring.

One might adjust for multiple comparisons in the presence of blocking in the same way as was done without blocking, as discussed in §4.5. That is, one might use the variances and covariances to show that under the null $(\bar{R}_i - \bar{R}_j)/\sqrt{K(LK + 1)/6}$ is approximately normal, and employ the method of Bonferroni, Fisher’s LSD, or Tukey’s HSD.

One might also use scoring methods, including normal and Savage scores, as before. Recall also that treating ranks as general scores, with tied observations given average ranks, makes tie handling automatic.

One might also use scores that are data values. In this case, the test statistic is similar to that of the normal-theory $F$ test, but the reference distribution is generated from all possible permutations of the data within blocks and among treatments.

Example 30 Refer again to Example 29. Test dependence of weight gain on protein source, blocking on level. There are eight observations at each protein level. Cycle through all $(\binom{8}{4})^3 = 34300$ ways to associate source labels within block, and compute the between sum of squares each time. The $p$-value is the number of rearrangements having this statistic meet or exceed the observed.

```r
library("MultNonParam")
# aov.P requires data sorted by block. Put block ends as the third argument.
chicken<-chicken[order(chicken$lev),]
aov.P(chicken$weight,as.numeric(as.factor(chicken$source)),
c(8,16,24))
The p-value is 0.182.
```

5.4 Tests for a Putative Ordering in Two-Way Layouts

Recall Friedman’s statistic $T \propto \sum_{k=1}^{K} [\bar{R}_k .. - 1/2(MK + 1)]^2$, where $\bar{R}_k ..$ are mean rankings for treatment $k$ when observations are ranked within blocks. The statistic $T$ treats deviation in any treatment in any direction similarly. If one wants to look specifically for alternatives suspected $a priori$ of being ordered according to the index $k$, in the balanced case use instead

$$T_L = \sum_{k=1}^{K} kR_{k ..}.$$  \hspace{1cm} (5.10)
This test was proposed by Page (1963) in the balanced case with one replicate per block-treatment pair, and is called Page’s test.

**Example 31** Refer again to the expense variability data of Example 28. Rank sums are 23, 36, 53, 57, 70, 70, 83, and the statistic is 23 \times 1 + 36 \times 2 + 53 \times 3 + 57 \times 4 + 70 \times 5 + 70 \times 6 + 83 \times 7 = 1883, with expected value 1568 and standard deviation under the null hypothesis 1829.333. The two-sided p-value is 2.90 \times 10^{-10}. Page’s test may be performed in R using `page.trend.test(expenses[,,-1],FALSE)`.

For more general replication patterns, the null expectation and variance for the statistic may be calculated, using (5.5), (5.6), and (5.7), and

\[
E_0[T_L] = \sum_{k=1}^{K} k E_0[R_{k..}] \quad (5.11)
\]

\[
\text{Var}_0[T_L] = \sum_{k=1}^{K} k^2 \text{Var}_0[R_{k..}] + \sum_{i=1}^{K} \sum_{k=1, k \neq i}^{K} ik \text{Cov}_0[R_{i..}, R_{k..}] \quad (5.12)
\]

In the balanced case, with \(M_{kl} = M \forall k, l\), moments simplify to

\[
E_0[T_L] = (K + 1)KLM(KM + 1)/4
\]

\[
\text{Var}_0[T_L] = LM^2(KM + 1) \left( \sum_{k=1}^{K} k^2(K - 1) - \sum_{i=1}^{K} \sum_{k=1, k \neq i}^{K} ik \right)
\]

\[
= LM^2(KM + 1) \left( K \sum_{k=1}^{K} k^2 - (\sum_{i=1}^{K} i)^2 \right)
\]

\[
= K^2(K + 1)LM^2(KM + 1)(K - 1)/12,
\]

using (3.18).

Page (1963) presents only the case with \(M_{kl} = 1\) all \(1\), and provides a table if the distribution of \(T_L\) for small values of \(K\) and \(L\). For larger values of \(K\) and \(L\) than appear in the table, Page (1963) suggests a normal approximation.

The scores in (5.10) are equally spaced. This is often a reasonable choice in practice. When the \(M_{kl}\) are not all the same, imbalance among numbers of ranks summed may change the interpretation of these scores, and a preferred statistic definition to replace (5.10) is

\[
T^*_L = \sum_{k=1}^{K} k \bar{R}_{k..} \quad (5.13)
\]

with moments given by

\[
E[\bar{R}_{k..}] = \sum_{i=1}^{L} M_{kl} \left( \sum_{j=1}^{K} M_{jl} + 1 \right)/2 / \sum_{i=1}^{L} M_{kl} \quad (5.14)
\]
Group Differences with Blocking

\[ \text{Var} \left[ \bar{R}_k \right] = \frac{\sum_{l=1}^{L} \{ M_{kl}(\sum_{j \neq k} M_{jl})(\sum_{j=1}^{K} M_{jl} + 1) \}}{12(\sum_{l=1}^{L} M_{kl})^2}. \] (5.15)

\[ \text{Cov} \left[ \bar{R}_k, \bar{R}_m \right] = -\sum_{l=1}^{L} [M_{kl}M_{ml}](\sum_{j=1}^{K} M_{jl} + 1)/(12 \sum_{l=1}^{L} M_{kl} \sum_{l=1}^{L} M_{ml}). \] (5.16)

**Example 32** Refer again to the chicken weight gain data of Example 29. In this case, test for an ordered effect of protein level on weight gain, blocking on protein source. In R, using `library(MultNonParam)`, perform the test using

```r
attach(chicken)
cat('n Page test with replicates n')
page.test.unbalanced(weight,source,lev)
```

In this balanced case, rank mean expectations are all 6.5, variances are 1.083, covariances are −0.542. The rank means by treatment level are 7.625, 7.250, 4.625, giving an overall statistic value of 36, compared with a null expectation of 39 and null variance of 3.25; the p-value is 0.096. Do not reject the null hypothesis.

### 5.5 Exercises

1. The data set
   
   http://ftp.uni-bayreuth.de/math/statlib/datasets/schizo
   
   contains data from an experiment using measurements used to detect schizophrenia, on non-schizophrenic patients. Various tests are given. Pick those data points with CS in the second column and compare the first and second gain ratios, in the third and fourth columns, by taking their difference. Test the null hypothesis of zero median difference using the Wilcoxon signed rank test, for the non-schizophrenic patients.

2. Calculate the asymptotic relative efficiency for the Wilcoxon signed rank statistic relative to the one-sample t-test (which you should approximate using the one-sample z-test). Do this for observations from
   a. the distribution uniform on the interval \([\theta - 1/2, \theta + 1/2]\),
   b. the logistic distribution, symmetric about \(\theta\), with variance \(\pi^2/3\) and density \(\exp(x - \theta)/(1 + \exp(x - \theta))^2\),

c. and the standard normal distribution, with variance 1 and expectation $\theta$.

3. The data set at

http://stat.rutgers.edu/home/kolassa/960-555/yarn.dat

represents strengths of yarn of two types from six bobbins. This file has three fields: strength, bobbin, and type. Apply the balanced variant of Friedman’s test to determine whether strength varies by type, controlling for bobbin.

4. The data at

http://stat.rutgers.edu/home/kolassa/960-555/friedman.dat

reflects expense variability by income level. Rows of this file correspond to expense categories, and columns represent income groups, in increasing order. Test the null hypothesis that expense variabilities do not depend on income group, treating categories as blocks, vs. the alternative that variabilities increases with income group.
6
Bivariate Methods

Suppose that independent random vectors \((X_i,Y_i)\) all have the same joint density \(f_{X,Y}(x,y)\). This chapter investigates whether \(X_i\) is independent of \(Y_i\) for each \(i\). That is, we test whether \(f_{X,Y}(x,y) = f_X(x)f_Y(y)\), without knowledge of \(f_{X,Y}\), or even of the null marginal densities.

This null hypothesis is most easily tested against alternative that, vaguely, large values of \(X\) are associated with large values of \(Y\) (or vice versa). Furthermore, if the null hypothesis is not true, we will measure the strength of the association between \(X_i\) and \(Y_i\).

6.1 Parametric Approach

Before developing nonparametric approaches to assessing relationships between variables, we review a standard parametric approach, that of the Pearson Correlation (Edgeworth, 1893):

\[
r_P = \frac{\sum_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y})}{\sqrt{\sum_{j=1}^{n}(X_j - \bar{X})^2 \sum_{j=1}^{n}(Y_j - \bar{Y})^2}},
\]

(6.1)
derived as what would now be termed the maximum likelihood estimator of the correlation for the bivariate normal distribution. This gives the slope of the least squares line fitting \(Y\) to \(X\), after scaling both variables by standard deviation. The Cauchy-Schwartz Theorem says that \(r_P\) is always in \([-1, 1]\). Perfect positive or negative linear association is reflected in a value for \(r_P\) of 1 or \(-1\) respectively.

6.1.1 Permutation Inference

Under the null hypothesis that \(X_j\) is independent of \(Y_j\), for all \(j\), then every permutation of the \(Y\) values among experimental units, with the \(X\) values held fixed, is equally likely. Under this permutation distribution, conditional on the collections \(\{X_1, \ldots, X_n\}\) and \(\{Y_1, \ldots, Y_n\}\), the variance of \(r_P\) may be calculated directly. Note that \(r_P\) depends only on the differences between the observations and their means, and so, without loss of generality, assume that...
\( \bar{X} = \bar{Y} = 0 \). Let \((Z_1, \ldots, Z_n)\) be a random permutation of \(\{Y_1, \ldots, Y_n\}\). Then

\[
\text{Cov} [Z_1, Z_2] = \sum_{i \neq j} \frac{1}{n(n-1)} Y_i Y_j \\
= \sum_{i,j} \frac{1}{n(n-1)} Y_i Y_j - \frac{1}{n(n-1)} Y_i^2 = -\frac{1}{n-1} \sigma^2_Y,
\]

for \( \sigma^2_Y = \sum_i \frac{1}{n} Y_i^2 \). So

\[
\text{Var} \left[ \sum_j X_j Z_j \right] = \sum_j X_j^2 \text{Var} [Z_j] + \sum_{j \neq i} X_i X_j \text{Cov} [Z_j, Z_i] \\
= \sum_j X_j^2 \sigma^2_Y - \frac{[\sum_j X_j^2 - \sum_i X_i^2] \sigma^2_Y}{n-1} \\
= \sum_j X_j^2 \sigma^2_Y + \sum_i X_i^2 \sigma^2_Y/(n-1) \\
= n^2 \sigma^2_X \sigma^2_Y/(n-1),
\]

where \( \sigma^2_X = \sum_i \frac{1}{n} X_i^2 \). So

\[
\text{Var} [r_P] = 1/(n-1), \tag{6.2}
\]

under the permutation distribution (Hotelling and Pabst, 1936). This value was determined by Student (1908), after fitting a parametric model to empirical data, rounding parameter estimates to values more consistent with intuition, and calculating the moments of this empirical distribution. Higher-order moments were determined by David et al. (1951) using a method of proof similar to that presented above for the variance.

Using this result, one might test the null hypothesis of independence vs. the two-sided alternative at level \( \alpha \) using \( r_P \), and rejecting the null hypothesis if

\[
|r_P| > z_{\alpha/2}/\sqrt{n-1}. \tag{6.3}
\]

### 6.2 Nonparametric Correlation

Pearson’s correlation \( p_r \) is designed to measure linear association between two variables, and fails to adequately reflect non-linear association. Furthermore, a family of distributional results, not recounted in this volume, depend on bivariate normality of the data summarized. Various nonparametric alternatives to the Pearson correlation have been developed to avoid these drawbacks.
6.2.1 Rank Correlation

Instead of calculating the correlation for the original variables, calculate the correlation of the ranks of the variables Spearman (1904). Under the null each X rank should be equally likely to be associated with each Y rank. Under the alternative, extreme ranks should be associated with each other. Let $R_j$ be the rank of the Y value associated with $X(j)$. Define the Spearman Rank correlation as

$$r_S = \frac{\sum_{j=1}^{n} (j - (n + 1)/2)(R_j - (n + 1)/2)}{\sqrt{\left(\sum_{j=1}^{n} (j - (n + 1)/2)^2\right) \left(\sum_{j=1}^{n} (R_j - (n + 1)/2)^2\right)}};$$

that is, $r_S$ is Pearson correlation on ranks.

Sums of squares in the denominator have the same value for every data set, and numerator can be simplified. Note that

$$\sum_{j=1}^{n} (j - (n + 1)/2)^2 = \sum_{j=1}^{n} j^2 - n(n+1)^2/4 = n(n+1)(2n+1)/6 - n(n+1)^2/4 = n(n^2 - 1)/12.$$  
Similarly, $\sum_{j=1}^{n} (R_j - (n + 1)/2)^2 = n(n^2 - 1)/12$. Furthermore, $\sum_{j=1}^{n} (j - (n + 1)/2)(n + 1)/2 = 0$, and

$$r_S = \frac{12}{n(n^2 - 1)} \sum_{j=1}^{n} (j - (n + 1)/2)R_j = \frac{12}{n(n^2 - 1)} \left(\sum_{j=1}^{n} jR_j - n(n+1)^2/4\right).$$

(6.4)

**Example 33** Consider again the twin brain data of Example 24, plotted in Figure 6.1. As before, the data set both has 10 records, reflecting the results from 10 pairs of twins. Ranks for twin brains are given in Table 6.1. The sum in the second factor of (6.4)

$$1 \times 1 + 4 \times 2 + 9 \times 9 + 5 \times 5 + 3 \times 4 + 6 \times 7 + 7 \times 3 + 10 \times 10 + 2 \times 6 + 8 \times 8 = 366,$$

and so the entire second factor is $366 - 10 \times 11^2/4 = 63.35$, and the Spearman correlation is $(12)(10 \times 99) \times 63.35 = 0.770$. Observed correlations may be calculated in R using

```r
cat(‘\n Permutation test for twin brain data \n’)
attach(both)
obsd<-c(cor(v1,v2),cor(v1,v2,method="spearman"))
```
to obtain Pearson and Spearman correlations 0.914 and 0.770 respectively. The Pearson correlation is given in (6.1), and the Spearman correlation
Bivariate Methods

TABLE 6.1: Twin Brain Volume Ranks

<table>
<thead>
<tr>
<th>Pair</th>
<th>1 2 3 4 5 5 7 8 9 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank of First</td>
<td>1 4 9 5 3 6 7 10 2 8</td>
</tr>
<tr>
<td>Rank of Second</td>
<td>1 2 9 5 4 7 3 10 6 8</td>
</tr>
</tbody>
</table>

was given in (6.4). Testing the null hypothesis of no association may be performed using a permutation test with either of these measures.

```r
tag <- array(NA, c(2, 20001))
dimnames(tag)[[1]] <- c("Pearson", "Spearman")
for(j in seq(dim(tag)[2])){
  newv1 <- sample(v1)
  tag[, j] <- c(cor(newv1, v2), cor(newv1, v2, method="spearman"))
}
cat("Monte Carlo One-Sided p value\n")
apply(apply(tag, 2, ">=", obsd), 1, "mean")
```
to obtain \( p \)-values \( 1.5 \times 10^{-4} \) and \( 6.1 \times 10^{-3} \). The asymptotic critical value, from (6.3) and using the normal approximation as motivated by Erdős and Réyni (1959), is given by

```r
cat("Asymptotic Critical Value\n")
-qnorm(0.025)/sqrt(length(v1)-1)
```
which gives 0.6533. Permutation tests based on either correlation method reject the null hypothesis. \( P \)-values may be approximated by dividing the observed correlations by \( \sqrt{n-1} \), but this approximation has poor relative behavior for small test levels. Function Kendall in package Kendall in R uses a modification of the method of Best and Roberts (1975), which corrects the normal approximation using an Edgeworth series. This construction is depicted in Figure 6.1.

```r
cat("P-values using t approximation\n")
c(cor.test(v1, v2)$p.value, cor.test(v1, v2, method="spearman")$p.value)
```
giving \( p \)-value approximations \( 2.1 \times 10^{-4} \) and \( 1.37 \times 10^{-2} \) respectively.

