Human–Computer Interaction Series

Judy Robertson
Maurits Kaptein *Editors*

Modern Statistical Methods for HCI

EXTRAS ONLINE

Springer

Judy Robertson · Maurits Kaptein
Editors

Modern Statistical Methods for HCI

Springer

Editors
Judy Robertson
Moray House School of Education
Edinburgh University
Edinburgh
UK

Maurits Kaptein
Donders Centre for Cognition
Radboud University Nijmegen
Nijmegen
The Netherlands

Additional material to this book can be downloaded from http://extras.springer.com.

ISSN 1571-5035
Human–Computer Interaction Series
ISBN 978-3-319-26631-2 ISBN 978-3-319-26633-6 (eBook)
DOI 10.1007/978-3-319-26633-6

Library of Congress Control Number: 2015958319

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer International Publishing AG Switzerland

Chapter 7
Nonparametric Statistics in Human–Computer Interaction

Jacob O. Wobbrock and Matthew Kay

Abstract Data not suitable for classic parametric statistical analyses arise frequently in human–computer interaction studies. Various nonparametric statistical procedures are appropriate and advantageous when used properly. This chapter organizes and illustrates multiple nonparametric procedures, contrasting them with their parametric counterparts. Guidance is given for when to use nonparametric analyses and how to interpret and report their results.

7.1 Introduction

The field of human–computer interaction (HCI) is diverse in many ways. Its researchers and practitioners come from all over the academic spectrum, from social sciences like psychology, sociology, and anthropology, technical endeavors like computer science, information science, and electrical engineering, and design disciplines like product design, graphic design, interaction design, and architecture. With such a wide range of backgrounds, methods, and phenomena of interest, it is no wonder that almost any kind of data may arise as part of a study in HCI.

Whether we are examining people's interactions with existing technology or evaluating new technologies that we invent, it is common to find that our data is not amenable to conventional parametric analyses. Such analyses, most commonly the familiar analysis of variance (ANOVA), are based on assumptions often violated by data arising from HCI studies. Different "nonparametric" analyses are needed to properly draw conclusions from such data.

J.O. Wobbrock (✉)
The Information School, University of Washington, Seattle, WA 98195-2840, USA
e-mail: wobbrock@uw.edu

M. Kay
Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-2350, USA
e-mail: mjskay@uw.edu

J. Robertson and M. Kaptein (eds.), *Modern Statistical Methods for HCI*,
Human–Computer Interaction Series, DOI 10.1007/978-3-319-26633-6_7

This chapter reviews nonparametric analyses, many of which are commonly found in HCI studies, and others that are more recently emerging but could be of value to HCI researchers and practitioners. For historical context, this chapter endeavors to cite the original articles where the analyses first appeared. But before plunging into the analyses themselves, let us first understand when to use nonparametric analyses.

7.2 When to Use Nonparametric Analyses

As Chap. 3 described, every data set should be explored with descriptive statistics and visual plots to ascertain the shape of its distribution. Preparing to analyze data with nonparametric procedures is no different. A lot can be learned by examining the shape of data to see whether it appears to conform to a normal distribution, also known as a Gaussian distribution or "bell curve." The concept behind most nonparametric analyses is that they do not assume a normal distribution and effectively destroy the distribution inherent in a data set by operating on ranks, rather than on the original data points themselves. Ranks destroy the intervals between values, so ascending data points like 1, 5, 125 and 1, 3, 5 both become ranks 1, 2, 3.

Certain types of measures in HCI tend to fall normally while others almost never do. Common non-normal data distributions that arise frequently in HCI studies are:

- Preference tallies, such as from studies of competing technologies
- Completion times, which may be skewed and long-tailed
- Error rates, which may have numerous zeroes amidst other values
- Ordinal scales, such as responses on Likert-type scales
- Rare events, such as recognition errors from an accurate gesture recognizer

The above examples are just some of the types of data that arise in HCI and may warrant nonparametric analyses.

7.2.1 Assumptions of Analysis of Variance (ANOVA)

The familiar analysis of variance procedure, or ANOVA, is often more powerful than analogous nonparametric procedures for the same data. The oft-used t-test and F-test are two examples. Such tests are therefore generally preferred to their nonparametric cousins. But t-tests and F-tests cannot always be used when data violates one or more of three underlying assumptions required by such analyses. HCI researchers and practitioners seeking to use ANOVA procedures should first ensure that their data conforms to the three assumptions. When violations occur, nonparametric procedures may be preferred.

The three underlying assumptions of ANOVA, and how to test for them, are:

1. *Independence.* Responses must be distinct measurements independent from one another, except as correlated in a within-subjects design. Put another way, the

value of one measure should not determine the value of any other measure. *How to test?* The independence assumption is not tested mathematically but is verified by an experiment design that ensures this assumption is met.

2. *Normality.* Residuals are the differences between model predictions and observed measures. The normality assumption requires that residuals are normally distributed. In practice, the normality assumption can be regarded as referring to the distribution of the response within each group under consideration. Mild deviations from normality often do not pose serious threats to drawing valid statistical conclusions, particularly when sample sizes are large. However, in the presence of substantial deviations from normality, especially with small sample sizes as are common in HCI, nonparametric procedures ought to be used. *How to test?* A histogram of the residuals or the data itself can often reveal obvious deviations from normality, such as data that conforms to log-normal, Poisson, or exponential data distributions. More formal tests of normality can be conducted, such as the Shapiro-Wilk test (Shapiro and Wilk 1965) or the Kolmogorov-Smirnov test (Kolmogorov 1933; Massey 1951; Smirnov 1939). R code for executing these tests is provided elsewhere in this chapter. A good review of these and other goodness-of-fit tests can be found in the literature (D'Agostino 1986).
3. *Equal variances.* The equal variances assumption is more formally known as the assumption of "homogeneity of variance" or "homoscedasticity." It requires that the variance, or equivalently the standard deviation, among different experimental groups should be about the same. *How to test?* A histogram of the data from each group being compared can reveal whether some groups have different variances than others. More formally, Levene's test can be used (Levene 1960). If homoscedasticity is violated, a Welch ANOVA or White-corrected ANOVA can be used (Welch 1951; White 1980), which do not have the equal variances assumption. An alternative is to use a nonparametric analysis, such as those covered in this chapter.

7.2.2 Table of Analogous Parametric and Nonparametric Tests

Many HCI researchers and practitioners are more familiar with parametric tests than nonparametric tests. It can be helpful to see how the two types of tests relate. By understanding the relation among parametric tests and their nonparametric equivalents, researchers and practitioners can more confidently choose which nonparametric test is right for their data.

This chapter is far from a comprehensive treatment of nonparametric statistics. There are myriad nonparametric tests that might benefit researchers and practitioners in HCI. This chapter focuses on the most common, widely available, and versatile nonparametric tests. For a complete approach to nonparametric statistics, the reader is directed to comprehensive treatments on the subject (Higgins 2004; Lehmann 2006).

Table 7.1 Parametric tests and their nonparametric cousins

Samples			Parametric test	Nonparametric test
1			• One-sample *t*-test	• One-sample chi-square test • Binomial test • Multinomial test
≥1				• *N*-sample chi-square test • *G*-test • Fisher's exact test
Factors	Levels	Between- or within-subjects	Parametric test	Nonparametric test
1	2	B	• Independent-samples *t*-test	• Median test • Mann-Whitney *U* test
1	≥2	B	• One-way ANOVA	• Kruskal-Wallis test
1	2	W	• Paired-samples *t*-test	• Sign test • Wilcoxon signed-rank test
1	≥2	W	• One-way repeated measures ANOVA	• Friedman test
≥1	≥2	B	• *N*-way ANOVA	• Aligned rank transform • Generalized linear models† –Multinomial logistic –Ordinal logistic –Poisson –Gamma
≥1	≥2	W	• *N*-way repeated measures ANOVA	• Aligned rank transform • Generalized linear mixed models† • Generalized estimating equations

†Generalized linear models and generalized linear mixed models may be considered parametric analyses but the distributions on which they operate may be non-normal

Table 7.1 categorizes the tests covered in this chapter based on the number of "factors" and their "levels." Factors are the independent variables manipulated in an experiment, such as *Device* when comparing mice to a trackballs. They may also be covariates, like *Sex*, which are not manipulated but are still of interest for their possible effect on dependent variables, or responses. Factors can be "between-subjects" or "within-subjects," owing to whether each participant is assigned only one level of the factor or more than one. Levels are the number of values any given factor can assume. For example, *Device*, above, has two levels (mouse, trackball), as does *Sex* (male, female). Note that *Device* could be a within-subjects factor if every participant utilized both devices. *Sex*, on the other hand, is generally considered only a between-subjects factor.

Besides factors and levels, another distinguishing feature of certain tests is whether they are "exact tests." Exact tests do not rely on approximations or asymptotic properties of the sampling distribution to derive p-values. Rather, they calculate their p-values directly and exactly. HCI studies often have small sample sizes, which can cause problems for asymptotic tests, and exact tests are then preferred. The Chi-Square test is a popular asymptotic test. It may underestimate p-values at less than 1000 samples, increasing the chance of falsely rejecting null hypotheses. Exact tests are currently under-utilized in HCI, largely due to conventions established before advances in computing made exact tests widely practicable. Where possible, we provide R code for running exact tests in this chapter.

Having established the criteria that are used to inform whether to use parametric or nonparametric analyses, we now turn to the analyses themselves. In each case, an (inappropriate) parametric analysis is conducted prior to any nonparametric analyses for comparisons. These parametric analyses are flagged with a ⊘ symbol to indicate caution. For continuity, the storyline about the sales teams using the new Mango smartwatches is used.

7.3 Tests of Proportions

Often in HCI studies, researchers and practitioners elicit responses from participants or users, count those responses, and then wish to draw conclusions from those counts. Responses of these kinds tend to be categorical in nature. For example, in a survey respondents may be asked to express a preference for one of a variety of technologies, like web browsers. Some number of respondents may choose Microsoft Internet Explorer, while others may choose Google Chrome, while still others may choose Mozilla Firefox or Apple Safari. One-sample tests of proportions can reveal whether responses differ significantly from chance or from known probabilities—in this case, perhaps the global market share percentage of each browser.

Going further, we may wish to know how respondents' browser preferences differ by country. We now would use a two-sample test of proportions. Known probabilities may now need to be adjusted by the market share of each browser in each country.

A three-sample test would allow us to determine whether sex plays a role. A four-sample test might include respondents' income bracket. And so on… In short, tests of proportions tell us whether observed proportions differ from chance or from otherwise hypothesized probabilities.

7.3.1 One-Sample Tests of Proportions

Let us introduce our scenario for our one-sample tests. At one point prior to the companywide adoption of Mango smartwatches, 75 sales representatives were recruited

for a pilot study in which each Mango smartwatch was outfitted with one of two email applications, A-mail or B-mail. After 3 weeks, the watches were updated to remove the first mail program and install the second. Another 3 weeks passed. At the end of 6 weeks, each sales representative had used both A-mail and B-mail. The representatives were then asked for their preference. As the data file `prefs1AB.csv` shows, A-mail was preferred by 46 sales representatives and B-mail was preferred by 29 representatives. The question is whether there was a significant preference for one email application over the other.

As stated above, for comparisons we briefly report an (inappropriate) parametric test prior to the preferred nonparametric tests.

⊘ ***One-sample t-test.*** The one-sample t-test is a simple parametric test that assumes the population response is normal (Student 1908). For our example, we can assume that if respondents showed no overall preference, $75/2 = 37.5$ respondents would vote for each email application. In other words, no overall preference would mean a 50 % chance of preferring one or the other applications. In our data, we have $46/75 = 61.3\,\%$ of respondents preferring A-mail, and $29/75 = 38.7\,\%$ preferring B-mail. Is this a statistically significant difference?

The R code for performing a one-sample t-test on `prefs1AB.csv` is:

```
> prefs1AB = read.csv("chapter7/prefs1AB.csv")
> t.test(prefs1AB$email_preference == "A-mail", mu=0.5)

      One Sample t-test

data: prefs1AB$email_preference == "A-mail"
t = 2.002, df = 74, p-value = 0.04895
alternative hypothesis: true mean is not equal to 0.5
95 percent confidence interval:
 0.5005335 0.7261332
sample estimates:
mean of x
0.6133333
```

The p-value is 0.049, which is less than the critical value of $\alpha = 0.05$, meaning the 46 votes for A-mail do represent a statistically significant preference for A-mail over B-mail. The test result is reported as $t(74) = 2.00$, $p < 0.05$.

One-sample Chi-Square test. The most common way to test proportions is using a one-sample Chi-Square test (Pearson 1900), which is a nonparametric alternative to the one-sample t-test. Unlike the t-test, the Chi-Square test does not require that the data be sampled from a normal distribution. However, it is an asymptotic test, not an exact test, so for small sample sizes it must be used with caution. The premise behind the test is the same as before, where we compare observed proportions to chance, i.e., to 50/50.

The R code for performing a one-sample Chi-Square test on `prefs1AB.csv` is:

```
# assuming prefs1AB.csv is already loaded
# chisq.test expects frequency tables as input: here we
# create a cross tabulation (hence xtabs) of the number of
# responses for each level of email_preference
> email_preferences = xtabs( ~ email_preference, data=prefs1AB)
> email_preferences
email_preference
A-mail B-mail
    46     29
> chisq.test(email_preferences)

    Chi-squared test for given probabilities

data: email_preferences
X-squared = 3.8533, df = 1, p-value = 0.04965
```

The truncated p-value is 0.049, again indicating that the 46 votes for A-mail represent a statistically significant preference over B-mail. The test result is reported as $\chi^2(1, N = 75) = 3.85$, $p < 0.05$, where 1 is the value of `df` above.

Binomial test. The binomial test is a nonparametric test used to compare two categories against expected probabilities, often called the probability of "success" or "failure." Unlike the Chi-Square test, which relies on approximations, the binomial test is an exact test. A common use of the binomial test is to see whether responses in two categories are equally likely to occur, such as testing whether a coin is fair from a series of tosses. We can use the binomial test to see whether the probability of someone preferring one of the email programs is significantly different from chance (i.e., 50 %).

The R code for performing a binomial test on `prefs1AB.csv` is:

```
# assuming prefs1AB.csv is already loaded
> email_preferences = xtabs( ~ email_preference, data=prefs1AB)
> binom.test(email_preferences)

    Exact binomial test

data:  email_preferences
number of successes = 46, number of trials = 75,
p-value = 0.06395
alternative hypothesis: true probability of success is not
equal to 0.5
95 percent confidence interval:
 0.4937958 0.7236319
sample estimates:
probability of success
             0.6133333
```

The p-value is 0.064, greater than the critical value of $\alpha = 0.05$, meaning we fail to reject the null hypothesis that the two categories are equally probable. In this

case, the one-sample Chi-Square test underestimated the p-value: we rejected the null hypothesis when using the asymptotic Chi-Square test and failed to reject when using the exact binomial test. Since modern computers allow us to easily run an exact test in this case, we should prefer the exact result to the Chi-Square result. The test result is reported as a binomial test of $N = 75$ responses, a 50/50 hypothesized response probability, and a p-value of 0.0640.

Multinomial test. What if there are more than two response categories? The binomial test cannot be used. In such cases, the nonparametric multinomial test is appropriate. Like the binomial test, the multinomial test is an exact test, and should be preferred to the Chi-Square test.

Suppose the original study had compared three email programs instead of two. Preferences were elicited from the 75 sales representatives. The data in `prefs1ABC.csv` indicates that 35 respondents preferred A-mail, 22 preferred B-mail, and 18 preferred C-mail. The question is whether these counts differ significantly from chance, i.e., a third of respondents in each category. The code for performing a multinomial test on `prefs1ABC.csv` is:

```
> library(XNomial)
> prefs1ABC = read.csv("chapter7/prefs1ABC.csv")
> email_preferences = xtabs(~ email_preference, data=prefs1ABC)
> xmulti(email_preferences, c(1/3, 1/3, 1/3), statName="Prob")

P value (Prob) = 0.04748
```

The p-value is 0.047, indicating that we should reject the null hypothesis that each email program is preferred equally. The test result is reported as a multinomial test of $N = 75$ responses, equal chance hypothesized response probability (i.e., a 1/3 chance of each email application being preferred), and a p-value less than 0.05.

The one-sample Chi-Square test we utilized above can also accommodate more than two response categories like the multinomial test. The following R code runs a one-sample Chi-Square test on `prefs1ABC.csv`:

```
# assuming prefs1ABC.csv is already loaded
> email_preferences = xtabs(~ email_preference, data=prefs1ABC)
> chisq.test(email_preferences)

     Chi-squared test for given probabilities

data:  email_preferences
X-squared = 6.32, df = 2, p-value = 0.04243
```

The p-value is 0.042. According to this Chi-Square test, there is a statistically significant difference between the observed preferences and chance. Observing the preference counts, we surmise a significant preference for A-mail over the other

two mail programs. Note that the Chi-Square test underestimates the p-value compared to the exact multinomial test. The multinomial test should be preferred where computationally feasible, typically for sample sizes of less than 1000.

7.3.2 N-Sample Tests of Proportions

Regardless of how many response categories there are, if only one dimension of data is considered, a multinomial test or a one-sample Chi-Square test is an option. But what if we wish to categorize responses along more than one dimension? Consider the question of whether preferences for the Mango email applications—A-mail, B-mail, or C-mail—were different in Sales Team X versus Sales Team Y. One dimension lies along the sales teams, with two possible categories. A second dimension lies along the email applications, with three possible categories. Thus, we have a 2×3 contingency table, also known as a "crosstabulation" or "crosstabs".

Data appearing in `prefs2ABC.csv` contains 75 responses from each of two sales teams. Its 2×3 contingency table is shown in Table 7.2.

The question is whether the email application preferences of the two sales teams differ significantly.