Pearson (1907) criticizes the Spearman correlation, in part on the grounds that, as it is designed to reflect association even when the association is non-linear, it also looses the interpretation as a regression parameter, even when the underlying association is linear.
6.2.1.1 Moments of the Spearman correlation

Note that

$$\sum_{j=1}^{n} j R_j = \sum_{j=1}^{n} j \left( 1 + \sum_{i \neq j} I(Y_j > Y_i, X_j > X_i) \right)$$

$$= n(n+1)/2 + \sum_{j=1}^{n} j \sum_{i \neq j} I(Y_j > Y_i, X_j > X_i).$$

Some algebra results in

$$r_S = 1 - \left( \frac{6}{n(n^2 - 1)} \right) S$$

for

$$S = \sum_{j=1}^{n} (j - R_j)^2. \quad (6.5)$$

Note that $r_S = 1$ if variables share exact same ordering, and $r_S = -1$ if ordering is exactly opposite. The Cauchy-Schwartz theorem says these are the extremes.

Under the null hypothesis of independence between $X_j$ and $Y_j$, for all $j$,
E[S] = 0, and the variance is given by (6.2). Under the alternative hypothesis,

\[ E[S] = \frac{12}{n(n^2-1)} \frac{n(n+1)/2}{n(n+1)/2} + \frac{\sum_{j=1}^{n} \sum_{i < j}^{n} p_1 - n(n+1)/2}{n(n-1)/4} \]

\[ = \frac{12}{n(n^2-1)} \left( (n-1)n(n+1)p_1/2 - n(n^2-1)/4 \right) = 3(2p_1 - 1). \]

for

\[ p_1 = P[Y_j > Y_i, X_j > X_i]. \]  

(6.6)

Hoeffding (1948) proves asymptotic normality of \( r_S \), including under alternative distributions.

6.2.2 Kendall’s Tau

Kendall (1938) constructs a new measure based on counts of concordant and discordant pairs. Consider the population quantity \( \tau = 2p_1 - 1 \), for \( p_1 \) of (6.6), called Kendall’s Tau. This quantity reflects the probability of a concordant pair minus the probability of a discordant pair. Denote the number of concordant pairs by

\[ U = \sum_{i < j} Z_{ij}, \]

for \( Z_{ij} = I((X_j - X_i)(Y_j - Y_i) > 0) \). Then \( U^+ = n(n-1)/2 - U \) is number of discordant pairs; equals number of rearrangements necessary to make all pairs concordant. Estimate \( \tau \) by the excess of concordant over discordant pairs, divided by maximum:

\[ r^\tau = \frac{U - (n(n-1)/2 - U)}{n(n-1)/2} = \frac{4U}{n(n-1)} - 1. \]  

(6.7)

Then \( E[U] = n(n-1)p_1/2 \), for \( p_1 \) of (6.6), and \( E[r^\tau] = 2p_1 - 1 \). The null value of \( p_1 \) is half, recovering

\[ E_0[U] = n(n-1)/4, \quad E_0 [r^\tau] = 0. \]  

(6.8)

As with the Spearman correlation \( r_S \), Kendall’s correlation \( r^\tau \) may be used to test the null hypothesis of independence, using asymptotic normality, but one requires the variance. Note that

\[ \text{Var} [U] = \sum_{i < j} \text{Var} [Z_{ij}] + \sum_{i < j}^* \text{Cov} [Z_{ij}, Z_{kl}]. \]

Here \( \sum^* \) the sum over three distinct indices \( i < j, k < l \). This sum consists of \( n^2(n-1)^2/4 - n(n-1)/2 - n(n-1)(n-2)(n-3)/4 = n(n-1)(n-2) \)
Nonparametric Correlation

Hence

\[ \text{Var}[U] = \sum_{i<j} \text{Var}[Z_{ij}] + \sum^* \text{Cov}[Z_{12}, Z_{13}] \]

\[ = \sum_{i<j} (p_1 - p_1^2) + \sum^* (p_3 - p_1^2) \]

\[ = n(n - 1)(p_1 - p_1^2)/2 + n(n - 1)(n - 2)(p_3 - p_1^2), \]

where

\[ p_3 = P[(X_1 - X_2)(Y_2 - Y_1) \geq 0, (X_1 - X_3)(Y_3 - Y_1) \geq 0]. \]

The null value of \( p_3 \) is \( \frac{5}{18} \), as can be seen by examining all 36 pairs of permutations of \( \{1, 2, 3\} \). Hence

\[ \text{Var}_0[U] = n(n - 1)/8 + n(n - 1)(n - 2)/36 = n(n - 1)(5 + 2n)/72, \quad (6.9) \]

and

\[ \text{Var}_0[r_\tau] = 2(2n + 5)/(9n(n - 1)). \quad (6.10) \]

The result of Hoeffding (1948) also proves asymptotic normality of \( r_\tau \), including under alternative distributions. El Maache and Lepage (2003) discuss multivariate normality of both \( r_\tau \) and \( r_S \) from collections of variables.

**Example 34** Consider again the twin brain data of Example 24, with ranks in Table 6.1. Discordant pairs are 2 – 5, 2 – 9, 4 – 7, 4 – 9, 5 – 7, 5 – 9, 6 – 7, and 7 – 9. Hence 8 of 45 pairs are discordant, and the remainder, 37, are concordant. Hence \( r_\tau = 4 \times 37/90 - 1 = 0.644 \), from (6.7). This may be calculated in R using

\[ \text{cor(both$v1, both$v2, method="kendall")} \]

A test of the null hypothesis of no correlation may be performed by calculating \( z = r_\tau/\sqrt{2(2n + 5)/(9n(n - 1))} = 0.644/\sqrt{2 \times 25/810} = 0.644/0.248 = 2.60; \) reject the null hypothesis of no association for a two-sided test of level 0.05.

Figure 6.2 shows two artificial data sets with functional relationships between two variables. In the first relationship, the two variables are identical, while in the second, the second variables is related to the arctangent of the first. Both variables relationships show perfect association. The first relationship represents perfect linear association, while the second reflects perfect nonlinear association. Hence the Spearman and Kendall association measures are 1 for both relationships; the Pearson correlation is 1 for the first relationship, and 0.937 for the second relationship.
6.3 Bivariate semi-parametric methods

Spearman’s Rank Correlation $r_S$ is the correlation of ranks of $X$’s and $Y$’s. This definition was motivated by the expectation that under the alternative hypothesis, large $X$’s would be associated with large $Y$’s, or with small $Y$’s.

Assume a linear model $Y_j = \beta_1 + \beta_2 X_i + e_i$, where $e_i$ has some distribution with median 0 (making $\beta_1$ identifiable). The correlation measures introduced above may be used to estimate $\beta_2$. Furthermore, one might obtain a confidence interval by inverting test of the null hypothesis that the residuals are independent of the explanatory variables. This approach could be addressed this using the Pearson, Spearman, or Kendall correlation. Using the Pearson approach gives the traditional least-squares estimate for $\beta_2$, and approximating the distribution of the errors as Gaussian leads to standard normal theory inference.

One could also perform inference using the permutation distribution for $\sum_{j=1}^n(X_j - \bar{X})R_j$ for $R_j$ the rank of the residuals, possibly after transforming the ranks.

6.3.1 Inversion of the test of zero correlation

One might invert the test using correlation between $X_i$ and $Y_i - \beta_2 X_i$, as a function of $\beta_2$, using any of the three preceding correlation meth-
ods, generically called $r$. That is, let $T(\beta_2) = r(X, Y - \beta_2 X)$, and note that this function is monotonic in $\beta_2$. First, calculate $t(L)$ and $t(U)$, such that $P[t(L) < T(\theta, \text{data}) < t(U)] < 1 - \alpha$, exactly as in (1.18). Then solve $r(\beta_2) = t(L)$, $r(\beta_2) = t(U)$.

In all of these approaches, inference on $\beta_1$ is impossible, because it washes out of ranks.

6.3.1.1 Pearson Correlation

The fully-parametric analysis, assuming bivariate normality for the data and $r_P$, reduces to the standard inference from ordinary least squares regression. Below we consider inference under the permutation distribution, to avoid distributional assumptions on the data apart from independence.

Both determining critical values for $r_P$, and inverting the test statistic as a function of $\beta_2$ to give confidence bounds, are tricky for Pearson correlation, under the permutation distribution. Exact (non-Monte Carlo) probability calculations for the permutation distribution of the Pearson correlation are as difficult as enumerating all permutations. These calculations are often simplified by rounding data values to a lattice. Alternatively, since the null mean and variance are known, and under the null hypothesis the distributions of these measures are approximately normal, quantiles could be approximated using normal distribution.

Also, $r_P$ has steps of irregular size, and so translating $q_{\alpha/2}$ and $q_{1-\alpha/2}$ into values of $\beta_2$ is difficult.

6.3.1.2 Kendall’s $\tau$

Calculations are easier for Kendall’s $\tau$. Best and Gipps (1974), using the algorithm of Kaarsemaker and van Wijngaarden (1953), present software for calculating all of the probability atoms, and hence quantiles, exactly for samples as large as 8, and with an Edgeworth series for larger samples. They note that, for $c(n, u)$ the count of permutation of $Y_1, \ldots, Y_n$ with $u$ concordant pairs, then

$$c(n, u) = \sum_{s=u-n+1}^{u} c(n-1, s), \text{ for } n > 1, \ n(n-1)/2 \geq u \geq 0, \ (6.11)$$

$$c(n, u) = 0 \text{ for } u < 0, \ c(0, 0) = 1,$$

$$c(n, u) = 0 \text{ for } u > n(n-1)/2. \ (6.12)$$

Relation (6.11) is motivated by considering the possible effects of adding a final observation for a data set with $n - 1$ observations. Furthermore, $P[U = u] = c(n, u)/n!$.

A recursion for counts of permutations satisfying $U \leq u$ are identical, except that

$$c(n, u) = 1 \text{ for } u > n(n-1)/2$$
replaces (6.12).

Quantile $1 - \alpha$ of $U$ can be calculated by inverting $P[U \leq u]$, for samples sizes corresponding to feasible exact calculations, or via

$$E_0[U] + z_\alpha \sqrt{\text{Var}_0[U]}, \quad (6.13)$$

for moments given by (6.8) and (6.9). Jumps in $r_{\tau}$, as a function of $\beta_2$, occur at pairwise slopes $W_{ij} = (Y_j - Y_i)/(X_j - X_i)$, since at these points the pairs $(X_i, Y_i - \beta_2 X_i)$ and $(X_j, Y_j - \beta_2 X_j)$ change from being concordant to discordant (Theil, 1950), and so an estimate $\hat{\beta}_2$ of $\beta_2$ is given by the median of these slopes, and the confidence interval endpoints are given as in §3.10.1. This last operation is more easily performed using the number of concordant pairs $U$, as in (6.7). Hence if $t^o$ satisfies $P_0[U \leq t^o] \geq \alpha/2$ and $P_0[U \geq t^o] \geq 1 - \alpha/2$, and if $\tilde{W}_1, \ldots, \tilde{W}_{n(n-1)/2}$ are the ordered values of the pairwise slopes $W_{ij}$, then $(\tilde{W}_{t^o}, \tilde{W}_{n(n-1)/2+1-t^o})$ is a $1 - \alpha$ confidence interval for $\beta_2$. The intercept $\beta_1$ may be estimated as that value $\hat{\beta}_1$ to give the residuals $Y_j - \hat{\beta}_1 - \hat{\beta}_2 X_j$ have zero median; that is,

$$\hat{\beta}_1 = \text{samplmed} [Y_1 - \hat{\beta}_2 X_1, \ldots, Y_n - \hat{\beta}_2 X_n]. \quad (6.14)$$

Generally, $\hat{\beta}_1 \neq \text{samplmed}[Y_1, \ldots, Y_n] - \hat{\beta}_2 \text{samplmed}[X_1, \ldots, X_n]$. Sen (1968) investigates this procedure further.

This approach to estimation of $\beta_2$ only holds in contexts with no ties among the explanatory variable values.

**Example 35** Sen applies this procedure to the data set

1 2 3 4 10 12 18
9 15 19 20 45 55 78

The following commands calculate a confidence interval for the slope parameter:

```r
tt<-c(1,2,3,4,10,12,18); xx<-c(9,15,19,20,45,55,78)
out<-rep(NA,length(tt)*(length(tt)-1)/2)
count<-0
for(ii in seq(length(tt)-1)) for(jj in (ii+1):length(tt)){
  count<-count+1
  out[count]<-(xx[jj]-xx[ii])/(tt[jj]-tt[ii])
}
There are $7 \times 6/2 = 21$ pairwise slopes $\tilde{W}_j$:
1.00 2.50 3.67 3.71 3.75 3.83 3.94 4.00 4.00
4.00 4.00 4.06 4.12 4.14 4.17 4.18 4.38 5.00 5.00 6.00

The estimate of $\beta_2$ in the regression model $\text{med}[X_j] = \beta_1 + \beta_2 t_j$ is the
median of these piecewise-slopes, 4. Confidence intervals may be exhibited using quantiles of the number of concordant pairs, given in package MultNonParam.

```
library(MultNonParam); qconcordant(0.025,7)
giving 4 as the quantile. Hence the 0.95 confidence interval is (\(\hat{W}_4, \hat{W}_{18}\)) = (3.71, 4.38). These calculations may be performed using `theil(xx,tt)`. It may also be performed using `theilsen(xx~tt)` from the package deming, which also gives a confidence interval for the intercept term. The intercept is estimated by (6.14). Plotting may be done via

```
plot(tt,xx);abline(theilsen(xx~tt))
```
Results are in Figure 6.3.

FIGURE 6.3: Thiel-Sen Estimator for Artificial Example

6.3.1.3 Spearman Correlation

Calculation for the Spearman correlation \(r_S\) is more complicated. van de Wiel and Di Bucchianico (2001) provide tools for the exact calculation of the null sampling distribution of the Spearman correlation \(r_S\); this algorithm is implemented in the R package pspearman, which gives probabilities for the random variable \(S\) of (6.5). Closed-form inversion of the function \(T(\beta_2) = r_S(X, Y - \beta_2 X)\) is more difficult.