N-Sample Chi-Square test. We have thus far been generating 1×2 and 1×3 contingency tables using the `xtabs` command. This command can generate crosstabs with an arbitrary number of dimensions, making N-Sample Chi-Square tests a simple extension of the procedures we have already employed above.

The R code for running a two-sample Chi-Square test of proportions is:

```
> prefs2ABC = read.csv("chapter7/prefs2ABC.csv")
# we specify multiple factors in the xtabs formula to get
# crosstabs of higher dimensions.
> email_preferences = xtabs( ~ email_preference + team,
                             data=prefs2ABC)
> chisq.test(email_preferences)
   Pearson's Chi-squared test
data: email_preferences
X-squared = 6.4919, df = 2, p-value = 0.03893
```

Table 7.2 A 2×3 contingency table of email application preferences by 75 members each of Sales Team X and Sales Team Y

		Email application preference			
		A-mail	B-mail	C-mail	Total
Sales team	X	35	22	18	75
	Y	21	35	19	75
	Total	56	57	37	**150**

The p-value is 0.039, indicating a significant difference in email application preferences between the two sales teams. The test result is reported as $\chi^2(2, N = 150) = 6.49$, $p < 0.05$.

Other N-sample tests. Here we highlight two alternative tests of proportions that use a similar syntax as `chisq.test`. One test that is gaining popularity is the *G*-test (Sokal and Rohlf 1981), which, although an asymptotic test, is considered more accurate than the Chi-Square test, which employs approximations where the *G*-test directly computes likelihood ratios. The test is conducted in R using the `G.test` function in the `RVAideMemoire` package.

Another popular test is Fisher's exact test (Fisher 1922), which is an exact test used primarily on 2×2 contingency tables but is capable of being extended to general $r \times c$ tables provided sufficient computational resources (Mehta and Patel 1983). The test is conducted in R using the `fisher.test` function, and is natively capable of handling general $r \times c$ contingency tables.

7.4 Single Factor Tests

In the previous section, we discussed tests for data that counted respondents—and more precisely, their preferences—as measures of interest. In the rest of this chapter, we consider the results of experimental designs in which people are assigned treatments and the measures of interest involve the behavior or attributes of those people under those treatments. We first consider single-factor tests. Before doing so, however, we introduce statistical tests for the assumptions of ANOVA, which may be used to determine whether nonparametric tests are warranted in the first place.

7.4.1 Testing ANOVA Assumptions

Recall from the outset of this chapter that the three assumptions of ANOVA are independence, normality, and equal variances. Here we examine these assumptions for the `salesXY.csv` data file, which arises from the following scenario. Let us assume members of one company sales team, Sales Team X, were given Mango smartwatches over a three-month period. During the same period, members of another company sales team, Sales Team Y, were not given the smartwatches so as to serve as a control group. Each representative's sales were measured during the three-month period and multiplied by four to reflect estimated annualized sales. Thus, we have a single between-subjects factor, *Team*, and a continuous measure for each participant, annualized sales, in dollars.

The first assumption is independence. Do the measures arise independent of one another? Assuming the salespeople work independently and that there are many more sales opportunities than sales representatives, and thus one salesperson's gain is not

inherently another salesperson's loss, we can trust that the independence assumption is met.

The second assumption is normality. Before conducting formal tests of normality, let us visually examine the distribution of data with a histogram, one for each sales team (Fig. 7.1).

The histograms are clearly non-normal in their appearance. As with most measures of income, the annualized sales data conforms to an exponential distribution. This hunch can be formally tested with goodness-of-fit tests. We briefly review two popular goodness-of-fit tests here. We recommend Shapiro-Wilk for testing normality; the Kolmogorov-Smirnov test can be used to test the goodness-of-fit of non-normal distributions. For more on goodness-of-fit tests, the reader is directed elsewhere (D'Agostino 1986).

Shapiro-Wilk test. The Shapiro-Wilk test considers whether data from a sample originated from a normal distribution (Shapiro and Wilk 1965). It has been shown to have the best power of the three tests considered here (Razali and Wah 2011).

The R code for conducting a Shapiro-Wilk test on each team in `salesXY.csv` is:

```
> salesXY = read.csv("chapter7/salesXY.csv")
> shapiro.test(salesXY[salesXY$team == "X",]$sales)
   Shapiro-Wilk normality test
data: salesXY[salesXY$team == "X", ]$sales
W = 0.8368, p-value = 1.238e-07
> shapiro.test(salesXY[salesXY$team == "Y",]$sales)
   Shapiro-Wilk normality test
data: salesXY[salesXY$team == "Y", ]$sales
W = 0.791, p-value = 6.013e-09
```

The p-value for both teams is $p < 0.0001$, indicating a statistically significant difference between the distribution of their annualized sales and a normal distribution.

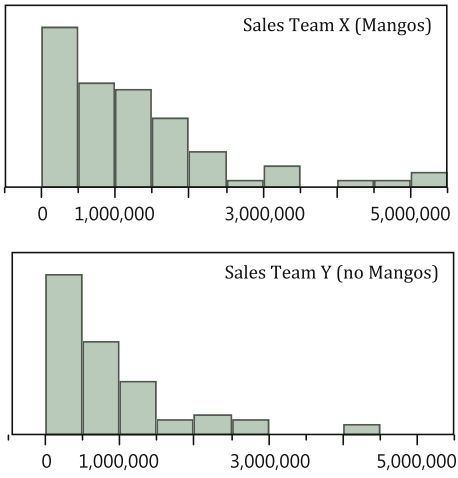

Fig. 7.1 The distribution of annualized sales for each sales team. Team X had Mango smartwatches. Team Y did not yet have the Mango smartwatches. Histograms can be generated using the `hist` function, as in `hist(salesXY[salesXY$team=="X",]$sales)`

Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test considers how a data sample compares to a given probability distribution (Kolmogorov 1933; Massey 1951; Smirnov 1939). It calculates the distance between the empirical distribution function of a sample and the cumulative distribution function of the given probability distribution. Thus, the Kolmogorov-Smirnov test can be used to test against non-normal distributions.

The R code for executing the Kolmogorov-Smirnov test is:

```
# assuming salesXY.csv is already loaded
> lillie.test(salesXY[salesXY$team == "X",]$sales)

    Lilliefors (Kolmogorov-Smirnov) normality test

data:   salesXY[salesXY$team == "X", ]$sales
D = 0.1445, p-value = 0.0005212
> lillie.test(salesXY[salesXY$team == "Y",]$sales)

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  salesXY[salesXY$team == "Y", ]$sales
D = 0.1791, p-value = 2.904e-06
```

The p-values for the teams are $p < 0.001$ and $p < 0.0001$, again indicating statistically significant departures from normality. Thus, from visual inspection and from both formal goodness-of-fit tests, we conclude the data violates the normality assumption of ANOVA.

The third assumption of ANOVA is equal variances. The standard deviation of annualized sales for Sales Team X is $1,169,590.80. For Sales Team Y, it is $904,175.91. Of course, with highly non-normal distributions, standard deviations are not particularly descriptive. So how might we formally test the assumption of equal variances?

Levene's test. Levene's test for homogeneity of variance, or homoscedasticity, is a formal method for testing the equal variances assumption (Levene 1960). The test determines the likelihood of whether two data samples are drawn from populations with equal variance. A significant p-value below the $\alpha = 0.05$ level indicates that the data samples being tested are unlikely to have come from populations with equal variances.

We conduct Levene's test on two data samples, the annualized sales of Sales Team X and of Sales Team Y. The R code for conducting Levene's test is:

```
# assuming salesXY.csv is already loaded
> library(car)
> leveneTest(sales ~ team, data=salesXY)
Levene's Test for Homogeneity of Variance (center = median)
       Df  F value   Pr(>F)
group   1   2.9567  0.08761  .
      148
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 0.088, above the $\alpha = 0.05$ threshold for declaring that the equal variances assumption has been violated. Even still, with a trend-level result such as this one, we ought to be wary of utilizing parametric tests. The result of Levene's test is reported as $F(1,148) = 2.96$, $p = 0.088$.

We have heretofore demonstrated that the data in `salesXY.csv` is not suitable to analyse via parametric ANOVA. Nonparametric tests are therefore warranted. Let us now turn to those tests.

7.4.2 Single-Factor Between-Subjects Tests

We continue with our scenario comparing the annualized sales of two sales teams, Sales Team X wearing Mango smartwatches and Sales Team Y without such watches. As above, for comparisons we briefly report an (inappropriate) parametric test prior to the preferred nonparametric options.

⊘ ***Independent-samples t-test.*** The independent-samples t-test is a parametric test for one-factor two-level between-subjects designs (Student 1908). Due to the violation of normality, the test is inappropriate for the data in `salesXY.csv`. Nevertheless, the R code for executing such an analysis is shown below. Note that by default, R uses the Welch t-test (Welch 1951), which does not require equal variances, having been formulated for this purpose.

```
# assuming salesXY.csv is already loaded
> t.test(sales ~ team, data=salesXY)
    Welch Two Sample t-test
data: sales by team
t = 2.2293, df = 139.173, p-value = 0.02739
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
  43049.83 718064.85
sample estimates:
mean in group X mean in group Y
      1250090.3        869532.9
```

The p-value is 0.027, which indicates that the annualized sales of the two teams are significantly different. Specifically, the sales of Sales Team X are higher than those of Sales Team Y, suggesting that the Mango smartwatches are having a positive effect. The test result is reported as $t(139.2) = 2.23$, $p < 0.05$.

Median test. A more appropriate test is the nonparametric median test (Brown and Mood 1948, 1951). The median test considers whether the medians from the populations from which two data samples are drawn are the same. A simple test, the median test counts each data point as above or below the median in the combined sample. Traditionally, then a Chi-Square test—although we can also use an exact test—is used to see whether the counts of data points from each sample differ. The median test is the preferred choice if any data points are extreme outliers.

The R code for conducting a median test is:

```
# assuming salesXY.csv is already loaded
# the distribution="exact" parameter specifies the exact
# version of this test, and can be dropped if an
# asymptotic test is needed (e.g., if this code
# takes too long to execute).
> library(coin)
> median_test(sales ~ team, data=salesXY, distribution="exact")
     Exact Median Test
data: sales by team (X, Y)
Z = -3.0923, p-value = 0.003157
alternative hypothesis: true mu is not equal to 0
```

The p-value is 0.003, indicating a significant difference in sales between the teams. The test result is reported as an exact median test $Z = -3.09$, $p < 0.01$.

Mann-Whitney U test. Like the median test, the nonparametric Mann-Whitney U test[1] operates on one-factor two-level between-subjects designs (Mann and Whitney 1947). It is more common in the field of HCI than the median test and usually more powerful. The test converts data to ranks and is generally more powerful than the parametric t-test for non-normal data.

The R code for conducting the test is:

```
# assuming salesXY.csv is already loaded
> library(coin)
> wilcox_test(sales ~ team, data=salesXY, distribution="exact")
    Exact Wilcoxon Mann-Whitney Rank Sum Test
data: sales by team (X, Y)
Z = 2.2346, p-value = 0.02521
alternative hypothesis: true mu is not equal to 0
```

[1] The Mann-Whitney U test has multiple and sometimes confusing names. It is also known as the Wilcoxon-Mann-Whitney test, the Mann-Whitney-Wilcoxon test, and the Wilcoxon rank-sum test. None of these should be confused with the Wilcoxon signed-rank test, which is for one-factor two-level *within*-subjects designs.

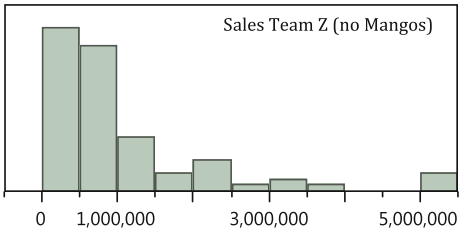

Fig. 7.2 The distribution of annualized sales for Sales Team Z

The p-value is 0.025, similar to that of the independent-samples t-test. The test result is reported as an exact Mann-Whitney $Z = 2.23$, $p < 0.05$.

The Mann-Whitney U test is a good option for analyzing one-factor two-level between-subjects designs. But what if we have one factor with more than two levels? For example, let us consider 75 additional sales representatives, this time from Sales Team Z, added to the study. Like Sales Team Y, Sales Team Z was not given Mango watches at this time. The distribution of the team's annualized sales is shown in Fig. 7.2.

We now have a one-factor three-level between-subjects design. The factor is *Team* and the levels are X, Y, and Z. We will use `salesXYZ.csv`, which extends the data table to include the data from Sales Team Z.

⊘ ***One-way ANOVA.*** The popular parametric analysis for one factor with more than two levels is a one-way ANOVA (Fisher 1921, 1925). As with the t-test above, this analysis is inappropriate for these data due to the violation of the normality assumption.

The R code for conducting a one-way ANOVA is:

```
> salesXYZ = read.csv("chapter7/salesXYZ.csv")
> summary(aov(sales ~ team, data=salesXYZ))
             Df        Sum Sq       Mean Sq      F value     Pr(>F)
team          2     5.438e+12     2.719e+12        2.345   0.0982 .
Residuals   222     2.574e+14     1.159e+12
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 0.098, which is not statistically significant at the $\alpha = 0.05$ level. The test result is reported as F(2,222) = 2.35, $p = 0.098$.

Kruskal-Wallis test. The nonparametric Kruskal-Wallis test extends the Mann-Whitney U test to one factor with more than two levels (Kruskal and Wallis 1952). Like the Mann-Whitney U test, the Kruskal-Wallis test operates on ranks. It is a more appropriate test to conduct on salesXYZ.csv than a one-way ANOVA.

The R code for executing a Kruskal-Wallis test is:

```
# assuming salesXYZ.csv is already loaded
library(coin)
kruskal_test(sales ~ team, data=salesXYZ,
             distribution="asymptotic")

Asymptotic Kruskal-Wallis Test

data:  sales by team (X, Y, Z)
chi-squared = 5.1486, df = 2, p-value = 0.07621
```

The p-value is 0.076, indicating no significant differences among groups. The test result is reported as a Kruskal-Wallis test $\chi^2(2, N = 225) = 5.15$, $p = 0.076$.

Many studies in HCI utilize designs in which multiple responses are received from each participant. We now turn to nonparametric tests for one-factor within-subjects designs.

7.4.3 Single-Factor Within-Subjects Tests

In an effort to collect more data per participant, and generally to increase the power of statistical tests on that data, many HCI researchers and practitioners prefer to utilize within-subjects factors rather than between-subjects factors. Within-subjects factors expose participants to more than one of their levels, for example, by having each sales representative not use *and* use a Mango smartwatch over different time periods. Such was the case for Sales Team Y, which initially was not given the Mango smartwatches to serve as a control for Sales Team X. After three months, Sales Team Y was given the watches, thereby enabling within-subjects comparisons for Sales Team Y.

The data in `salesYY.csv` contains the same pre-Mango sales data for Sales Team Y as shown in Fig. 7.1. It also contains post-Mango sales data for each representative, shown in Fig. 7.3. Thus, we have a single within-subjects factor, *Watch*, and a continuous measure for each participant: their annualized sales, in dollars.

As before, we begin with an (inappropriate) parametric test for comparisons.

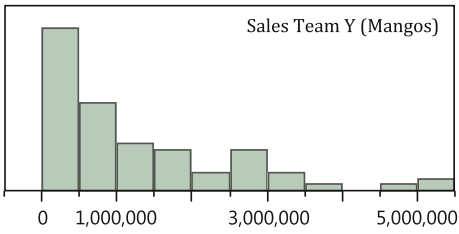

Fig. 7.3 The distribution of annualized sales for Sales Team Y after the adoption of the Mango smartwatches

⊘ ***Paired-samples t-test.*** A paired-samples t-test is a parametric within-subjects test when two measures are taken from each participant (Student 1908). Due to the violation of normality, the test is inappropriate for the data in `salesYY.csv`. Nevertheless, the R code for executing such an analysis is:

```
> salesYY = read.csv("chapter7/salesYY.csv")
> library(reshape2) #for dcast
# for a paired t-test we must use a wide-format table. Most
# functions in R do not require a wide-format table, but the
# dcast function offers a quick way to translate long-format
# into wide-format when we do need it. This creates "pre"
# and "post" columns containing pre- and post-watch sales.
> salesYY_wide = dcast(salesYY, subject ~ watch,
                       value.var="sales")
> t.test(salesYY_wide$pre, salesYY_wide$post, paired=TRUE)
    Paired t-test
data: salesYY_wide$pre and salesYY_wide$post
t = -2.4309, df = 74, p-value = 0.01748
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -732056.10  -72552.05
sample estimates:
mean of the differences
              -402304.1
```

The p-value is 0.017, which indicates that after the adoption of the Mango smartwatches, the annualized sales of Sales Team Y were different. The t-test result is reported as $t(74) = -2.43$, $p < 0.05$. By examining the means and distributions, we can see that the sales went up, from a median of about $580,000 before the watches to about $870,000 after the watches:

```
# generate summary statistics for the sales, split
# into groups according to the levels of watch
> ddply(salesYY, ~ watch, function(data) summary(data$sales))
   watch    Min.   1st Qu.  Median      Mean   3rd Qu.     Max.
1   post   10000    342500  870300   1272000   1911000  5000000
2    pre   10000    239800  580300    869500   1091000  4351000
```

Sign test. The sign test is a nonparametric alternative to the paired-samples t-test (Dixon and Mood 1946; Stewart 1941). It is analogous to the median test but for paired data rather than unpaired data. The test is particularly useful when paired values do not have scalar magnitudes but simply a greater-than or less-than relationship, even coded as just 1 or 0.

The intuition behind the sign test is if the paired samples are not significantly different, then subtracting one value from its paired value should result in a positive number (vs. a negative number) about 50 % of the time. A binomial test is then used to test for significant departures from an equal number of positive versus negative differences.