**Example 36** Consider data on changes in systolic and diastolic blood pressure after treatment by the drug captopril (Cox and Snell, 1981, Ex-
ample E). We calculate the Theil estimate and confidence interval for the linear dependence of change in diastolic blood pressure, $Y_i$, on the change in systolic blood pressure, $X_i$. Pearson, Spearman, and Kendall correlations between change in systolic blood pressure, $Y_i$, and the change in diastolic blood pressure $X_i$, are 0.105, 0.144, and 0.067 respectively. The three correlations between $X_i$ and $Y_i - \beta_2 X_i$, as a function of the slope, are plotted in Figure 6.4. The relationship for the Pearson correlation is smooth; the relationships for the Spearman and Kendall correlations are step functions. Horizontal lines corresponding to asymptotic normal critical values for the test of zero correlation are plotted; the permutation variance (6.2) is used for the Pearson and Spearman correlation. The two lines are slightly offset in the picture to display both lines. The Kendall correlation uses variance (6.10). Vertical lines pass through the point at which the horizontal lines cross the correlation curve, and their intersections with the horizontal axis determine the end points of the regression confidence interval. Figure 6.5 displays the resulting regression lines.

FIGURE 6.4: Construction of Regression Estimates for Blood Pressure Data

Statistic Value

Regression Parameter

Statistic Quantiles are approximated using Normal
6.4 Exercises

1. The data set

http://ftp.uni-bayreuth.de/math/statlib/datasets/federalistpapers.txt

represents the result from an analysis of a series of documents. The first column gives document number, the second gives the name of a text file, the third gives a group to which the text is assigned, the fourth represents a measure of the use of first person in the text, the fifth presents a measure of inner thinking, the sixth presents a measure of positivity, and the seventh presents a measure of negativity. There are other columns that you can ignore. Plot use of positivity vs. negativity. Calculate the three correlation measures, and relate their relative values to the shape of the curve. Test at \( \alpha = .05 \) the null hypothesis of zero population Spearman correlation for these two variables, vs. the general alternative. (The version on line, above, has odd line breaks. A fixed version can be found at stat.rutgers.edu/home/kolassa/960-555/federalistpapers.txt).

2. The data set at
reflects brain volumes of twins. Using the method of Theil, estimate the slope parameter in the regression of second brain volume on the first brain volume, and give a confidence interval.
Suppose that one observes \( n \) subjects, indexed by \( i \in \{1, \ldots, n\} \), and, for subject \( i \), observes responses \( X_{ij} \), indexed by \( j \in \{1, \ldots, J\} \). Potentially, covariates are also observed for these subject.

This chapter explores explaining the multivariate distribution of \( X_{ij} \) in terms of these covariates. Most simply, these covariates often indicate group membership.

### 7.1 Multivariate Estimation

Often one wants to infer about a the vector of population central values for each of the \( j \) responses on the various subjects. In this section, we assume that the vectors are independent and identically distributed.

Standard parametric analyses presuppose distributions of data well-enough behaved that location can be well-estimated using a sample mean. Denote the mean by the vector \( \hat{\mu}(X) \), where component \( j \) of this vector is given by \( \hat{\mu}_j(X) = \frac{\sum_{i=1}^{n} X_{ij}}{n} \). Then \( \hat{\mu}(X) \) is the method of moments estimator for \( \mu = E[X] \). Assumptions guaranteeing that \( \hat{\mu}(X) \) has an asymptotically normal distribution generally include the existence of some moment of the distribution greater than the second moment.

In the absence of such parametric assumptions, one might instead measure location using the median. Central limit theorems for the sample median generally require only that the density of the raw observations be positive in a neighborhood of the median.

Define the multivariate population median \( \nu \) to be the vector of univariate medians; that is, \( \nu_j \) satisfies

\[
P[X_{ij} \leq \nu_j] \geq \frac{1}{2}\text{ and } P[X_{ij} \geq \nu_j] \geq \frac{1}{2}.
\]

Such medians always exists. A sufficient condition for the to be unique is that \( X_{ij} \) has a positive density around one root of (7.1); another sufficient condition for uniqueness is that \( X_{ij} \) have a positive point mass on one candidate median.

An estimator \( \hat{\nu}(X) \) of \( \nu \) may constructed by applying (7.1) to the empirical distribution of \( X_{ij} \), formed by putting point mass on each of the \( n \) values. In this case, with \( n \) odd, the median is the middle value, and, with \( n \) even,
even, (7.1) fails to uniquely define the estimator. In this case, the estimator is conventionally defined to be the average of the middle two values.

Alternatively, one might define \( \nu \) so to minimize the sum of distances from the median:

\[
\tilde{\nu}(X) = \arg\min_{\nu} \sum_{i=1}^{n} \sum_{j=1}^{J} |X_{ij} - \nu_j| \tag{7.2}
\]

that is, the estimate minimizes the sum of distances from data vectors to the parameter vector, with distance measured by the sum of absolute values of component-wise differences. Because one can interchange the order of summation in (7.2), the minimizer in (7.2) is the vector of component-wise minimizers. Furthermore, the minimizer for each component is the traditional univariate median as above.

A summary of multivariate median concepts is given by Small (1990).

### 7.1.1 Invariance Properties

In the univariate case (that is, \( J = 1 \)), both the mean and the median are equivariant with respect to affine transformations of the raw data. That is, if \( \tilde{X}_{i1} = a + bX_{i1} \), then \( \tilde{\mu}(\tilde{X})_1 = a + b\tilde{\mu}(X)_1 \), with the same relation holding with \( \tilde{\nu} \) in place of \( \tilde{\mu} \). Equivariance to affine transformations in the multivariate case holds for the mean: for a vector \( a \) and a matrix \( B \) with \( J \) columns, \( \tilde{\mu}(a + BX) = a + B\tilde{\mu}(X) \). A similar equality fails to hold for \( \tilde{\mu} \), unless \( B \) is diagonal; hence \( \tilde{\mu} \) is not equivariant under affine transformations.

### 7.2 One-Sample Hypothesis Testing

Consider a null hypothesis stating that the marginal median vector \( \nu \) takes on a value specified in advance; without loss of generality, take this to be zero.

#### 7.2.1 The Standard Parametric Approach

First consider the case when \( \tilde{X} \) follows a central limit theorem, and in which one knows the variance matrix \( \Sigma = \text{Var}[(X_{i1}, \ldots, X_{iJ})] \), and \( \Sigma \) is nonsingular. Then one can use as a test statistic \( T^2 = \tilde{X}^\top \Sigma^{-1} \tilde{X} \), and can approximate its distribution under the null hypothesis by \( \chi^2_J \). If \( \Sigma \) is unknown, and if one can estimate it as \( \hat{\Sigma} \) using the usual sum of squares, and

\[
T^2 = \tilde{X}^\top \hat{\Sigma}^{-1} \tilde{X} \tag{7.3}
\]

has an \( F \) distribution, with \( J \) numerator degrees of freedom (Hotelling, 1931). If the distribution of \( (X_{i1}, \ldots, X_{iJ}) \) has a density, then \( P[\hat{\Sigma} \text{ singular}] = 0 \).
If $\Sigma$ unknown, and is best estimated by a nonsingular $\tilde{\Sigma}$, which is other than the sum of squares estimator, then generally $\bar{X}^\top \tilde{\Sigma}^{-1} \bar{X} \sim \chi^2_J$, approximately. These techniques require approximate multivariate normality of $\bar{X}$, which is stronger than approximate marginal normality.

7.2.2 Nonparametric Asymptotic Approaches

The statistic (7.3) represents the combination of separate location test statistics for the various components of the random vectors, and its distribution depends on multivariate normality of the underlying observations. Note that marginal normality for each variable does not imply multivariate normality; a simulated example is given in Figure 7.1.

**FIGURE 7.1:** Univariate Normal Data that are Not Bivariate Normal
A nonparametric hypothesis test can be constructed by assembling component-wise nonparametric statistics into a vector $T$, analogous to $\bar{X}$, and centered so that $E_0[T] = 0$. One might combine sign test statistics, or signed rank statistics if one assumes symmetry, often in the context of paired data. That is,

$$T(X_{ij}, \ldots, X_{nj}) = \sum_{i=1}^{n} \tilde{s}(X_{ij} > 0), \quad (7.4)$$

for $\tilde{s}(u) = \begin{cases} 1 & \text{for } u > 0 \\ -1 & \text{for } u < 0 \end{cases}$. Or, define $R_{ij}(X)$ to be the marginal rank of $|X_{ij}|$ among $\{|X_{1j}|, \ldots, |X_{nj}|\}$, and set

$$T_j(X_{1j}, \ldots, X_{nj}) = \sum_{i=1}^{n} R_{ij}(X) \tilde{s}(X_{ij} > 0). \quad (7.5)$$

The multivariate version of the sign test is

$$T(X) = (T_1(X_{11}, \ldots, X_{n1}), \ldots, T_J(X_{1J}, \ldots, X_{nJ}))^\top.$$  

Then combine components of $T$ from (7.4) to give the multivariate sign test statistic, or from (7.5) to give the multivariate sign rank test, using

$$W = T^\top \text{Var}_0[T] T. \quad (7.6)$$

As in §2.3, in the case that the null location value is 0, the null distribution for the multivariate sign test statistic is generated by assigning equal probabilities to all $2^n$ modifications of the data set by multiplying the rows $(X_{1i}, \ldots, X_{ji})$ by +1 or −1. That is, the null hypothesis distribution of $T(X)$ is generated by placing probability $2^{-n}$ on all of the $2^n$ elements of

$$\mathcal{X} = \{ \bar{X} \text{ an } n \times J \text{ matrix} | (\bar{X}_{1i}, \ldots, \bar{X}_{ji}) = \pm (X_{1i}, \ldots, X_{ji}) \}. \quad (7.7)$$

In previous applications of these central limit results, in the case of univariate sign and signed rank one-sample tests, in the case of two-sample Mann-Whitney-Wilcoxon tests, and in the case of of Kruskal-Wallis testing, all dependence of the permutation distribution on the original data was removed through ranking. This is not the case for $T$, since this distribution involves correlations between ranks of the various response vectors. These correlations are not specified by the null hypothesis. The separate tests are generally dependent, and dependence structure depends on distribution of raw observations. The asymptotic distribution of (7.6) relies on this dependence via the correlations between components of $T$. The correlations must be estimated.

Furthermore, the distribution of $W$ of (7.6) under the null hypothesis is dependent on the coordinate system for the variables, but, intuitively, the analysis should not. For example, suppose that $(X_{1i}, X_{2i})$ is distributed approximately multivariate normally, with expectation $\mu$, and variance $\Sigma$, with
One-Sample Hypothesis Testing

Consider the null hypothesis $H_0 : \mu = 0$. Then the canonical test is $(X_{1i}, X_{2i})\Sigma^{-1}(X_{1i}, X_{2i})^\top$, and it is unchanged if we base test on $(U_i, V_i)$ for $U_i = X_{1i} + X_{2i}$ and $V_i = X_{1i} - X_{2i}$. Hence the parametric analysis is independent of the coordinate system.

The first of this difficulty is readily addressed. For marginal sign tests, $T_j$ is given by (7.4). Then, marginally, under $H_0$, $T_j/\sqrt{n} \approx \mathcal{G}(0, 1)$. One might estimate $\text{Cov}[\tilde{s}(X_{ij}), \tilde{s}(X_{ij'})] = E[\tilde{s}(X_{ij})\tilde{s}(X_{ij'})]$ by

$$\hat{\sigma}_{jj'} = \sum_i \tilde{s}(X_{ij})\tilde{s}(X_{ij'})/n.$$  \hspace{1cm} (7.8)

The solution for the Wilcoxon signed rank test is

$$\hat{\sigma}_{jj'} = \sum_i R_{ij}R_{ij'}\tilde{s}(X_{ij})\tilde{s}(X_{ij'})/n.$$  \hspace{1cm} (7.9)

Combine the components of $T(X)$ to construct the statistic

$$W = W(X) = T^\top\Sigma^{-1}T,$$  \hspace{1cm} (7.10)

for an estimate $\hat{\Sigma}$ of $\Sigma = \text{Var}[T]$ as in (7.8), or similarly for the signed rank statistic.

The multivariate central limit theorem of Hájek (1960), and the quality of approximation to $\Upsilon$, justifies approximating the null distribution of $W$ by $\chi_j^2$ distribution. The test statistic that rejects the null hypothesis of zero component-wise medians when

$$W(X) > \chi_{j,\alpha}^2$$

for the $1 - \alpha$ quantile of the $\chi_j^2$ distribution.

**Example 37** Consider the data of Example 36. We test the null hypothesis that the joint distribution of systolic and diastolic blood pressure changes is symmetric about $(0, 0)$, using Hotelling’s $T^2$ and the two asymptotic tests that substitutes signs and signed ranks for data. This test is performed in R using

```r
# For Hotelling and multivariate rank tests resp:
library("Hotelling"); library("ICSNP")
cat("\n One-sample Hotelling Test\n")
HotellingsT2(bp[,c("spd","dpd")])
cat("\n Multivariate Sign Test\n")
rank.ctest(bp[,c("spd","dpd")],scores="sign")
cat("\n Multivariate Signed Rank Test\n")
rank.ctest(bp[,c("spd","dpd")])
```
Multivariate Analysis

P-values for Hotelling’s $T^2$, the marginal sign rank test, and marginal sign test, are $9.839 \times 10^{-6}$, $2.973 \times 10^{-3}$, and $5.531 \times 10^{-4}$.

Tables 7.1 and 7.2 contain attained levels and powers for one-sample multivariate tests with two manifest variables of nominal level 0.05, from various distributions. Tests compared are Hotelling’s $T^2$ tests, and test (7.10) applied to the sign and signed rank tests. Tests have close to their nominal levels, except for Hotelling’s test with the Cauchy distribution; furthermore, the agreement is closer for sample size 40 than for sample size 20. Furthermore, the sign test power is close to that of Hotelling’s test for Gaussian variables, and the signed rank test has attenuated power. Both nonparametric tests have good power for the Cauchy distribution, although Hotelling’s test performs poorly, and the perform better than Hotelling’s test for Laplace variables.

Some rare data sets simulated to create Tables 7.1 and 7.2 include some for which $\Upsilon$ is estimated as singular. Care must be taken to avoid difficulties; in such cases, $p$-values are set to 1.

7.2.3 More General Permutation Solutions

One might address this problem using permutation testing. First, select an existing parametric test statistic $U(X)$, perhaps a Hotelling statistic, or a rank-based statistic. Under the permutation null distribution, the sampling distribution puts equal weight $2^{-n}$ to all $2^n$ values of the statistic evaluated at each element of (7.7); these $2^n$ values need not all be unique. For $n$ large enough to make exhaustive evaluation prohibitive, a random subset of elements of (7.7) may be selected. The $p$-value is reported as the proportion of data sets with permuted signs having a the test statistic value as large as, or larger than, that observed. In this way, the analysis of the previous subset

<table>
<thead>
<tr>
<th>Test</th>
<th>Sample Size 20</th>
<th>Sample Size 40</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>Cauchy</td>
</tr>
<tr>
<td>$T^2$</td>
<td>0.04845</td>
<td>0.01550</td>
</tr>
<tr>
<td>Sign test</td>
<td>0.04215</td>
<td>0.04130</td>
</tr>
<tr>
<td>Sign rank test</td>
<td>0.04115</td>
<td>0.03895</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>Sample Size 20</th>
<th>Sample Size 40</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>Cauchy</td>
</tr>
<tr>
<td>$T^2$</td>
<td>0.74375</td>
<td>0.08090</td>
</tr>
<tr>
<td>Sign test</td>
<td>0.70330</td>
<td>0.26290</td>
</tr>
<tr>
<td>Sign rank test</td>
<td>0.52615</td>
<td>0.32605</td>
</tr>
</tbody>
</table>
for the sign test, and by extension the signed-rank test, can be extended to
general rank tests, including tests with data as scores.

7.3 Confidence Regions for a Vector Shift Parameter

Proceed analogously to the one-dimensional confidence interval construction
as in §2.4. Introduce a shift parameter to move the data to a conventional null
hypothesis. In this one-sample case, one may apply the one sample test for the
null hypothesis that marginal medians are 0 to the shifted data
\( X - 1_n \otimes \mu \),
where 1\(_n\) is vector of ones of length \( n \), and \( \otimes \) is outer product, so 1\(_n\) \( \otimes \) \( \mu \) is the
matrix with entry \( \mu_j \) in column \( j \) for all rows. That is, calculate
\( T(X - 1_n \otimes \mu) \) from the data set
\( X - 1_n \otimes \mu \). Calculate (7.8) or (7.9), as appropriate, from
the shifted data set. Calculate \( W(X - 1_n \otimes \mu) \) from (7.10). Then the random
set
\[ I = \{ \mu | W(X - 1_n \otimes \mu) \leq \chi^2_{J, \alpha} \} , \]
satisfies \( P[\mu \in I] = 1 - \alpha \) using the test inversion argument of (1.17). This is
the one-sample case of the region proposed by Kolassa and Seifu (2013).

Example 38 Recall again the blood pressure data set of Example 36.
Figure 7.2 exhibits the 0.05 contour of \( p \)-values for the multivariate test
constructed from sign rank tests for each of systolic and diastolic blood
pressure, and forms a 95% confidence region. Note the lack of convexity.

7.4 Two-sample Methods

Two-sample methods are generally of more interest than the preceding one-
sample methods. Consider a multivariate data set \( X_{ij} \) for \( i \in \{1, \ldots, M_1 + M_2\} \)
and \( j \in \{1, \ldots, J\} \), with data from the first sample occupying the first \( M_1 \)
rows of this matrix, and data from the second sample occupying the last \( M_2 \)
rows. Assume that the vectors \( (X_{i1}, \ldots, X_{iJ}) \) and \( (X'_{i1}, \ldots, X'_{iJ}) \), for all \( i \),
are independent if \( i \neq i' \). Assume further than the vectors \( (X_{i1}, \ldots, X_{iJ}) \) all
have the same distribution for \( i \leq M_1 \), and that the vectors \( (X_{i1}, \ldots, X_{iJ}) \) all
have the same distribution for \( i > M_1 \). Let \( g \) be a vector indicating group
membership; \( g_i = 1 \) if \( i \leq M_1 \), and \( g_i = 2 \) if \( i > M_1 \). As in §7.1, consider
testing and confidence interval questions.
7.4.1 Hypothesis Testing

Combine the techniques for one-dimensional two-sample testing testing of Chapter 3 and for multi-dimensional one-sample testing of §7.1. Consider the null hypothesis $H_0$ that the distribution of $(X_{i1}, \ldots, X_{iJ})$ is the same for all $i$.

7.4.1.1 Traditional normal-theory approach

Let $\bar{X}_1$ be the vector of sample means for observations in group 1, with components $\bar{X}_{1j} = \sum_{i|g_i=1} X_{ij}/M_1$. Let $\bar{X}_2$ be the vector of sample means for observations in group 2, with components $\bar{X}_{2j} = \sum_{i|g_i=2} X_{ij}/M_2$. Let $C_{1,j,j'}$ be the sample covariance for the group 1 values between responses $j$ and $j'$: $\sum_{i|g_i=1}(X_{ij} - \bar{X}_j)(X_{ij'} - \bar{X}'_j)/(M_1 - 1)$. Let $C_{2,j,j'}$ be the sample covariance for the group 2 values between responses $j$ and $j'$: $\sum_{i|g_i=2}(X_{ij} - \bar{X}_j)(X_{ij'} - \bar{X}'_j)/(M_2 - 1)$. Let $C_{j,j'}$ be the pooled sample covariance for the all observations: $C_{j,j'} = ((M_1 - 1)C_{1,j,j'} + (M_2 - 1)1_{Y,j,j'})/(M_1 + M_2 - 2)$. Then the
Hotelling two-sample statistic

\[ T^2 = \frac{M_1 M_2}{M_1 + M_2} (\bar{X}_1 - \bar{X}_2)^\top C^{-1} (\bar{X} - \bar{Y}) \]  

(7.11)

measures difference between sample mean vectors, in a way that accounts for sample variance, and combines the response variables. Furthermore, under the null hypothesis of equality of distribution, and assuming that this distribution is multivariate normal, \( \frac{M_1 + M_2 - J - 1}{M_1 + M_2 - 2} T^2 \sim \chi^2_{J, M_1 + M_2 - J - 1} \). Equality of distribution implies that the variance matrices of observations in the two groups are the same.

### 7.4.1.2 Permutation Testing

Under the null hypothesis of equality of distributions across the two groups, all assignments of the observed vectors among the two groups that keep the sizes of groups 1 and 2 at \( m \) and \( n \) respectively, are equally likely. Hence a permutation test that involves evaluating the Hotelling statistic, or any other parametric statistic, at each of the \( \binom{M_1 + M_2}{M_1} \) such reassignments of the observations to the groups,

\[ X^* = \{(X, g) | g \text{ has } M_1 \text{ entries that are 1 and } M_2 \text{ that are 2}\}; \]

\( p \)-values are calculated by counting the number of such statistics with values equal to or exceeding the observed value, and dividing the count by \( \binom{M_1 + M_2}{M_1} \).

Other marginal statistics may be combined; for example, one might use the Max \( t \)-statistic, defined by first calculating univariate \( t \)-statistics for each manifest variable, and reporting the maximum. This statistic is inherently one-sided, in that it presumes an alternative in which each marginal distribution for the second group is systematically larger than that of the first group. Alternatively, one might take the absolute value of the \( t \)-statistics before maximizing. One might do a parallel analysis with either the maximum of Wilcoxon rank sum statistics or the maximum of their absolute values, after shifting to have a null expectation zero.

Finally, one might apply permutation testing to the statistic (7.11), calculated on ranks instead of data values, to make the statistic less sensitive to extreme values.

### Example 39

Consider the data on wheat yields, in metric tons per hectare (Cox and Snell, 1981, Set 5), reflecting yields of six varieties of wheat grown at ten different experimental stations. Two of these varieties, Huntsman and Aton, are present at all ten stations, and so the analysis will include only these. Stations are included from three geographical regions of England; we compare those on the north to those elsewhere. The standard Hotelling two-sample test may be performed in R using
wheat<-as.data.frame(scan("set5.data",what=list(variety="",
y0=0,y1=0,y2=0,y3=0,y4=0,y5=0,y6=0,y7=0,y8=0,y9=0),
na.strings="-"))
# Observations here are represented by columns rather than by
# rows. Swap this. New column names are in first column.
dimnames(wheat)[[1]]<-wheat[,1]
wheat<-as.data.frame(t(wheat[,-1]))
dimnames(wheat)[[1]]<-c("E1","E2","N3","N4","N5","N6","W7",
"E8","E9","N10")
wheat$region<-factor(c("North","Other")[1+((
substring(dimnames(wheat)[[1]],1,1)!='N')],c("Other","North"))
attach(wheat)
plot(Huntsman,Atou,pch=(region=="North")+1,
main="Wheat Yeilds")
legend(6,5,legend=c("Other","North"),pch=1:2)

Data are plotted in Figure 7.3. The normal-theory p-value for testing
equality of the bivariate yeild distributions in the two regions is given by
library(Hotelling)#for hotelling.test
print(hotelling.test(Huntsman+Atou~region))

The p-value is 0.0327. Comparing this to the univariate results
t.test(Huntsman~region);t.test(Atou~region)
gives two substantially smaller p-values; in this case, treatment as a
multivariate distribution did not improve statistical power. On the other
hand, the normal quantile plot for Atou yeilds shows some lack of nor-
mality. Outliers do not appear to be present in these data, but if they
were, they could be addressed by performing the analysis on ranks, either
using asymptotic normality:
cat('Wheat rank test, normal theory p-values')
print(hotelling.test(rank(Huntsman)+rank(Atou)~region))

Permutation testing may be done to avoid the assumption of multi-
variate normality:
#Brute-force way to get estimate of permutation p-value for both T2 and the max t statistic.
cat('Permutation Tests for Wheat Data, Brute Force')
obsh<-hotelling.test(Huntsman+Atou~region)$stats$statistic
obst<-max(c(t.test(Huntsman~region)$statistic,
t.test(Atou~region)$statistic))
out<-array(NA,c(1000,2))
dimnames(out)<-list(NULL,c("Hotelling","t test"))
for(j in seq(dim(out)[[1]]){
  newr<-sample(region,length(region))
  hto<-hotelling.test(Huntsman+Atou~newr)
  out[j,1]<-hto$stats$statistic>=obsh
  out[j,2]<-max(t.test(Huntsman~newr)$statistic,
t.test(Atou~newr)$statistic)>=obst
}
apply(out,2,mean)
giving permutation p-values for the Hotelling and max-t statistics of 0.023 and 0.003 respectively. The smaller max-t statistic reflects the strong association between variety yields across stations. If one wants only the Hotelling statistic significance via permutation, one could use print(hotelling.test(Huntsman+Atou~region,perm=T, progBar=FALSE))

The argument progBar will print a progress bar, if desired, and an additional argument controls the number of random permutations.

FIGURE 7.3: Wheat Yeilds

7.4.1.3 Approximations to the Permutation Distribution
Let $T_j$ be the Mann-Whitney-Wilcoxon statistic using manifest variable $j$, for $j \in \{1,\ldots,J\}$. Let $\mathbf{T} = (T_1,\ldots,T_J)$. Let $\mathbf{\Sigma}$ be the variance matrix of
Multivariate Analysis

this statistic, \( \text{Var} [T] \), with elements defined as in
\[
\begin{pmatrix}
\sigma_{11} & \cdots & \sigma_{1J} \\
\vdots & \ddots & \vdots \\
\sigma_{J1} & \cdots & \sigma_{JJ}
\end{pmatrix}.
\]
The diagonal elements \( \sigma_{jj} \) are all known (to equal \( M_1 M_2 (M_2 + M_1 + 1)/12 \), but that’s not important here). The remaining entries of \( \Sigma \) must be estimated. For \( i = 1, \ldots, M_1 + M_2 \), let \( F_{ij} \) be the number of observations in group 2 that beat observation \( i \) on variable \( j \) if \( i \) is in group 1, and the number of observations in group 1 that \( i \) beats on variable \( j \), if \( i \) is in group 2. Then \( 4/(M_2 + M_1) \) times covariance matrix for \( F \) estimates the variance matrix of \( T \). Kawaguchi et al. (2011) provide details of these calculations. Superior performance can be obtained using known diagonal values, and estimated correlations for the remaining entries of the variance matrix (Chen and Kolassa, 2018).

**Example 40** Consider again the wheat yield data of Example 39. Asymptotic nonparametric testing is performed using

```r
library("ICSNP")#For rank.ctest and HotelingsT2.
rank.ctest(cbind(Huntsman,Atou)~region)

to obtain a \( p \)-value of 0.039, or