The R code for conducting a sign test on `salesYY.csv` is:

```
# assuming salesYY_wide was constructed as above
# We can conduct a sign test simply by cross-tablulating the
# number of times post-watch sales are greater than pre-watch.
> post_sales_greater = xtabs( ~ post > pre, data=salesYY_wide)
> binom.test(post_sales_greater)

   Exact binomial test

data:  post_sales_greater
number of successes = 31, number of trials = 75, p-value =
0.1654
alternative hypothesis: true probability of success is not
equal to 0.5
95 percent confidence interval:
 0.3007536 0.5329729
sample estimates:
probability of success
             0.4133333
```

The p-value is 0.165, indicating that the Mango smartwatches did not statistically significantly affect the probability an individual team member's annualized sales increased on Sales Team Y. The test result is reported as a sign test of $N = 75$ paired observations and $p = 0.165$.

Recognizing the relative statistical weakness of the sign test, Wilcoxon developed a more powerful test, the signed-rank test, which considers not just direction of paired differences but their magnitude as well.

Wilcoxon signed-rank test. The Wilcoxon signed-rank test, not to be confused with the Wilcoxon rank-sum test (see Footnote 1), is a powerful and widely used nonparametric test for one within-subjects factor with two levels (Wilcoxon 1945). In HCI studies, the test is often used when individual participants try each of two alternatives, say input devices or webpage designs, and the best alternative is to be determined. Like many nonparametric tests, it operates on ranks rather than on raw observations.

The R code for executing a Wilcoxon signed-rank test is:

```
# assuming salesYY.csv is already loaded
> library(coin)
# here we specify the response variable (sales), the within-
# subjects variable (watch), and the variable identifying each
# subject (subject).
> wilcoxsign_test(sales ~ watch | subject, data=salesYY,
                  dist="exact")

   Exact Wilcoxon-Signed-Rank Test

data:  y by x (neg, pos)
    stratified by block
Z = -2.2178, p-value = 0.02615
alternative hypothesis: true mu is not equal to 0
```

The p-value is 0.026, indicating a statistically significant difference in annualized sales for Sales Team Y pre- and post-adoption of the Mango smartwatches. The test result is reported as an exact Wilcoxon $Z = -2.22$, $p < 0.05$. We note that whereas the sign test did not find a statistically significant difference, the Wilcoxon signed-rank test did, confirming the greater power of this test and a reason for its preference.

The Wilcoxon signed-rank test is powerful but limited in an important way: it can only compare two levels of a single factor. What if there are more than two levels to be compared at once? Let us imagine that the company, pleased with the increase in annualized sales due to the Mango smartwatches, decided to have Sales Team Y conduct a third three-month experiment in which sales representatives would wear *two* Mango smartwatches, one on each wrist. (Perhaps in the hope that if one smartwatch is good, two might be better!)

The distribution of annualized sales is shown in Fig. 7.4.

The data for two watches represented above are captured in `salesYY2.csv`. They are accompanied by the data for Sales Team Y for no watch (Fig. 7.1) and one watch (Fig. 7.3). We therefore have one factor, *Watch*, now with three levels: none, one watch, and two watches.

As above, we begin with an (inappropriate) parametric test for comparisons.

⊘ ***One-way repeated measures ANOVA.*** The parametric repeated measures ANOVA can be used when multiple measures are taken from the same participant. It

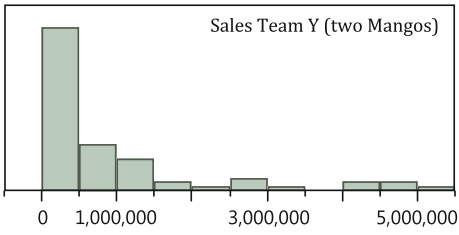

Fig. 7.4 The distribution of annualized sales for Sales Team Y when each sales representative wore two Mango smartwatches, one on each wrist

is important to use a corrected test such as a Greenhouse-Geisser correction (Greenhouse and Geisser 1959) in the event that sphericity is violated. Sphericity is a property of the data related to the covariance among experimental groups that can be tested with Mauchly's test of sphericity (Mauchly 1940). When sphericity is not violated, an uncorrected test can be used.

The R code for conducting Mauchly's test of sphericity and the ensuing repeated measures ANOVA is:

```
> salesYY2 = read.csv("chapter7/salesYY2.csv")
> library(ez)
# here we specify the dependent variable (sales), within-
# subjects variables (watch), and the variable that
# identifies subjects (subject).
> m = ezANOVA(dv=sales, within=watch,wid=subject,data=salesYY2)
# we then check the model for violations of sphericity
> m$Mauchly
  Effect             W             p p<.05
2  watch     0.9518013     0.1647936
# and given no violations, examine the uncorrected ANOVA. If
# violations were found, we would instead look at m$Sphericity.
> m$ANOVA
  Effect   DFn   DFd          F             p p<.05        ges
2  watch     2   148   2.973636   0.05418002       0.02593512
```

Mauchly's test of sphericity gives $W = 0.952$, $p = 0.165$, indicating that sphericity is not violated and that an uncorrected test can be used. The repeated measures ANOVA gives F(2,148) = 2.97, $p = 0.054$, falling just shy of statistical significance. Of course, given the violations of normality, we know this result to be specious. A nonparametric test should be used instead.

Friedman test. The nonparametric Friedman test is a rank-based test for a single within-subjects factor of any number of levels (Friedman 1937). The test is particularly useful in HCI studies where participants work with more than two variations of a user interface. For our example, we will use it to compare the annualized sales of Sales Team Y when its representatives wore zero, one, and two Mango smartwatches.

The R code for initiating a Friedman test on `salesYY2.csv` is:

```
# assuming salesYY2.csv is already loaded
> library(coin)
> friedman_test(sales ~ watch | subject, data=salesYY2)

      Asymptotic Friedman Test

data:   sales by
      watch (none, one, two)
      stratified by subject
chi-squared = 6.9067, df = 2, p-value = 0.03164
```

Table 7.3 Pairwise comparisons of annualized sales among the three levels of *Watch* using Holm's sequential Bonferroni procedure

	No watch	One watch	Two watches		
Median Sales	$580,320.00	$870,290.00	$337,919.00		
Comparison	Wilcoxon W		p-value	Holm's α	Significant?
One watch versus two watches	457.0		0.0153	$\alpha/3 = 0.0167$	Yes
No watch versus one watch	420.0		0.0262	$\alpha/2 = 0.0250$	No
No watch versus two watches	190.0		0.3189	$\alpha/1 = 0.0500$	No

The p-value is 0.032, indicating a statistically significant difference in annualized sales among the levels of *Watch*. The test result is reported as $\chi^2(2, N = 75) = 6.91$, $p < 0.05$.

With three levels of *Watch*, we may wish to know which pairwise comparisons are significant. Three pairwise comparisons may be conducted, but we must be careful to apply a correction to avoid inflating the Type I error rate—the possibility of false positives. A correction such as Holm's sequential Bonferroni procedure can avoid inflating the Type I error rate (Holm 1979).[2] We conduct the pairwise comparisons using Wilcoxon signed-rank tests, the results of which are shown in Table 7.3.

We can conduct all pairwise comparisons in R and use the `p.adjust` function to apply Holm's sequential Bonferroni procedure follows:

```
# assuming salesYY2.csv is already loaded
> library(plyr)
# get all pairwise combinations of levels of the watch factor,
# equivalent to combn(levels(salesYY2$watch),2,simplify=FALSE).
comparisons = list(c("none", "one"), c("none", "two"),
c("one", "two"))
# run wilcoxon signed-ranks on each pair of levels, collecting
# the test statistic and the p-value into a single table.
> post_hoc_tests = ldply(comparisons, function(watch_levels){
    wt = wilcoxsign_test(sales ~ factor(watch) | subject,
        data=salesYY2[salesYY2$watch %in% watch_levels,],
        dist="exact")
    data.frame(comparison = paste(watch_levels,collapse=" - "),
        z = statistic(wt), pvalue = pvalue(wt)
    )
})
```

[2]Holm's sequential Bonferroni procedure for three pairwise comparisons uses a significance threshold of $\alpha = 0.05/3$ for the lowest p-value, $\alpha = 0.05/2$ for the second lowest p-value, and $\alpha = 0.05/1$ for the highest p-value. Should a p-value compared in that ascending order fail to be statistically significant, the procedure halts and any subsequent comparisons are regarded as statistically non-significant.

```
# derive adjusted p-values using Holm's sequential Bonferroni
# procedure
> post_hoc_tests$adjusted_pvalue =
                p.adjust(post_hoc_tests$pvalue, method="holm")
> post_hoc_tests
  comparison            z         pvalue      adjusted_pvalue
1 none - one     2.217834     0.02615427           0.05230854
2 none - two    -1.003306     0.31893448           0.31893448
3  one - two    -2.413214     0.01532255           0.04596764
```

Corrected pairwise comparisons show that one Mango smartwatch produced different sales than two Mango smartwatches. Looking at median sales, it is clear that two smartwatches *hindered* sales compared to one smartwatch. Perhaps information overload had a deleterious effect on sales representatives' productivity!

We have thus far considered nonparametric tests of proportions and nonparametric single-factor tests with two or more levels. Our final consideration in this chapter is nonparametric multifactor tests—those used when more than one factor is being tested in the same experimental design.

7.4.4 Multifactor Tests

Modern experiments in HCI often involve more than one factor. Multifactor experimental designs examine more than one factor simultaneously. Each factor may have two or more levels. Chief among statistical concerns are tests for "interactions," wherein levels of one factor interact with levels of another factor to differentially affect responses. For example, perhaps one Mango smartwatch email application creates higher sales for Sales Team X, while a different email application creates higher sales for Sales Team Y. This situation would result in a statistically significant *Team* × *Application* interaction.

Nonparametric statistical methods for multifactor designs can be quite complex and are a topic of active statistical research (Sawilowsky 1990). This chapter offers a pragmatic but cursory review of four techniques: the Aligned Rank Transform, Generalized Linear Models, Generalized Linear Mixed Models, and Generalized Estimating Equations. For full treatments, the reader is directed to books on nonparametric statistics (Higgins 2004; Lehmann 2006).

7.4.5 ⊘ N-Way Analysis of Variance

As above, we begin with an (inappropriate) parametric analysis for comparisons. Let us reuse the data from Sales Team Y with zero, one, and two Mango smartwatches, but now embellished with the city in which each sales representative operated: Babol,

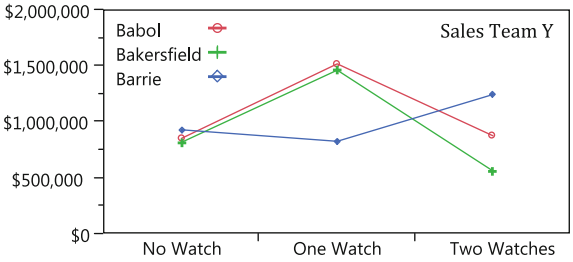

Fig. 7.5 Annualized sales for Sales Team Y by *Watch* and by *City*. A significant *Watch* × *City* interaction is suggested by this graph. A similar graph can be generated by the R code `with(salesYY2city, interaction.plot(watch,city, sales))`

Bakersfield, or Barrie. Thus, *City* is a three-level between-subjects factor, as each sales representative worked only in one city. As before, *Watch* is a three-level within-subjects factor. We therefore have a two-factor mixed design. These data are shown in `salesYY2city.csv`.

The R code for executing a two-way factorial ANOVA with one within-subjects factor, *Watch*, and one between-subjects factor, *City*, is:

```
> salesYY2city = read.csv("chapter7/salesYY2city.csv")
> library(ez)
> m = ezANOVA(dv=sales, between=city, within=watch,
              wid=subject, data=salesYY2city)
> m$Mauchly
      Effect         W         p p<.05
3      watch 0.9619982 0.2527474
4 city:watch 0.9619982 0.2527474
> m$ANOVA
      Effect  DFn  DFd         F           p p<.05        ges
2       city    2   72 0.2622249   0.77006952       0.00255353
3      watch    2  144 3.1014944   0.04800009   *   0.02717733
4 city:watch    4  144 2.5909015   0.03915842   *   0.04459346
```

Although *City* alone was not a statistically significant main effect, there was a significant *Watch* × *City* interaction as indicated by p-value of 0.039, meaning each level of *Watch* resulted in different annualized sales depending on the city in which the sales representative worked. Although a formal analysis would use pairwise comparisons to draw conclusions, for our purposes we simply eyeball the graph shown in Fig. 7.5. The graph suggests that although sales in all three cities were similar without a Mango smartwatch, with one watch, sales in Babol and Bakersfield improved, but not in Barrie. But with two Mango smartwatches, sales in Barrie improved, but were worse in Babol and Bakersfield. (Perhaps the Barrie salespeople

were all ambidextrous!) Differential results like these are what cause the statistically significant *Watch* × *City* interaction.

Given the known violations of normality, the above parametric ANOVA is an inappropriate analysis. We now turn to multifactor nonparametric procedures.

7.4.6 Aligned Rank Transform (ART)

Rank transforms have been utilized extensively in nonparametric procedures (Conover and Iman 1981). However, ANOVAs applied to rank transforms are known to drastically inflate Type I error rates for interaction effects (the chance of declaring a significant interaction effect when there is not one), and therefore are not suitable as multifactor analyses (Higgins and Tashtoush 1994; Mansouri 1999b; Salter and Fawcett 1993).

One rank-based procedure that avoids this problem is called the *Aligned Rank Transform (ART)*. The nonparametric ART procedure originated in the 1980s (Fawcett and Salter 1984; Salter and Fawcett 1985) and has been a subject of attention ever since (Higgins et al. 1990; Higgins and Tashtoush 1994; Mansouri 1999a; Mansouri et al. 2004; Richter 1999; Salter and Fawcett 1993). Before ranking, the ART procedure "aligns" the data separately for each effect by subtracting estimates of all effects other than the effect of interest from each response (Hodges and Lehmann 1962). The idea is to "strip out" any effects except one from the data. The aligned data is then ranked and a factorial ANOVA is performed with the aligned ranks as the response. Importantly, only the effect for which the responses were aligned is examined in the effects table; others are ignored. Thus, a separate aligning-ranking-ANOVA process is conducted for every effect of interest, whether a main effect or interaction. As aligning and ranking for every effect is tedious, tools have been developed to automate the process (Wobbrock et al. 2011).

The authors of this chapter have created an R package for performing the ART procedure called `ARTool`, based on a prior tool for Microsoft Windows of the same name (Wobbrock et al. 2011). The package performs the aligning-and-ranking process and automates running an ANOVA for each main effect and interaction.[3] Using this package, the ART procedure is run on the data in `salesYY2city.csv` using the following code:

[3] Rather than using traditional repeated measures ANOVAs, `ARTool` uses mixed-effects analyses of variance, explained below in the section on Generalized Linear Mixed Models.

```
# assuming salesYY2city.csv is already loaded
> library(ARTool)
> m = art(sales ~ watch * city + (1|subject),data=salesYY2city)
> anova(m)
Aligned Rank Transform Anova Table (Type III tests)

Response: art(sales)

                F  Df    Df.res     Pr(>F)
watch      4.3132   2       144    0.01516 *
city       0.0165   2        72    0.98361
watch:city 2.0173   4       144    0.09510 .
---
Signif. codes: 0'***' 0.001'**' 0.01'*' 0.05 '.' 0.1'' 1
```

As with the parametric two-way ANOVA above, we have a statistically significant effect of *Watch* on annualized sales $F(2,144) = 4.31$, $p < 0.05$. Also as above, *City* alone did not exhibit a statistically significant effect. However, unlike above, the *Watch* × *City* interaction is not statistically significant here but only a trend. This result is to be trusted over the parametric result, given the evident violation of normality.

The ART procedure can facilitate post hoc comparisons, provided levels are compared within the aligned-and-ranked effects from which they came (Mansouri et al. 2004). We can conduct pairwise comparisons for all levels of the *Watch* factor using the following R code:

```
# assuming m is the result of the call to art() above
> library(lsmeans)
> lsmeans(artlm(m, "watch"), pairwise ~ watch)
$lsmeans
watch   lsmean        SE  df   lower.CL  upper.CL
none  108.5600  7.359735 216   94.05391  123.0661
one   130.0133  7.359735 216  115.50724  144.5194
two   100.4267  7.359735 216   85.92057  114.9328

Results are averaged over the levels of: city
Confidence level used: 0.95

$contrasts
 contrast     estimate       SE  df t.ratio   p.value
 none - one -21.453333 10.40824 144  -2.061    0.1017
 none - two   8.133333 10.40824 144   0.781    0.7150
 one -  two  29.586667 10.40824 144   2.843    0.0141

Results are averaged over the levels of: city
P value adjustment: tukey method for a family of 3 means
```

The `lsmeans` procedure reports p-values corrected for multiple comparisons using Tukey's method (Kramer 1956; Tukey 1949, 1953). As for the post hoc tests conducted above, we find the only significant difference is between one and two watches, reported as $t(144) = 2.84$, $p < 0.05$.

7.4.7 Generalized Linear Models (GLM)

The classic ANOVA we have been utilizing in this chapter can be mathematically formulated by what is called the General Linear Model (LM).[4] This model describes a family of analyses where the dependent variables are a linear combination of independent variables[5] plus normally-distributed errors. However, when the assumption of normality is not met, the General*ized* Linear Model (GLM) can utilize non-normal response distributions (Nelder and Wedderburn 1972). GLMs are specified with a distribution and a link function, which describe how factors relate to responses. The LM is subsumed by the GLM when the distribution is "normal" and the link function is "identity." Many distribution-link function combinations are possible. Here we review four common uses of the GLM for data arising in HCI studies. It is important to note that such models are suitable only for between-subjects factors. For within-subjects factors, we must add random effects to the models, as described in the next subsection.