```r
rank.ctest(cbind(Huntsman,Atou)~region,scores="sign")
```

The final calculation above is a multivariate version of Mood's median test.

The above calculations may also be performed by first separating the matrix of yeilds into subsets by region, and giving `rank.ctest` two arguments corresponding to the two data matrices.

---

### 7.5 Exercises

1. The data set

   ```
   http://ftp.uni-bayreuth.de/math/statlib/datasets/federalistpapers.txt
   ```

   gives data from an analysis of a series of documents. The first column gives document number, the second gives the name of a text file, the third gives a group to which the text is assigned, the fourth represents a measure of the use of first person in the text, the fifth presents a measure of inner thinking, the sixth presents a measure of positivity, and the seventh presents a measure of negativity. There are other columns that you can ignore. (The version on line, above, has odd line breaks. A fixed version can be found at [stat.rutgers.edu/home/kolassa/960-555/federalistpapers.txt](http://stat.rutgers.edu/home/kolassa/960-555/federalistpapers.txt)).
Exercises

a. Test the null hypothesis that the multivariate distribution of first person, inner thinking, positivity, and negativity, are the same between groups 1 and 2, using a permutation test. Test at $\alpha = .05$.

b. Construct new variables, the excess of positivity over negativity, and the excess of thinking ahead over thinking behind, by subtracting variable six minus variable seven, and variable eight minus variable nine. Test the null hypothesis that the multivariate distribution these two new variables have a median zero, vs. the general alternative, using the multivariate version of the sign test. Test at $\alpha = .05$. 
Density Estimation

Suppose one observes a sample of independent and identical observations from density $f(x)$, and suppose one wants to estimate the density.

8.1 Histograms

The most elementary approach to this problem is the histogram, which represents the density as a bar chart.

The bar chart represents frequencies of the values of a set of categorical variable in various categories as the heights of bars. When the categories are ordered, one places the bars in the same order. To construct a histogram for a continuous random variable, then, coarsen the continuous variable into a categorical variable, whose categories are ranges of the scale of the original variable. Construct a bar plot for this categorical variable, again, with bars ordered according to the order of the intervals.

Because the choice of intervals is somewhat arbitrary, the boundary between bars is deemphasized by making the bars butt up against their neighbors. The most elementary version of the histogram has the height of the bar as the number of observations in the interval. A more sophisticated analysis makes the height of the bar represent the proportion of observations in the bar, and a still more sophisticated representation makes the area of the bar equal to the proportion of observation in the bar; this allows bars of unequal width while keeping the area under a portion of the curve to approximate the proportion of observation in that region. Such unequally-sized bars are unusual.

Construction of a histogram, then, requires selection of bar width and bar starting point.

One generally chooses the end points of intervals generating the bars to be round numbers. A deeper question is the number of such intervals. An early argument (Sturges, 1926) involved determining the largest number so that if the data were in proportion to a binomial distribution, every interval in range would be non-empty. This gives a number of bars proportional to the log of the sample size.

The histogram is determined once the bin width $\Delta_n$, and any bin sep-
aration point, is selected. Once $\Delta_n$ is selected, many bin separation points determine the same histogram; without loss of generality, choose the smallest non-negative value and denote it by $t_n$.

Scott (1979) calculates optimal sample size by minimizing mean square error of the histogram as an estimator of the density. Construct the histogram with bars of uniform width, and such that the bar area equals the proportion of observations out of the entire sample falling in the particular bar. Let $\hat{f}(x)$ represents the height of the histogram bar containing th potential data value $x$. The integrated mean squared error of the approximation is then

$$
\int_{-\infty}^{\infty} E \left[ |f(x) - \hat{f}(x)|^2 \right] dx.
$$

(8.1)

The quality of the histogram depends primarily on the width of the bin $\Delta_n$; the choice of $t_n$ is less important. Scott (1979) takes $t_n = 0$, and, using Taylor approximation techniques within the interval, shows that the integrated mean squared error of the approximation is

$$
1/(n\Delta_n) + \frac{1}{12} \Delta_n^2 \int_{-\infty}^{\infty} f'(x)^2 dx + O(1/n + \Delta_n^3).
$$

(8.2)

The first term in (8.2) represents the variance of the estimator, and the second term represents the square of the bias. Minimizing the sum of the first two terms gives the optimal bin size $h^* = [6/\int_{-\infty}^{\infty} f'(x)^2 dx]^{1/2} n^{-1/3}$, and, furthermore, approximates $\int_{-\infty}^{\infty} f'(x)^2 dx$ by using the value of the normal density with variance matching that of the data, to obtain $h^* = 3.49 sn^{-1/2}$, for $s$ the standard deviation of the sample.

### 8.2 Kernel density estimates

Histograms are easy to construct, but their choppy shape generally does not reflect our expectations about true density. A more sophisticated approach is kernel density estimation (Rosenblatt, 1956). Estimate the density as the average of densities centered at data points:

$$
\hat{f}(x) = (\Delta_n n)^{-1} \sum_{i=1}^{n} w((X_i - x)/\Delta_n).
$$

(8.3)

Here $w$ is a density, also called a kernel; that is, a non-negative function integrating to 1.

This estimator depends on the kernel $w$, and the smoothing parameter $\Delta_n$.

Some plausible choices for $w$ are the (standard) Gaussian density, the Epanechnikov kernel $w(x) = a - \frac{16a^3 x^2}{9}$ for $|x| \leq \frac{3}{4a}$; this is scaled this to have
unit variance with \( a = \frac{3}{4\sqrt{5}} \). This kernel minimizes the integrated mean square error (8.1) among symmetric kernels (Epanechnikov, 1969).

Another plausible kernel is the triangle kernel \( w(x) = 1/\sqrt{6} - |x/6| \), for \( |x| \leq \sqrt{6} \), and 0 otherwise. One might also consider the box kernel; consider the simplest version, \( w(x) = 1 \) if \( |x| \leq 1/2 \), rather than the one standardized to unit variance, \( w(x) = \sqrt{3}/2 \) if \( |x| \leq \sqrt{3} \). This kernel will be considered for pedagogical reasons below, although in practical terms its use abandons the advantages of the smoothness of the result.

**Example 41** Refer again to the nail arsenic data from Example 2. Figure 8.1 displays kernel density estimates with a default band width and for a variety of kernels. This figure was constructed using

```r
cat("\n Density estimation \n")
attach(arsenic)
# I am saving the density object at the same time I plot it.; R
plot(a<-density(nails),lty=1,
main="Density for Arsenic in Nails for Various Kernels")
lines(density(nails,kernel="epanechnikov"),lty=2)
lines(density(nails,kernel="triangular"),lty=3)
lines(density(nails,kernel="rectangular"),lty=4)
legend(1,2, lty=rep(1,4),
    legend=c("Normal","Quadratic","Triangular","Rectangular"))
```

Note that the rectangular kernel is excessively choppy, and fits the poorest. This figure would have been more clearly drawn in color.

Generally, any other density, generally symmetric and with a finite variance, may be used. The parameter \( \Delta_n \) is called the band width. This bandwidth \( \Delta_n \) should depend on spread of data, and \( n \); the spread of the data might be described using the standard deviation or interquartile range, or, less reliably, on the sample range.

The choice of bandwidth balances effects of variance and bias of the density estimator, just as the choice of bin width did for the histogram. If the bandwidth is too high, the density estimate will be too smooth, and hide features of data. If the bandwidth is too low, the density estimate will provide too much clutter to make understanding the distribution possible.

We might consider minimizing the mean squared error for \( x \) fixed, rather than after integration. That is, choose \( \Delta_n \) to minimize the mean square error of the estimate,

\[
\text{MSE}[\hat{f}(x)] = \text{Var} \left[ \hat{f}(x) \right] + \left( E \left[ \hat{f}(x) \right] - f(x) \right)^2 .
\]

Using the box kernel, the estimate \( \hat{f}(x) \) has a rescaled binomial distribution, and \( E \left[ \hat{f}(x) \right] = p/\Delta_n \) for \( p = F(x + \Delta_n/2) - F(x - \Delta_n/2) \). Expanding \( F(x) \) as
a Taylor series, noting that $F'(x) = f(x)$, and cancelling terms when possible,

$$p/\Delta_n = f(x) + \Delta_n^2 f''(x^*)/24 \text{ for some } x^* \in [x - \Delta_n/2, x + \Delta_n/2].$$  \hfill (8.4)

The bias of the estimator is $\Delta_n^2 f''(x^*)/24$. Hence, generally, if $\Delta_n$ does not converge to zero as $n \to \infty$, then bias does not converge to zero, and the estimate is inconsistent; hence consider only strategies for which $\lim_n \Delta_n = 0$.

Furthermore, since $\lim_n \Delta_n = 0$, then bias is approximated by $\Delta_n^2 C_1(x)$ for $C_1(x) = f''(x^*)/24$.  \hfill (8.5)

Furthermore, $\text{Var}\left[\hat{f}(x)\right] = p(1-p)/(n\Delta_n^2)$, and applying (8.4),

$$\text{Var}\left[\hat{f}(x)\right] = \left(1 - \Delta_n f(x) - \frac{1}{24} \Delta_n^3 f'''(x^*)\right) \left(f(x) + \frac{1}{24} \Delta_n^2 f''(x^*)\right),$$

and using the convergence of $\Delta_n$ to zero, we see

$$\text{Var}\left[\hat{f}(x)\right] \approx C_2(x)/(\Delta_n^2) \text{ for } C_2(x) = f(x).$$  \hfill (8.6)

Minimizing the mean square requires minimizing

$$C_2(x)/(\Delta_n^2) + C_1(x)^2 \Delta_n^4.$$  \hfill (8.7)
Kernel density estimates

Differentiating and setting to zero implies that $C_2(x)\Delta_n^{-2}/n = 4C_1(x)^2\Delta_n^3$, or

$$\Delta_n = 2^{-2/5}C_2(x)^{1/5}C_1(x)^{-2/5}n^{-1/5}$$

or

$$\Delta_n = f(x)^{1/5}2^{-2/5}(f''(x))^{-2/5}n^{-1/5}.$$  

(8.8)

This gives a bandwidth dependent on $x$; a bandwidth independent of $x$ may be constructed by minimizing the integral of the mean squared error (8.7), $C_2/(\Delta_n^2n) + C_1^2\Delta_n^4$, for $C_2 = \int_{-\infty}^{\infty} C_2(x) \, dx$ and $C_1 = \sqrt{\int_{-\infty}^{\infty} C_1^2(x) \, dx}$.

**Example 42** Refer again to the arsenic nail data of Examples 2 and 41. The kernel density estimate, using the suggested bandwidth, and bandwidths substantially larger and smaller than the optimal, are given in Figure 8.2. The excessively small band width follows separate data points closely, but obscures the way the clusters together. The excessively large bandwidth wipes out all detail from the data set. Note the large probability assigned negative arsenic concentrations by the estimate with the excessively large bandwidth. This plot was drawn in R using

```r
plot(density(nails,bw=a$bw/10),xlab="Arsenic",type="l",
     main="Density estimate with inopportune bandwidths")
lines(density(nails,bw=a$bw*5),lty=2)
legend(1,2,legend=paste("Band width default",c("/10","*5")),
        lty=c(1,2))
```

the object *a* is the bandwidth with the default bandwidth determined in Example 8.1.

**FIGURE 8.2:** Density estimate with inopportune bandwidths

Silverman (1986, §3.3) discusses a more general kernel $w(t)$. Equations (8.5)
and (8.6) hold, with

\[ C_1(x) = f(x) \int_{-\infty}^{\infty} w(t)^2 \, dt, \quad C_2(x) = \frac{1}{2} f''(x) \int_{-\infty}^{\infty} t^2 w(t) \, dt, \]

and (8.8) continues to hold. Sheather and Jones (1991) further discuss the constants involved.

Epanechnikov (1969) demonstrates that (8.3) extends to multivariate distributions; estimate the multivariate density \( f(x) \) from independent and identically distributed observations \( X_1, \ldots, X_n \), where \( X_i = (X_{i1}, \ldots, X_{id}) \), using

\[ \hat{f}(x) = n^{-1} \left( \prod_{j=1}^{d} \Delta_{nj} \right)^{-1} \sum_{i=1}^{n} w((X_{i1} - x_1)/\Delta_{n1}, \ldots, (X_{id} - x_d)/\Delta_{nd}). \] (8.9)

Here \( w \) is a density over \( \mathbb{R}^d \), and \( \Delta_{nj} \) are dimension-dependent bandwidth parameters. Epanechnikov (1969) uses multivariate densities that are products of separate densities for each dimension, but surveys work using more general kernels in two dimensions. The number of observations needed for precise estimation grows dramatically as the dimension \( d \) increases.

### 8.3 Exercises

1. The data at

   http://lib.stat.cmu.edu/datasets/CPS.85.Wages

reflects wages from 1985. The first 27 lines of this file represent an explanation of variables; delete these lines first, or skip them when you read the file. The first six fields are numeric. The sixth is hourly wage; you can skip everything else. Fit a kernel density estimate to the distribution of wage, and plot the result. Comment on what you see.
9

Regression Function Estimates

Consider modeling a collection of responses $Y_j$ as a function of explanatory variables $X_j = (X_{j1}, \ldots, X_{jK})$. That is, express $Y_j = g(X_j) + \epsilon_j$, with errors $\epsilon_j$ independent, and with various constraints on $g$. In this chapter, assume further that these errors are identically distributed. The least restrictive constraint allows for an arbitrary conditional expectation of $Y_j$; since in many cases each unique value of $X_j$ appears only a small number of times in the data set, fitting such a model reliably is difficult.

9.1 Standard Regression Inference

The most restrictive constraint is an affine function of $X_j$; that is, $g(X_j) = \mathbf{\beta}^\top X_j$; here $\mathbf{\beta}$ and $X_j$ are column vectors. Generally, the first component of each of the $X_j$ is 1, making the first component of $\mathbf{\beta}$ an intercept parameter.

The notation above, with $X_i$ capitalized, implies that the explanatory variables are random. Analyses considered in this chapter are generally performed conditionally on these response variables, and so their random nature might be ignored.

The standard approach chooses the regression parameters to minimize

$$\sum_{i=1}^{n} (Y_i - \mathbf{\beta}^\top X_i)^2;$$

that is,

$$\hat{\mathbf{\beta}} = \arg\min_{\mathbf{\beta}} \sum_{i=1}^{n} (Y_i - \mathbf{\beta}^\top X_i)^2 = (X^\top X)^{-1} X^\top Y, \quad (9.1)$$

for $X$ the $n \times K$ matrix with rows given by $X_i^\top$, and $Y$ is the column vector with entries $Y_i$. The estimator (9.1) is defined only if $X$ is of full rank; that is, if the inverse in (9.1) exists. Otherwise, the model is not identifiable, in that two different parameter vectors $\mathbf{\beta}$ and $\mathbf{\gamma}$ give the same fitted values, or $X\mathbf{\beta} = X\mathbf{\gamma}$.

When the errors $\epsilon_j$ are normally distributed, then the vector $\hat{\mathbf{\beta}}$ has a multivariate normal distribution, exactly, and

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{s^2 v_i}} \sim \mathcal{N}_{n-K}, \quad (9.2)$$
for \( s^2 = \sum_{i=1}^{n} (Y_i - \beta^\top X_i)^2 / (n - K) \), and \( v_i \) the entry in row and column \( i \) of \( X^\top X \)^{-1}. When the errors \( \epsilon_j \) are not normal, under certain circumstances, a central limit theorem justifies asymptotic normality for \( \hat{\beta} \), and (9.2) holds, approximately.

### 9.2 Kernel and Local Regression Smoothing

An intermediate level of constraint has \( g(x) \) continuous and differentiable, with curves that turn quickly discouraged. One method of fitting under such a constraint is kernel smoothing, and specifically as Nadaraya-Watson Smoothing (Nadaraya, 1964; Watson, 1964). One obtains an expression that is explicit rather than implicit; estimate \( g(x) \) as

\[
\hat{g}(x) = \frac{n}{\sum_{j=1}^{n} w((x - X_j)/\Delta_n)} \sum_{j=1}^{n} w((x - X_j)/\Delta_n).
\] (9.3)

The weight function can be the same was used for kernel density estimation. This weight function is often a normal density, or a uniform density centered at 0. Fan (1992) discusses a local regression smoother

\[
\hat{g}(x) = \sum_{\ell=0}^{L} \hat{\beta}_\ell x^\ell,
\] (9.4)

for \( L = 1 \) and

\[
\hat{\beta} = \arg\min \left\{ \sum_{j=1}^{n} \left( Y_j - \sum_{\ell=0}^{L} \hat{\beta}_\ell X_j^\ell \right)^2 w((x - X_j)/\Delta_n) \right\},
\] (9.5)

and argues that this estimator has smaller bias than (9.3). Köhler et al. (2014) present considerations for bandwidth selection.

**Example 43** Example 2 presents nail arsenic levels from a sample; arsenic levels in drinking water were also recorded, and we investigate the dependence of nail arsenic on arsenic in water. The data may be plotted in R using

```
attach(arsenic)
plot(water,nails,main="Nail and Water Arsenic Levels",
     xlab="Water",ylab="Nails",
     sub="Smoother fits. Bandwidth chosen by inspection.")
```

Kernel smoothing may be performed in R using the function `ksmooth`. Bandwidth was chosen by hand.
The function `loclpoly` from library `KernSmooth` applies a local regression smoother:

```r
library(KernSmooth)
lines(loclpoly(water,nails,bandwidth=.05,degree=1),lty=3)
```

Results are given in Figure 9.1. The box kernel performs poorly; a sample this small necessarily undermines smoothness of the box kernel fit. Perhaps the normal smoother is under-smoothed, but not by much. Library `KernSmooth` contains a tool `dpill` for automatically selecting bandwidth. The documentation for this function indicates that it sometimes fails, and, in fact, `dpill` failed in this case. This local regression smoother ignored the point with the largest values for each variable, giving the curve a concave rather than convex shape.

These data might have been jointly modeled on the square-root scale, to avoid issues relating to the distance of the point with the largest values for each variable from the rest of the data. An exercise suggests exploring this; in this case, the automatic bandwidth selector `dpill` returns a value. Figure 9.2 demonstrates the results of selecting a bandwidth too small. In this case, the Gaussian kernel gives results approximately constant in the neighborhood of each data point, and the box kernel result is not defined for portions of the domain, because both numerator and denominator in (9.3) are zero.

An alternative procedure uses as \( \hat{g}(x) \) the fitted value at \( x \) for low-degree (viz., linear or quadratic) regression of points with \( X_j \) near \( x \) (Cleveland, 1979; Savitzky and Golay, 1964). Cleveland and Devlin (1988) refer to this procedure as locally weighted regression (loess). One specifies the number of points \( k \), and up-weights points near \( x \) and down-weights them away from \( x \).

The weighting function is scaled to make the point in neighborhood farthest from \( x \) have a weight going down to zero.

The estimate is now (9.4), for

\[
\hat{\beta} = \arg\min_k \left( \sum_{j \in N(x)} \left( Y_j - \sum_{\ell=0}^{L} \beta_{\ell} X_j^\ell \right)^2 w((x - X_j)/\Delta_n(x)) \right),
\]

for \( N(x) \) the indices of the \( k \) closest points to \( x \), \( L = 2 \), and

\[
\Delta_n(x) = \max\{|X_j - x|, j \in N(x)\}.
\]

The local linear regression fit (9.5) with (9.3) differs from (9.6) with (9.3)
in the restriction of the consideration of points defining local regression parameter estimates to points near the point of interest, rather than using all points with positive values for a weighting function. Note further that for the loess procedure, bandwidth is determined implicitly, and depends on the point at which the smoother is applied.
Isotonic regression

A common weight function is \( w(x) = (1 - |x|^3)^3 \). The estimator \( \hat{g}(x) \) cannot be expressed in closed form. Cleveland and Devlin (1988) extend this loess technique to higher dimensions.

Example 44 Return again to the arsenic data previously examined in Example 43. Again, plot the data in R using

```r
attach(arsenic)
plot(water,nails,main="Nail and Water Arsenic Levels",
 xlab="Water",ylab="Nails", sub="Loess fits")
```

Loess smoothing may be performed in R using the function `loess`. Bandwidth was chosen by hand. Unlike in the case of `ksmooth`, `loess` does not provide output that can be plotted directly. A set of points at which to calculate the smoother must be specified; this is stored in the variable `x` below. Instead of specifying a bandwidth for the loess procedure, one specifies the number of observations contributing to each neighborhood from which the fit is calculated. In R, this is specified as the proportion of the total sample, via the input parameter `span`. Hence the second call below uses the entire data set.

```r
x<-min(water)+(0:50)*diff(range(water))/50
lines(x,y=predict(loess(nails~water),x))
lines(x,y=predict(loess(nails~water,span=1),x),lty=2)
```

Values of `span` above 1 specify for the bandwidth to be expanded beyond (9.7), thus increasing smoothing:

```r
lines(x,y=predict(loess(nails~water,span=10),x),lty=3)
legend(0.0,2.0,legend=c("Default span .75",
 "Span 1","Span 10"),lty=1:3)
```

Results are given in Figure 9.3. Note that the solution using the default proportion of the data 0.75 appears to under-smooth the data, and that raising this parameter above 1 misses detail of the shape of the relationship.

9.3 Isotonic regression

Many contexts justify a non-parametric relationship between variables, assuming that it is non-decreasing.

In the case in which \( K = 1 \), one might choose \( \hat{Y}_j \) to minimize \( \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 \) subject to \( \hat{Y}_j \geq \hat{Y}_i \) whenever \( X_j \geq X_i \). This model fitting is an example of quadratic programming. When \( K > 1 \), then the space of possible covariates

```
Loess fits do not have a natural ordering. One may construct a partial ordering; that is, for some distinct covariate vectors \( x \) and \( y \), neither \( x \) nor \( y \) is ordered higher. For example, one may define the partial ordering \((x_1, \ldots, x_K) \preceq (y_1, \ldots, y_K)\) if \( x_i \preceq y_i \) for all \( i \); that is, consider one vector as greater than or equal to the other if and only if each component of the first vectors does not exceed the same component of the second vector. Such an ordering lacks the property that any two elements of the set may be compared, and so is called a partial ordering. For example, neither \((2, 1) \preceq (1, 2)\) nor \((1, 2) \preceq (2, 1)\).

Extending this back to the univariate case, since for any two distinct real numbers, one can be determined to be the smaller and the other is the larger. Brunk (1955) considers introduces the notion of model fitting that respects the partial ordering \( g(x) \leq g(y) \) if \( x \preceq y \). Dykstra (1981) reviews an algorithm for fitting such a model, called the pooled adjacent violators (PAV) algorithm, and produces theoretical justification for this algorithm. Best and Chakravarti (1990) reviews more general algorithmic considerations.

Example 45 Example 2 presents arsenic levels in water and nails. Isotonic regression may be performed in R using the function \texttt{isoreg}, and \texttt{R} provides a plotting method for isotonic regression results. Hence the regression output may be plotted directly, using

```r
attach(arsenic)
plot(isoreg(water,nails),
     main="Nail Arsenic and Water Arsenic",
     xlab="Water As",ylab="Nail As")
```
Results are given in Figure 9.4.

FIGURE 9.4: Nail Arsenic and Water Arsenic

9.4 Splines

Splines represent a technique for drawing a smooth curve approximating the relationship between an explanatory variable $x$ and a response variable $y$, based on observed pairs of points $(X_1, Y_1), \ldots, (X_n, Y_n)$. The objective is to describe the dependence of $y$ on $x$ between two points $x_0$ and $x_N$. One estimates the dependence by picking $N - 1$ intermediate points $x_1 < x_2 < \cdots < x_{N-2} < x_{N-1}$. The intermediate points are (called knots. One then determines a polynomial of degree $M$ between $x_{j-1}$ and $x_j$, constrained so that the derivatives of order up to $M - 1$ match up at knots. Denote the fitted mean by $\hat{g}(x)$

Taken to an extreme, if all $X_j$ are unique, then one can fit all $n$ points with a polynomial of degree $n - 1$; this, however, will yield a fit with unrealistically extreme fluctuations. Instead, choose the polynomials to minimize

$$\sum_{j=1}^{n} (Y_j - \hat{g}(X_j))^2 + \lambda \int_{X_{(1)}}^{X_{(n)}} \hat{g}''(x) \, dx.$$
Example 46 Again revisit the arsenic data of Example 43. The spline is fit using R as

```
attach(arsenic)
plot(water,nails, main="Nail Arsenic and Water Arsenic")
hgrid<-min(water)+(0:50)*diff(range(water))/50
lines(predict(spl<-smooth.spline(water,nails),hgrid))
lines(predict(smooth.spline(water,nails,spar=.1),hgrid),lty=3)
lines(predict(smooth.spline(water,nails,spar=.5),hgrid),lty=2)
legend(1250,1110,lty=1:3,col=1:3,
legend=paste("Smoothing",c(round(spl$spar,3),.5,.1),
"Default","=",""))
```

Here `hgrid` is a set of points at which to calculate the spline fit. As with most of the smoothing methods, R contains a lines method that will plot predicted values directly. See Figure 9.5. The optimal smoothing parameter value yields an almost straight line. Smaller smoothing parameters yield over-fitted curves.

FIGURE 9.5: Nail Arsenic and Water Arsenic

Schoenberg (1946) introduced this technique.
Quantile Regression

9.5 Quantile Regression

Least squares regression, as in (9.1), provides parameter estimates minimizing the sum of squares of errors in fitting. Such an approach is often criticized as allowing excessive influence to outliers, and, hence, is non-robust. As an alternative, consider minimizing the sum of differences

$$\sum_{j=1}^{n} |Y_j - \beta_1 - \beta_2 X_j|.$$  \hspace{1cm} (9.8)

Note the absence of $\cdot^2$, which would be present in regular least squares regression.

Equivalently, one might minimize

$$\sum_{j} e_j^+ + e_j^-,$$  \hspace{1cm} (9.9)

for

$$e_j^+ \geq 0, \ e_j^- \geq 0, \ e_j^+ - e_j^- = Y_j - \beta_1 - \beta_2 X_j \forall j.$$  \hspace{1cm} (9.10)

Figure 9.6 demonstrates the estimation of an intercept (in this case, for the arsenic example to be discussed below). Here the intercept is held at its optimal value. The figure is drawn in R using

```r
attach(arsenic)
rqo<-summary(rq(nails~water))$coef
regp<-rqo[2,2]+diff(rqo[2,2:3])*(0:100)/100
m<-apply(abs(outer(nails,rep(1,length(regp)))-
outer(water,regp)-rqo[1,1]),2,sum)
plot(regp,m,type="l",xlab="Regression Parameter",
ylab="Sum of absolute values of residuals",
main="L1 fit for Arsenic Large Scale",
sub="Intercept fixed at optimum")
```

The objective function is not differentiable; the optimum is at the point. For other data sets, the optimum might lie along a horizontal line segment, and so often the minimizer of (9.8) is not unique.

This solution is called $L^1$ regression, after the power on $|Y_j - \beta_1 - \beta_2 X_j|$. This solution is also called quantile regression. To understand the motivation behind this term, note that if $\beta_2 = 0$, then the best $\beta_1$ is median. More generally, this technique fits parameters so that the median, rather than the expectation, of the residuals is zero. Just as median is not always uniquely defined, these estimates are not necessarily uniquely defined.

Stigler (1984) attributes the suggestion to fit the linear model using (9.9) to Roger Boscovich, and a partial solution to the minimization problem to
Regression Function Estimates

FIGURE 9.6: L1 fit for Rabbit Data Large Scale

TABLE 9.1: Achieved Coverage for 90% Confidence Intervals for $L_1$ and $L_2$ Regression, with Various Error Distributions

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Cauchy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_1$</td>
<td>0.898</td>
<td>0.948</td>
</tr>
<tr>
<td>$L_2$</td>
<td>0.894</td>
<td>0.885</td>
</tr>
</tbody>
</table>

Thomas Simpson, both circa 1760. Stigler (1973) attributes the solution of the problem when $\beta_1 = 0$ to Laplace (1818). Koenker (2000) traces further developments through the work of Edgeworth.

One can adapt this technique to fit quantiles of the errors other than the median. One can replace the objective (9.9) by

$$
\tau \sum_j e_j^+ + (1 - \tau) \sum_j e_j^-, \quad (9.11)
$$

still subject to constraints (9.10). This change causes the regression line to run through the $1 - \tau$ quantile of $Y|X$.

9.5.1 Fitting the Quantile Regression Model

This minimization of (9.9) or (9.11), subject to (9.10), is an example of linear programming. The solution to this minimization is more computationally intensive than the minimization of (9.1), and, unlike as in the least squares
Quantile Regression

Case, does not have a closed-form expression. This added difficulty arises because the objectives (9.9) or (9.11) are non-differentiable. At the minimizer $\beta$, the objective function should not decrease as $\beta$ moves away from $\hat{\beta}$. Since $\sum_{j=1}^{n} |Y_j - \beta_1 - \beta_2 X_j|$ is a piecewise linear function of $\beta_1$ and $\beta_2$, the function $S(\beta)$ giving “derivative” should be piecewise constant, with jumps. Then $\hat{\beta}$ should satisfy $S(\hat{\beta}) \approx 0$ The estimate either sets the function to 0, or is a point where it jumps across 0. This score function $S(\beta)$ may be expressed as a rank statistic, and the null hypothesis $H_0: \beta = \beta^0$ may be tested by comparing $S(\beta^0)$ to 0, either exactly or asymptotically. This test can be inverted to give confidence sets for $\beta$.

Solution to the more general problem seems to have been an early application of general linear programming methods.

Example 47 Consider again the arsenic data of Example 43. The quantile regression fitter is found in R in library quantreg. The following commands fit this model:

```r
library("quantreg")#Need for rq
rq(nails~water)

to give the results

Call: rq(formula = nails ~ water)

tau: [1] 0.5

Coefficients: coefficients lower bd upper bd
(Intercept) 0.11420 0.10254 0.14093
water 15.60440 3.63161 18.07802

Hence the best estimate of the slope of the linear relationship of arsenic in nails to arsenic in water is 15.604, with 90% confidence interval (3.632,18.078).

Compare this to the Theil-Sen estimator:

library(deming)
attach(arsenic); theilsen(nails~water,conf=.9)$coef

yielding

(Intercept) water
0.1167158 14.2156504

A confidence interval for the slope parameter is calculated using

library(MultNonParam);theil(water,nails)
giving the confidence interval (8.12, 16.41); note the higher precision of this approach.

Example 48  Consider the blood pressure data set of Example (36). Explain systolic blood pressure after treatment in terms of systolic blood pressure before treatment:

```r
#Expect warnings here about nonunique estimators. This is OK.
rqout <- rq(bp$sp~bp$spb); summary(rqout)
```

```r
             coefficients lower bd upper bd  
(Intercept)   -11.41026     -34.76720  29.00312  
bp$spb       0.94872        0.69282  1.11167
```

Hence blood pressure post-treatment increases with blood pressure pre-treatment; the best estimate is indicates that a one-for-one increase is plausible, in that 1 sits inside the confidence interval. One can also fit the .2 quantile:

```r
rqouttt <- rq(bp$sp~bp$spb, tau=.2); summary(rqouttt)
```

```r
Coefficients:
             coefficients lower bd upper bd  
(Intercept)     17.25000     -68.48262  23.34294  
bp$spb          0.75000        0.40129  1.20640
```

One can plot these relationships:

```r
plot(bp$spb, bp$spa, main="Systolic Blood Pressure", xlab="Before Treatment", ylab="After Treatment")
abline(rqout); abline(rqouttt, lty=2)
legend(150, 200, legend=c("Median Regression", ".2 Quantile Regression"), lty=1:2)
```

Results are plotted in Figure 9.7. Note the extreme lack of parallelism.

9.6 Exercises

1. The data at

   http://lib.stat.cmu.edu/datasets/CPS_85_Wages
a. Fit a relationship of wage on educational level, in such a way as to minimize the sum of absolute value of residuals. Plot this, and compare with the regular least squares regression.

b. Fit a relationship of wage on educational level, in such a way as to enforce monotonicity. Superimpose a plot of this relationship on a plot of the data.

c. Fit a kernel smoother to estimate the dependence of wage on educational level, and compare it to the results for the LOESS smoother.

d. Fit a Smoothing spline to estimate the dependence of wage on educational level, and compare it to the results for the LOESS smoother and the kernel smoother.

2. Repeat the analysis of Example 43 for arenic levels on the square root scale. As an intermediate step, calculate the optimal bandwidth for the local linear smoother. Plot your results. Compare with the results in Figure 9.1.
This chapter addresses the construction of a general procedure for constructing confidence intervals for a parameter with minimal assumptions about its distribution.

The solution is to treat data as though they are the population of interest. Often, a plausible estimator of the population parameter is the sample analog of that parameter. For example, an estimator of a population expectation is the sample mean, which is the expectation of the population formed by placing equal weights on each sample point. As a second example, an estimator of a population median is the sample median, which is the median of population formed by placing equal weights on each sample point.

The most elemental version of the bootstrap procedure is to use the distribution of an estimator formed from independent draws of this empirical distribution as a substitute for the unknown sampling distribution of the estimator. This distribution is, in principle, known exactly, but, in practice, if a sample of size \( n \) comes from a continuous distribution, and contains no ties, the distribution one wants for computation is a function of as many as \( n^n \) vectors of length \( n \), and may contain as many as \( n^n \) estimator values.

Generally, one proceeds via random sampling. Choose a number of random samples \( B \) to draw. One draws a new random samples from the population represented by the original sample, with replacement. This sampling with replacement distinguishes the bootstrap from previous permutation techniques.

For random sample \( i \), evaluate the estimator on this sample, and call it \( \hat{\theta}_{B,i} \). Express the ordered values of \( \hat{\theta}_{B,i} \) by \( \hat{\theta}_{B,(i)} \).

### 10.1 Univariate Bootstrap Techniques

I now review various strategies for using these samples in the most simple univariate contexts, and will use terminology consistent with the R package `boot`. 

159
10.1.1 The Normal Method

Consider the standard error estimate $\hat{\phi}^2 = \sum_{i=1}^{B}(\theta_{B,i} - \theta)^2/B$. Unfortunately, we don’t know $\theta$. Using $\hat{\theta}_B = \sum_{i=1}^{B}\theta_{B,i}/B$ in place of $\theta$ gives undercoverage. Use $B - 1$ instead of $B$ in denominator of $\hat{\phi}^2$, or $\hat{\theta}$ in place of $\theta$, to respond to this undercoverage. Then the standard deviation estimate $\hat{\phi}$ is the sample standard deviation of the bootstrap samples (Efron, 1981). Then a $1 - \alpha$ confidence interval is $\hat{\theta} \pm \hat{\phi} z_{1-\alpha/2}$ for $\varphi$ the sample standard deviation of bootstrap samples.

10.1.2 Basic Interval

Choose an estimate $\hat{\theta}$ for a parameter $\theta$. Suppose that the distribution of $\hat{\theta} - \theta$ does not depend on $\theta$. We use the distribution of $\hat{\theta}_{B,i} - \hat{\theta}$ as a proxy for that of $\hat{\theta} - \theta$. Let $v_L$ and $v_C$ satisfy $P[v_L \leq \hat{\theta} - \theta \leq v_U] = 1 - \alpha$. Then

$$P[\hat{\theta} - v_U \leq \theta \leq \hat{\theta} - v_L] = 1 - \alpha.$$  

Let $u_L$ and $u_U$ be quantiles of $\hat{\theta}_{B,i}$.

Order statistics from the bootstrap sample are used to estimate quantiles of the bootstrap distribution (again, supported on as many as $n^n$ points, and computationally intractable exactly for large $n$.) The most naive approach uses $\hat{\theta}_{B(i)}$ to represent quantile $i/B$ of the bootstrap distribution. By this logic, $\hat{\theta}_{B(1)}$ estimates the $1/B$ quantile, and $\hat{\theta}_{B(B)}$ represents 1 quantile. Conceptually, the estimation problem ought to be symmetric if we swap the order of bootstrap observations, but these naive quantiles are not. The upper quantile is wrong, since either for residual or for percentile bootstrap, the population bootstrap observations, but these naive quantiles are not. The upper quantile definition symmetric by making $\hat{\theta}_{B(i)}$ to represent quantile $i/(B+1)$ of this distribution.

Then, for an interval with confidence $1 - \alpha$,

$$u_L = \theta_{B,(\alpha(B+1)/2)} \quad \text{and} \quad u_U = \theta_{B,(1-\alpha/2)(B+1)}.$$  

(10.1)

Then quantiles $v_L$ and $v_C$ of $\hat{\theta}_{B,i} - \hat{\theta}$ are $v_L = u_L - \hat{\theta}$, and $v_C = u_U - \hat{\theta}$. Then a confidence interval for $\theta$ is

$$(\hat{\theta} - v_U, \hat{\theta} - v_L) = (2\hat{\theta} - \theta_{B,(1-\alpha/2)(B+1)}), (2\hat{\theta} - \theta_{B,(\alpha(B+1)/2)}).$$

(10.2)

Davison and Hinkley (1997, p. 29) refer to this as the basic bootstrap confidence interval method.

If $\varphi = \psi(\theta)$ for a monotone function $\psi$, then one might use $\hat{\varphi} = \psi(\hat{\theta})$ as an estimate of $\varphi$. The bootstrap values are $\psi(\hat{\theta}_{B,i})$, and a confidence interval for $\varphi$ is given by $(2\psi(\hat{\theta}) - \psi(\theta_{B,(\alpha(B+1)/2)}), 2\psi(\hat{\theta}) - \psi(\theta_{B,(1-\alpha/2)(B+1)}))$.

10.1.3 The Percentile Method

Suppose that $\hat{\theta}$ has distribution symmetric about $\theta$. In this case, we assume treat $\hat{\theta} - v_U$ and $-\hat{\theta} + v_L$ as equivalent approximations to interchange, and so
one can use $v_U = \hat{\theta} - u_L$, $v_L = \hat{\theta} - u_U$. The confidence interval is now

$$(u_L, u_U) = (\theta_{B, \alpha(B+1)/2}, \theta_{B, (1-\alpha/2)(B+1)})$$


The comment in §10.1.2 about transformations of parameters corresponding to the same transformation applied to interval endpoints still holds, and implies that the percentile method applies if such a transformation to symmetry (independent of $\theta$) exists, even if the transformation is unknown.

This method is referred to as the percentile method (Efron, 1981), and as the usual method by Shao and Tu (1995, p. 132f).

**Example 49** Consider again the arsenic data of Example 2. We calculate a confidence interval for the median.

arsenic<-as.data.frame(scan('arsenic.dat',
what=list(age=0,sex=0,drink=0,cook=0,water=0,nails=0)))
cat('\n Bootstrap for Nail Arsenic \n')
cat('\n Percentile Bootstrap for Median\n')
medians<-rep(NA,999)
attach(arsenic)
for(j in seq(length(medians)))
  medians[j]<-median(sample(nails,length(nails),replace=TRUE))
ci<-quantile(medians,probs=c(.025,.975))
cat(ci,\n"
"
"
gives the confidence interval (0.119 0.310), and

cat(\n"
 Residual Bootstrap for Median\n"
ci<-2*median(arsenic$nails)-rev(ci)
cat(ci,"\n"
"
gives the confidence interval (0.040,0.231). Recall that the estimate of the density for the nail arsenic values was plotted in Figure 8.1. This distribution is markedly asymmetric, and so the percentile bootstrap is not reliable; use the residual bootstrap.

### 10.1.4 $BC_a$ Method

This method was introduced by Efron (1987), who called it the $BC_a$ method; it extends his $BC$ method, which he terms “bias-corrected”; Efron and Tibshirani (1993) refer to the method of this section as bias corrected and accelerated. Again, suppose one desires a confidence interval for $\theta$, with estimator $\hat{\theta}$. As in §10.1.3, suppose $\hat{\theta}$ can be transformed to symmetry using a transformation $\phi$ (which need not be known), adding an additional restriction, assume that the transformation leads to normality. Suppose further that $\phi(\hat{\theta})$ has a standard deviation that depends linearly on $\phi(\theta)$, and that $\phi(\hat{\theta})$ has
a bias that depends linearly on the standard deviation of $\phi(\hat{\theta})$. That is, assume that there exists a transformation $\phi$, and constants $a$ and $\zeta$ such that 

$$
\frac{(\phi(\hat{\theta}) - \phi(\theta))}{1 + a\phi(\theta)} + \zeta \text{ is approximately standard Gaussian. That is,}
$$

$$
P\left[ \frac{\phi(\hat{\theta}) - \phi(\theta)}{1 + a\phi(\theta)} + \zeta \leq x \right] \approx \Phi(x)
$$

and

$$
P\left[ \phi(\hat{\theta}) \leq y \right] \approx \Phi\left( \frac{y + \zeta(1 + a\phi(\theta)) - \phi(\theta)}{1 + a\phi(\theta)} \right). \quad (10.3)
$$

Let $\theta^*$ be the value of $\theta$ giving quantile $1 - \alpha$ for $\hat{\theta}$. Substituting $\theta^*$ for $\theta$ into (10.3), and equating $\theta$ with $\hat{\theta}$, gives

$$
\Phi\left( \frac{-az\alpha + \zeta(a\zeta - 2) + z\alpha}{a(\zeta - z\alpha) - 1} \right) = \Phi\left( \zeta + \frac{\zeta - z\alpha}{1 - a(\zeta - z\alpha)} \right). \quad (10.4)
$$

Equating the bootstrap distribution function to this tail probability, the corresponding quantile is (10.4). The quantity $a$ is called the acceleration constant.

The bias $\zeta$ may be estimated by the difference between the estimate and the median of the bootstrap samples, and $a$ may be estimated using the skewness of the bootstrap sample.

**Example 50** Return again to the nail arsenic values of the previous example. We again generate a confidence interval for the median. The $BC_a$ method, and the previous two methods, may be obtained using the package boot.

```r
library("boot")
gives boot and boot.ci.

# Define the function to be applied to data sets to get the parameter to be bootstrapped. Second component # is an estimate of its scale, squared.
mymedian <- function(x, index) return(median(x[index]))
boot.ci(boot(arsenic$nails, mymedian, 9999)) to give

Intervals :
Level Normal Basic
95% ( 0.0158, 0.2733 ) ( 0.0400, 0.2310 )

Level Percentile BCA
95% ( 0.119, 0.310 ) ( 0.118, 0.277 )

In this case, the normal and percentile intervals are suspect, because of the asymmetry of the distribution. In this case, the more reliable interval is the bias corrected and accelerated interval.

Recall that an exact confidence interval may be constructed using
library(MultNonParam); exactquantileci(arsenic$nails)

to obtain the interval (0.118, 0.354). Efron (1981) notes that this exact
interval will generally agree closely with the percentile bootstrap approach.

Example 51 A similar approach may be taken to a confidence interval
for the standard deviation of nail arsenic values. In this case, first change
to the log scale.

logscale<function(x,index) return(log(sd(x[index])))
sdbootsamp<boot(arsenic$nails,logscale,9999)
sdoutput<boot.ci(sdbootsamp)

Figure 10.1 shows the bootstrap samples; the plot produced by
plot(density(sdbootsamp$t))

shows a highly asymmetric distribution, and the BCa correction for asym-
metry is strong. (The actual bootstrap distribution is supported on a large
but finite number of values, and is hence discrete and does not have a den-
sity; the plot is heuristic only.) The output from boot.ci contains some
information not generally revealed using its default printing method. In
particular, sdoutput$bca is a vector with five numeric components. The
first of these is the confidence level. The fourth and fifth are the result-
ing confidence interval end points. The second and third give quantiles
resulting from (10.4). The upper quantile is very close to the maximum
value of 9999; boot.ci gives a warning, and serious application of BCa
intervals would better be done using more bootstrap samples. Rerunning
with

boot.ci(boot(arsenic$nails,logscale,99999))$bca

gives the BCa 0.95 interval (-1.647,-0.096) for the log of standard devi-
ation of nail arsenic.

10.1.5 Summary So Far, and More Examples

Table 10.1 contains the observed coverages for nominal 0.90 confidence inter-
vals for the medians of simulated data sets, for various distributions. Random
samples of size 20 were drawn 1000 times from each distribution, and in each
case, 9999 bootstrap replicates were constructed. For each random distribu-
tion, the interval was checked to see whether it contained the true population
median. Calculations were performed using

cat("n Performance of Bootstrap for median\n")
source("common.R")
fun.testboot(mymedian, nsamp=1000, sampsize=20, 
dists=c("rexp","runif","rcauchy"),
true=c(log(2),.5,0), alpha=.1)

TABLE 10.1: Observed coverage for nominal 0.90 Bootstrap Intervals, 10 Observations

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Normal</th>
<th>Basic</th>
<th>Percentile</th>
<th>BCa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>0.863</td>
<td>0.772</td>
<td>0.891</td>
<td>0.895</td>
</tr>
<tr>
<td>Uniform</td>
<td>0.833</td>
<td>0.754</td>
<td>0.901</td>
<td>0.900</td>
</tr>
<tr>
<td>Cauchy</td>
<td>0.917</td>
<td>0.835</td>
<td>0.875</td>
<td>0.862</td>
</tr>
</tbody>
</table>

The percentile bootstrap performed remarkably well for the exponential distribution, in light of the interval's construction assuming symmetry. None showed significant degradation when data came from a very heavy-tailed distribution.
10.2 Bootstrapping Multivariate Data Sets

The bootstrapping idea of the previous section extends directly to more complicated data contexts. The most immediate extension involves resampling data vectors as a group. When applied in regression contexts, responses per subject are often denoted by $Y_j$, and explanatory vector by $X_j$. Our first approach sample the ensembles $(X_j, Y_j)$ as a whole; that is, if a subject's response is selected for the sample, the explanatory variables will be selected, and with the same multiplicity. This is called a random $X$ bootstrap.

**Example 52** Consider again the brain-volume data of Example 24. We use a random $X$ bootstrap to get confidence intervals for the Pearson correlation between first and second twin brain volumes. First, define the correlation function

```r
rho<-function(d,idx) return(cor(d[idx,1],d[idx,2]))
```

and calculate the bootstrap samples

```r
bootout<-boot(both[,c("v1","v2")],rho,9999)
```

Figure 10.2 presents a histogram of the sample, with vertical lines marking the estimate and percentile confidence interval end points:

```r
hist(bootout$t,freq=FALSE)
legend(-.4,6,lty=1:2,legend=c("Estimate","Confidence Bounds"))
abline(v=bootout$t0)
abline(v=sort(bootout$t)[(bootout$R+1)*c(.025,.975)],lty=2)
```

Confidence intervals may be calculated using

```r
boot.ci(bootout)
```

to give

<table>
<thead>
<tr>
<th>Intervals</th>
<th>Level</th>
<th>Normal</th>
<th>Basic</th>
</tr>
</thead>
<tbody>
<tr>
<td>95%</td>
<td></td>
<td>( 0.7463, 1.1092 )</td>
<td>( 0.8516, 1.1571 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level</th>
<th>Percentile</th>
<th>BCa</th>
</tr>
</thead>
<tbody>
<tr>
<td>95%</td>
<td>( 0.6718, 0.9773 )</td>
<td>( 0.5609, 0.9719 )</td>
</tr>
</tbody>
</table>

The normal and basic (that is, residual) intervals fall outside the allowable values for the correlation coefficient; Figure 10.2 makes this predictable. The basic interval is the percentile interval, represented in the figure by broken lines, reflected about the estimate. This forces the upper end point
above the upper bound. Similarly, the variance used for the normal interval is increased by the long tail to the left, pushing the upper bound above 1. The BC$_a$ approach is tailored to the asymmetry in the bootstrap replicates, and is more reliable.