Multinomial logistic regression. Multinomial logistic regression, also referred to as nominal logistic regression, is used for categorical (nominal) responses[6] (Nelder and Wedderburn 1972). In this respect, it can be used on data also suited to N-sample Chi-Square tests of proportions. Recall the contingency data in Table 7.2, contained in `prefs2ABC.csv`. We can use multinomial logistic regression with *Team* as a two-level factor and *Preference* as a response to discover whether there were statistically significant differences in preference between sales teams.

In terms of the GLM, multinomial logistic regression uses a "multinomial" distribution and "logit" link function. The R code for executing such an analysis is:

---

[4] General Linear Models are often called "linear models" and may be abbreviated "LM." These should not be confused with General*ized* Linear Models, which may be abbreviated "GLM." However, some texts use "GLM" for linear models and "GZLM" for generalized models. Readers should take care when encountering this family of abbreviations.

[5] While not covered in this chapter, LMs and GLMs also offer the ability to use continuous independent variables, not just categorical independent variables (see Chap. 11).

[6] Multinomial logistic regression—when used with dichotomous responses such as Yes/No, True/False, Success/Fail, Agree/Disagree, or 1/0—is called "binomial regression." The GLM for binomial regression uses a "binomial" distribution and "logit" link function. It can be conducted using the `glm` function in much the same way as Poisson regression explained below, except with the parameter `family=binomial`.

```
> prefs2ABC = read.csv("chapter7/prefs2ABC.csv")
> library(nnet) #for multinom
> library(car) #for Anova
> m = multinom(email_preference ~ team, data=prefs2ABC)
> Anova(m)
Analysis of Deviance Table (Type II tests)

Response: email_preference

      LR Chisq   Df   Pr(>Chisq)
team    6.5556    2      0.03771 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 0.038, indicating a significant difference in preference. The test result is reported as a multinomial logistic regression $\chi^2(2, N = 150) = 6.56$, $p < 0.05$.

Multinomial logistic regression is useful for multifactor analyses as well. The file `prefs2ABCsex.csv` has the same preferences data now embellished with the sex of each sales representative. Thus, we have *Team* with two levels (Sales Team X or Y) and *Sex* with two levels (male or female).

The R code for executing multinomial logistic regression on `prefs2ABC sex.csv` is:

```
> prefs2ABCsex = read.csv("chapter7/prefs2ABCsex.csv")
> library(nnet) #for multinom
> library(car) #for Anova
# set contrasts for each factor to be sum-to-zero contrasts.
# this is necessary for the Type III Anova we will use.
> contrasts(prefs2ABCsex$team) <- "contr.sum"
> contrasts(prefs2ABCsex$sex) <- "contr.sum"
> m = multinom(email_preference ~ team * sex,data=prefs2ABCsex)
# We use a Type III Anova since it simplifies interpreting
# significant main effects in the presence of interactions.
> Anova(m, type=3)
Analysis of Deviance Table (Type III tests)

Response:  email_preference

          LR Chisq   Df   Pr(>Chisq)
team        6.4923    2      0.03892  *
sex         0.1029    2      0.94985
team:sex    0.5136    2      0.77354
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As with our contingency table analysis, here there was a statistically significant effect of *Team* on email application preference ($\chi^2(2, N = 150) = 6.49$, $p < 0.05$). However, as one would expect, there was no statistically significant effect of *Sex* or significant *Team* × *Sex* interaction.

Ordinal logistic regression. Ordinal logistic regression, also called ordered logit, proportional odds logistic regression, or the cumulative link model, is analogous to multinomial logistic regression but for ordered responses rather than unordered categories (McCullagh 1980). Such responses occur frequently in HCI studies that utilize Likert scales, e.g., with subjective responses ranging from "strongly disagree" to "strongly agree." Ordinal logistic regression is an extension of multinomial logistic regression to ordered response categories.

Let us assume that each sales representative was asked to indicate how much they liked the email application they most preferred. On a 1–7 scale with endpoints "strongly disagree" to "strongly agree," they rated their agreement with the statement, "I love my preferred Mango smartwatch email application." The data in `prefs2ABClove.csv` reflects their responses.

The R code for running ordinal logistic regression on `prefs2ABClove.csv` is:

```
> prefs2ABClove = read.csv("chapter7/prefs2ABClove.csv")
> library(MASS)
> library(car)
> contrasts(prefs2ABClove$team) <- "contr.sum"
> contrasts(prefs2ABClove$sex) <- "contr.sum"
# transform numeric variable into an ordinal variable
> prefs2ABClove$love = ordered(prefs2ABClove$love)
> m = polr(love ~ team * sex, data=prefs2ABClove)
> Anova(m, type=3)
Analysis of Deviance Table (Type III tests)

Response: love

           LR Chisq    Df     Pr(>Chisq)
team         0.0149     1         0.9029
sex         29.5134     1      5.553e-08 ***
team:sex     0.0285     1         0.8659
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There was no significant effect of *Team* on how much sales representatives love their preferred Mango email application. There was also no *Team* × *Sex* interaction. But there was a significant effect of *Sex*, reported as ordinal logistic $\chi^2(1, N = 150) = 29.51$, $p < 0.0001$. The average 1–7 Likert rating for males was 4.45; for females it was 5.70. Perhaps the female sales representatives were a more positive bunch!

Poisson regression. Poisson regression is used for nonnegative integers that represent count data (Nelder and Wedderburn 1972; Bortkiewicz 1898). A common use of Poisson regression in HCI is for counts of rare events. For example, accurate gesture recognizers or automatic spelling correction systems that produce relatively few errors from every 100 attempts may lend themselves to Poisson regression.[7]

For our scenario, let us pretend that the Mango smartwatch email applications tracked and counted the number of customers to whom each sales representative failed to respond within 48 hours. Since sales representatives are trained to respond quickly to customers, such occurrences should be relatively rare. The file `prefs2ABClate.csv` contains the email application preferences data embellished with the number of late responses for that sales representative during the three-month study. Now, the preferred email application is treated not as a response but as an independent variable potentially influencing the new response. Thus, we have a $2 \times 2 \times 3$ three-factor design with *Team* (X, Y), *Sex* (M, F), and *Preference* (A-mail, B-mail, C-mail).

In the GLM, Poisson regression uses a "Poisson" distribution and "log" link function, specified by the `family` argument to the `glm` function. It is executed in R with the following:

```
> prefs2ABClate = read.csv("chapter7/prefs2ABClate.csv")
> contrasts(prefs$ABClate$team) <- "contr.sum"
> contrasts(prefs2ABClate$sex) <- "contr.sum"
> contrasts(prefs2ABClate$email_preference) <- "contr.sum"
> m = glm(late_responses ~ team * sex * email_preference,
          data=prefs2ABClate, family=quasipoisson)
> Anova(m, type=3)
Analysis of Deviance Table (Type III tests)

Response: late_responses

                           LR Chisq     Df    Pr(>Chisq)
team                        20.8684      1     4.919e-06 ***
sex                          1.2934      1       0.25541
email_preference             3.3016      2       0.19190
team:sex                     2.0317      1       0.15405
team:email_preference        0.1471      2       0.92907
sex:email_preference         5.2803      2       0.07135 .
team:sex:email_preference    4.1328      2       0.12664
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[7] Given data with a large number of zeroes, it is prudent to consider an extension to Poisson regression called "`zero-inflated`" Poisson regression. This model incorporates binomial regression to predict the probability of a zero alongside Poisson regression to model counts. See the `zeroinfl` function in the `pscl` package.

There was a significant effect of *Team* on number of late responses, reported as a Poisson regression $\chi^2(1, N = 150) = 20.87$, $p < 0.0001$.

```
> ddply(prefs2ABClate, ~ team, function(data)
        summary(data$late_responses))
  team  Min. 1st Qu.   Median     Mean 3rd Qu. Max.
1    X     0       0        1    1.080       2    4
2    Y     0       1        2    2.307       3    8
```

The average number of late responses for Sales Team X was 1.08; for Sales Team Y, which did not have the Mango smartwatch yet, it was 2.31. We can also extract the estimated ratio of rates of late responses between the two teams:

```
> library(multcomp)  #for glht
> library(lsmeans)   #for lsm
> team_effect = confint(glht(m, lsm(pairwise ~ team)))

    Simultaneous Confidence Intervals

Fit: glm(formula = late_responses ~ team * sex *
email_preference, family = quasipoisson, data = prefs2ABClate)
Quantile = 1.96
95% family-wise confidence level

Linear Hypotheses:

            Estimate     lwr     upr
X - Y == 0   -0.6594 -0.9465 -0.3723

# effects in a Poisson model are on a log scale (because of the
# log link), so we often exponentiate them to interpret them.
> exp(team_effect$confint)
       Estimate       lwr       upr
X - Y 0.5171661 0.3880856 0.6891797
attr(,"conf.level")
[1] 0.95
attr(,"calpha")
[1] 1.959964
```

Thus, we should expect members of Team X to have about 0.517 times the rate of late responses as Team Y (95 % confidence interval: [0.388, 0.689]).