FIGURE 10.2: Histogram of Bootstrap Samples of Brain Volume Correlations

10.2.1 Regression Models and the Studentized Bootstrap Method

We apply random $X$ bootstrap techniques to inference on a regression parameter. The ordinary least-squares regression estimator has a standard error that depends on the covariate patterns in the data set, and hence regression parameters from random bootstrap samples from a data set will have different precisions associated with them. Hence members of a collection of bootstrapped regression parameters are not identically distributed.

Recall the Gaussian-theory approach to constructing confidence intervals for a regression parameter $\hat{\theta} \pm ts.e.(\hat{\theta})$, where $t$ is a $T$ quantile, with an appropriate degree of freedom. The ratio $(\hat{\theta} - \theta)/s.e.(\hat{\theta})$ is called studentized, in that it is divided by an empirical estimator of its variance in order to make
it comparable to a reference distribution. In this section we apply this idea to bootstrap sampling.

Suppose that bootstrap samples are still independent, but, as in the previous section, the variance of the estimator may vary with the unknown parameter. Suppose furthermore that a practical estimate of the variability of the estimator also exists. Then, re-sample $B$ data sets as above, and for each re-sampled data set, calculate the estimate $\hat{\theta}_{B,i}$, and an estimate of variability $\hat{\sigma}_{B,i}$. Let empirical distribution of $T_{B,i} = (\hat{\theta}_{B,i} - \hat{\theta})/\hat{\sigma}_{B,i}$ stand in for distribution of $T = (\hat{\theta} - \theta)/\sigma$. This distribution doesn’t involve unknown parameters; such a quantity is called pivotal. One might apply this in conjunction with the percentile method, as in §10.1.3. In order to construct a $1 - \alpha$ confidence interval, let $t_L$ and $t_U$ be the $\alpha/2$ and $1 - \alpha/2$ quantiles from bootstrap distribution. One might use, for example, $\theta = E[X_j]$ and $\hat{\theta} = \bar{X}$, $\hat{\sigma} = S/\sqrt{n}$.

**Example 53** Consider again the brain-volume data of Example 24. I calculate the studentized interval. The biggest innovation here is to add standard error information to the bootstrap samples. In R, this is done by having the function to be run on the data samples return a vector with two components. The second component is the measure of variability, and should be on the squared scale. Hence I return the square of the standard error.

```r
regparam<-function(d,idx)
  return(summary(lm(d[idx,2]~d[idx,1]))$coefficients[2,1:2]^(1:2))
```

Now the calculations may be run as before:

```r
boot.ci(boot(both[,c("v1","v2")],regparam,9999))
```

to obtain

<table>
<thead>
<tr>
<th>Intervals</th>
<th>Normal</th>
<th>Basic</th>
<th>Studentized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 95%</td>
<td>0.2836, 1.1388</td>
<td>0.1746, 0.9239</td>
<td>0.4733, 1.0067</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 95%</th>
<th>Percentile</th>
<th>BCa</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6296, 1.3788</td>
<td>0.5962, 1.2082</td>
<td></td>
</tr>
</tbody>
</table>

*Note that the warning about missing information for the studentized interval no longer appears. The appropriate interval is the Studentized interval; use that one.*
10.3 Exercises

1. The data set

\[ http://ftp.uni-bayreuth.de/math/statlib/datasets/lupus \]

gives data on 87 lupus patients. The fourth column gives transformed disease duration.

a. Give a 90\% bootstrap confidence interval for the mean transformed disease duration, using the basic, studentized, and BCa approaches.

b. Give a jackknife estimate of the bias of the mean transformed disease duration.

2. The data set at

\[ http://ftp.uni-bayreuth.de/math/statlib/datasets/federalistpapers.txt \]

gives data from an analysis of a series of documents. The first column gives document number, the second gives the name of a text file, the third gives a group to which the text is assigned, the fourth represents a measure of the use of first person in the text, and the fifth presents a measure of inner thinking. There are other columns that you can ignore. (The version on line, above, has odd line breaks. A fixed version can be found at

\[ \text{stat.rutgers.edu/home/kolassa/960-555/federalistpapers.txt} \]

).  

a. Calculate a bootstrap confidence interval, with confidence level .95, for the regression coefficient of inner thinking regressed on first person. Test at \( \alpha = .05 \). Provide basic, studentized, and BCa intervals. Do the fixed-X bootstrap.

b. Calculate a bootstrap confidence interval, with confidence level .95, for the regression coefficient of inner thinking regressed on first person. Provide basic, studentized, and BCa intervals. Do not do the fixed-X bootstrap; re-sample pairs of data.

c. Calculate a bootstrap confidence interval, with confidence level .95, for the \( R^2 \) statistic for inner thinking regressed on first person. Provide basic and BCa intervals. Do not do the fixed-X bootstrap; re-sample pairs of data.
Analysis using the SAS System

The preceding chapters detailed statistical computations using R. This appendix describes parallel computations using SAS. Some of the computations require the use of macros, available at

stat.rutgers.edu/home/kolassa/960-555/common.sas.

Include this file before running any macros below, by first downloading into the current working directory, and adding

%include "/folders/myfolders/common.sas";

to your SAS program before using the macros. Adjust the /folders/myfolders/ to reflect your local configuration. Furthermore, the examples use data sets that need to be read into SAS before analysis; reading the data set is given below the first time the data set is used.

A.1 Table 2.1

Table 2.1 gives levels for various tests applied to data from various distributions. This simulation may be invoked by running the commands

%test1(10,1000,0); proc print data=size noobs; run;
%test1(40,1000,0); proc print data=size noobs; run;

The leading percent sign identifies test1 as a macro. This macro simulates one-sample testing, and gives the proportion of samples leading to rejection of the null hypothesis. The first argument is the number of observations in each sample. The second is the number of Monte Carlo samples. The third is the offset from the population median; 0 represents the null hypothesis, to simulate level. The third argument could be replaced by an alternative value to give a table in parallel with Table 2.3; note, however, that this macro is simpler than the R function, in that it requires the alternative value to be the same for every distribution.

Macro evaluations do the calculations, but do not display output. Results of test1 are placed into the data set size. This data set is then printed using
the SAS procedure **proc print**. In common with most SAS procedures, data that **proc print** is to manipulate is specified to the procedure by **data=**, followed by the data set name.

### A.2 Example 2

Perform the sign test to evaluate the null hypothesis that the population median for nail arsenic levels is .26. This is done using **proc univariate**:

```sas
/* Data from http://lib.stat.cmu.edu/datasets/Arsenic */
/* reformatted into ASCII on the course home page. Data re- */
/* flect arsenic levels in toenail clippings; covariates in- */
/* clude age, sex (1=M), categorical measures of quantities */
/* used for drinking and cooking, arsenic in the water, and */
/* arsenic in the nails. To make arsenic.dat from Arsenic, do*/
/*antiword Arsenic|awk '((NR>39)&&(NR<61)){print}'>arsenic.dat*/
/* Potential threshold for ill health effects for toenails is */
data arsenic; infile '/folders/myfolders/arsenic.dat';
  input age sex drink cook water nails;
run;
proc univariate data=arsenic mu0=.26 ciquantdf(alpha=.05);
  var nails;
run;
```

Material between /* and */ are ignored by SAS, and are presented above so that information describing the data set may be included with the analysis code. The option **mu0=.26** specifies the null hypothesis, and **ciquantdf(alpha=.05)** specifies that distribution free intervals for quantiles, with confidence .95=1-.05, should be produced. The **var** statement indicates which variable should be summarized.

The hypothesis test might also have been done using **proc freq**:

```sas
data arsenic; set arsenic;y=nails>.26; run;
proc freq data=arsenic; tables y/binomial(p=.5 level=1);
  exact binomial; run;
```

### A.3 Example 3

The **proc univariate** call above gives intervals for a variety of quantiles, including quartiles, but not for arbitrary user-selected quantiles. The following
code gives arbitrary quantiles. Manually edit the code below to replace 21 with the actual number of observations, 0.025 with half the desired complement of the confidence, and .75 with the desired quantile.

```sas
proc sort data=arsenic; by nails; run;
data ci; set arsenic;
   a=quantile("binomial",.025,.75,21);
   b=21+1-quantile("binomial",.025,1-.75,21);
   if _N_=a then output; if _N_=b then output;
run;
title 'Upper quartile .95 CIs for nail arsenic';
proc print data=ci; run;
```

A.4 Example 5

Calculations for the empirical cumulative distribution function may be performed using

```sas
proc univariate data=arsenic; var nails; cdfplot nails; run;
```

and graphical output may be adjusted to your local SAS installation as necessary.

A.5 Example 6

This example calculates Mood’s median test for the subset of yarn strength data coming from bobbin 3. Note that the variable denoting type is a character variable; this is designated by `. First, construct the reduced data set. Then, use `proc npar1way` to run the test. This procedure does the bulk of nonparametric group comparison calculations available in SAS; two-group comparisons are a special case. Group membership is specified by the variable specified in the `class` statement. The `median` keyword triggers Mood’s median test.

```sas
/***********************************************************/
/* Yarn strength data from Example Q of Cox & Snell */
/* (1981). Variables represent strength of two types of */
/* yarn collected from six different bobbins. */
/***********************************************************/
data yarn; infile '/folders/myfolders/yarn.dat';
   input strength bobbin type $; run;
data yarn1; set yarn;
```
Analysis using the SAS System

if bobbin=3 then; else delete; run;
title 'Median test applied to bobbin 3 yarn data';
proc npar1way data=yarn1 median ; class type;
   exact; var strength; run;

Compare this with the t-test results.

proc ttest data=yarn1 ; class type; var strength; run;

A.6 Example 7

This example applied the Wilcoxon rank sum test to the investigating differences between the strengths of yarn types. At present, analysis is restricted to bobbin 3, since this data set did not have ties. These score tests are calculated using proc npar1way , with the option wilcoxon. The first has continuity correction turned off, with the option correct=no. Correction is on by default, and so no such option appears in the second run, with correction. Exact values are given using the exact statement, followed by the test whose exact p-values are to be calculated.

title 'Wilcoxon Rank Sum Test, yarn strength by type, bobbin 3';
title2 'Approximate values, no continuity correction';
proc npar1way data=yarn1 wilcoxon correct=no ;
   class type; var strength; run;
title2 'Approximate values, continuity correction';
proc npar1way data=yarn1 wilcoxon ; class type;
   var strength; run;
title2 'Exact values';
proc npar1way data=yarn1 wilcoxon ; class type;
   exact wilcoxon; var strength; run;

A.7 Example 8

Scores are requested as part of proc npar1way statement.

proc npar1way data=arsenic vw savage ;
   class sex; var nails; run;
A.8 Example 9
Permutation testing may be done using \texttt{proc npar1way} and \texttt{scores=data}:

\begin{verbatim}
title 'Permutation test for Arsenic Data';
proc npar1way data=arsenic scores=data ;
   class sex; exact scores=data; var nails; run;
\end{verbatim}

A.9 Table 3.4
Test sizes and Powers may be calculated using the macro \texttt{test2} in the file \texttt{common.sas}, as above. First and second arguments are the group sizes. The third argument is the Monte Carlo sample size. The last argument is the offset between groups.

\begin{verbatim}
title 'Monte Carlo Assessment of Test Sizes';
\%include "/folders/myfolders/common.sas";
*Third argument below is MC sample size. Set higher for;
  *more realistic simulation. SAS Studio gives issues.;
\%test2(10,10,1000,0);
proc print data=size noobs; run;
title 'Monte Carlo Assessment of Test Powers';
\%test2(10,10,10000,1);
proc print data=size noobs; run;
\end{verbatim}

A.10 Tables 3.4 and 3.5
Test sizes and Powers may be calculated using the macro \texttt{test2} in the file \texttt{common.sas}, as above. First and second arguments are the group sizes. The third argument is the Monte Carlo sample size. The last argument is the offset between groups.

\begin{verbatim}
title 'Monte Carlo Assessment of Test Sizes';
\%include "/folders/myfolders/common.sas";
\%test2(10,10,10000,0); proc print data=size noobs; run;
title 'Monte Carlo Assessment of Test Powers';
\%test2(10,10,10000,1); proc print data=size noobs; run;
\end{verbatim}
A.11 Example 11
The npar1way procedure also performs the Siegel-Tukey and Ansari-Bradley tests, using the keywords `ab` and `st` in the statement.

```
proc npar1way data=yarn ab st ; class type;
   var strength; run;
```

A.12 Example 12
Hodges-Lehmann estimation is performed by npar1way, using the `hl` option to the proc npar1way statement. Adding the `exact hl` first inverts the Mann-Whitney-Wilcoxon test to determine which order statistics make up the interval; otherwise, a Gaussian approximation to the Mann-Whitney-Wilcoxon test is used to get critical values, and confidence interval end points are interpolated between order statistics.

```
proc npar1way data=arsenic hl wilcoxon plots=none;
   class sex; exact hl ; var nails; run;
```

A.13 Example 13
Kolmogorov-Smirnov and Cramer Cramér-von Mises tests are also performed by proc npar1way, by specifying `edf` in the statement:

```
proc npar1way data=yarn edf; class type; var strength; run;
```

A.14 Examples 14 and 15
The Kruskal-Wallis test, and the variants using Savage and van der Waerden score test, can be performed using proc npar1way.

```
data maize; infile '/folders/myfolders/T58.1';
   input exno tabno lineno loc $ block $ plot
treatment $ ears weight;
   nitrogen=substr(treatment,1,1); if weight<0 then weight=.;
```
A.15 Example 16

Here are two calls to npar1way that give in principle the same answer. The exact command causes exact p-values to be computed. If you specify scores (in this case data) in the exact statement, you’ll get that exact p-value. If you don’t specify scores, you’ll get the exact p-value for the scores specified in the proc npar1way statement. If you don’t specify either, you’ll get ranks without scoring. As we saw before, exact computations are hard enough that SAS quits. The second proc npar1way presents compromise: random samples from the exact distribution. Give the sample size you want after /mc. The two calls give different answers, since one is approximate. Successive calls will give slightly different answers.

```plaintext
title 'Permutation test for Maize Data';
proc npar1way data=tean scores=data plots=none; class treatment;
   exact; var weight; run;
title1 'Maize Permutation test, Monte Carlo Version';
proc npar1way data=tean scores=data plots=none; class treatment;
   exact scores=data/mc n=100000; var weight; run;
```

A.16 Example 23

Here we test for an ordered effect of nitrogen, and compare with the unordered version.
title 'Jonckheere-Terpstra Test for Maize';
proc freq data=tean noprint; tables weight*nitrogen/jt;
   output out=jttab jt; run;
proc print data=jttab noobs; run;
title 'K-W Test for Maize, to compare with JT';
proc npar1way data=tean plots=none; class nitrogen;
   var weight; run;
Bibliography


Clopper, C. and E. Pearson (1934, 12). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika* 26(4), 404–413.


Bibliography


Neyman, J. and E. S. Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231, 289–337.


Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika* 23(1/2), 114–133.


Bibliography


Terpstra, T. (1952). The asymptotic normality and consistency of Kendall’s test against trend, when ties are present in one ranking. *Indagationes Mathematicae 14*, 327–333.