Gamma regression. For data conforming to a Gamma distribution, a GLM can be fitted with a "log" link function.[8] A Gamma distribution applies to skewed, continuous data with a theoretical minimum, often zero. It is defined by two parameters, "shape" and "scale." The inverse of the scale parameter is called the "rate." The exponential distribution is a special case of Gamma distribution where the shape parameter equals one.

The data contained in `salesXY.csv` and graphed in Fig. 7.1 can be modeled by a Gamma distribution. That data has now been embellished with the sex of each salesperson in `salesXYsex.csv`. The question now is how *Team* and *Sex* may have affected annualized sales.

The R code for executing a GLM with a Gamma distribution and log link function is:

```
> salesXYsex = read.csv("chapter7/salesXYsex.csv")
> contrasts(salesXYsex$team) <- "contr.sum"
> contrasts(salesXYsex$sex) <- "contr.sum"
> m = glm(sales ~ team * sex, data=salesXYsex,
          family=Gamma(link="log"))
> Anova(m, type=3)
Analysis of Deviance Table (Type III tests)

Response: sales

          LR Chisq     Df   Pr(>Chisq)
team        5.1408      1      0.02337 *
sex         1.9128      1      0.16666
team:sex    0.1767      1      0.67420
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results indicate that there was a statistically significant effect of *Team* on annualized sales ($\chi^2(1, N = 150) = 5.14$, $p < 0.05$). The same conclusion was reached with the Mann-Whitney U test, whose p-value of 0.025 was similar. It does not appear that *Sex* or *Team* × *Sex* had any effect on annualized sales.

7.4.8 Generalized Linear Mixed Models (GLMM)

The GLMs reviewed in the last section are powerful models but with a major limitation—they are unable to handle within-subjects factors because they cannot

[8] Although the canonical link function for the Gamma distribution is actually the "inverse" function, the "log" function is often used because the inverse function can be difficult to estimate due to discontinuity at zero. The two functions provide similar results.

account for correlations among measures of the same participant. This restriction limits their utility in HCI studies in which the same participants are measured repeatedly in one session or over time.

The Generalized Linear Mixed Model (GLMM) is an extended model that allows for within-subjects factors (Gilmour et al. 1985; Stiratelli et al. 1984). A "mixed-effects model" refers to the combination of both "fixed" and "random" effects. Thus far in this chapter, we have only considered fixed effects, which are those whose levels are purposefully and specifically chosen as treatments. By contrast, random effects have levels whose values are not themselves of interest, but that represent a random sample from a larger population about which we wish to generalize. In HCI studies with repeated measures, the random effects are almost always the human participants in the experiment. By modeling *Subject* as a random effect, the correlation among measures taken from the same participant can be accounted for.

In other respects, GLMMs are similar to the GLMs reviewed above. Distributions and link functions can be specified for non-normal data.

Let us again consider the annualized sales for Sales Team Y with no Mango smartwatch, one Mango smartwatch, and two Mango smartwatches in `salesYY2city.csv`. (See Fig. 7.5.)

The R code for executing a factorial GLMM with a Gamma distribution, log link function, and *Subject* as a random effect[9] is:

```
> salesYY2city = read.csv("chapter7/salesYY2city.csv")
> library(lme4)       #for glmer
> library(car)        #for Anova
> contrasts(salesYY2city$city) <- "contr.sum"
> contrasts(salesYY2city$watch) <- "contr.sum"
# here (1|subject) indicates a random intercept
# dependent on subject.
> m = glmer(sales ~ city * watch + (1|subject),
            data=salesYY2city, family=Gamma(link="log"))
> Anova(m, type=3)
Analysis of Deviance Table (Type III Wald chisquare tests)

Response: sales

                   Chisq  Df  Pr(>Chisq)
(Intercept)   37079.7322   1     < 2e-16 ***
city              1.0191   2     0.60077
watch             5.4790   2     0.06460 .
city:watch       10.9741   4     0.02686 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05  '.' 0.1 ' ' 1
```

[9] This model uses an intercept-only random effect. There are other types of random effects such as slopes-and-intercept random effects that are described in Chap. 11.

The results show a statistically significant *Watch* × *City* interaction ($\chi^2(4, N = 225) = 10.97$, $p < 0.05$) and a nonsignificant main effect of *Watch* ($\chi^2(2, N = 225) = 5.48$, $p = 0.065$). These results differ somewhat from those of the Aligned Rank Transform, which showed a statistically significant main effect of *Watch* ($F(2,144) = 4.31$, $p < 0.05$) and a nonsignificant *Watch* × *City* interaction ($F(4,144) = 2.02$, $p = 0.095$). In neither analysis was *City* statistically significant. The discrepancies in these results indicate the degree to which statistical conclusions may vary depending on the analyses used.

7.4.9 Generalized Estimating Equations (GEE)

When the relationship among factors and responses is not known or there is no discernable structure, a Generalized Estimating Equation (GEE) can be used (Liang and Zeger 1986; Zeger et al. 1988). Unlike GLMMs, GEEs are less sensitive to covariance structure specification and can handle unknown correlation among outcomes. Responses may be continuous, nominal, or ordinal. Statistical inference is commonly performed with the Wald test (Wald 1943).

The R code for using a GEE with a Gamma distribution and log link function on `salesYY2city.csv` is:

```
> salesYY2city = read.csv("chapter7/salesYY2city.csv")
> library(geepack)
# geeglm requires data sorted by grouping variable, so we sort
# by subject (so that all rows for a given subject are
# contiguous).
> salesYY2city = salesYY2city[order(salesYY2city$subject),]
> m = geeglm(sales ~ city * watch, id=subject,
             data=salesYY2city, family=Gamma(link="log"))
> anova(m)
Analysis of 'Wald statistic' Table
Model: Gamma, link: log
Response: sales
Terms added sequentially (first to last)

           Df      X2    P(>|Chi|)

city        2    0.46        0.795
watch       2    7.94        0.019 *
city:watch  4   12.03        0.017 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results show a statistically significant *Watch* main effect ($\chi^2(2, N = 225) = 7.94$, $p < 0.05$) and *Watch* × *City* interaction ($\chi^2(4, N = 225) = 12.03$, $p < 0.05$).

However, `geeglm` does not support Type-III ANOVAs,[10] and as a result, our interpretation of these tests is slightly different: The significant effect of *Watch* can be interpreted as a significant effect only outside the presence of a significant *Watch* × *City* interaction, which this model contains. Therefore, we ignore the significant *Watch* main effect and focus any further analysis on the interaction. As above, *City* is statistically nonsignificant.

7.5 Summary

The field of human-computer interaction is a field devoted both to invention and to science. Its researchers and practitioners often transition fluidly from inventing new interactive technologies to scientifically evaluating the behavior of people with interactive technologies. The ability to correctly draw conclusions about this behavior is therefore of paramount importance in focusing the efforts of these professionals. Although not all studies in HCI require statistical inference, those that do must utilize it correctly or risk missing actual benefits or proclaiming phantom ones.

With the wide variety of data collected in HCI studies, nonparametric statistics are rife with opportunity for broad application. Such statistics may be understood best by their relationship to more familiar, but often inapplicable, parametric statistics. This chapter has provided an overview of nonparametric statistics useful in HCI at an exciting time when the appreciation of their utility is growing in the field.

References

Anderson TW, Darling DA (1952) Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann Math Stat 23(2):193–212

Anderson TW, Darling DA (1954) A test of goodness of fit. J Am Stat Assoc 49(268):765–769

Brown GW, Mood AM (1948) Homogeneity of several samples. Am Stat 2(3):22

Brown GW, Mood AM (1951) On median tests for linear hypotheses. In: Proceedings of the second Berkeley symposium on mathematical statistics and probability, Berkeley, California. University of California Press, Berkeley, California, pp 159–166

Conover WJ, Iman RL (1981) Rank transformations as a bridge between parametric and nonparametric statistics. Am Stat 35(3):124–129

D'Agostino RB (1986) Tests for the normal distribution. In: D'Agostino RB, Stephens MA (eds) Goodness-of-fit techniques. Marcel Dekker, New York, pp 367–420

Dixon WJ, Mood AM (1946) The statistical sign test. J Am Stat Assoc 41(236):557–566

Fawcett RF, Salter KC (1984) A Monte Carlo study of the F test and three tests based on ranks of treatment effects in randomized block designs. Commun Stat Simul Comput 13(2):213–225

[10] The ANOVA type indicates how the sums-of-squares are computed. In general, Type III ANOVAs are preferred because they can support conclusions about main effects in the presence of significant interactions. For Type I and Type II ANOVAs, significant main effects cannot safely be interpreted in the presence of significant interactions.

Fisher RA (1921) On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1(4):3–32
Fisher RA (1922) On the interpretation of χ^2 from contingency tables, and the calculation of P. J R Stat Soc 85(1):87–94
Fisher RA (1925) Statistical methods for research workers. Oliver and Boyd, Edinburgh
Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32(200):675–701
Gilmour AR, Anderson RD, Rae AL (1985) The analysis of binomial data by a generalized linear mixed model. Biometrika 72(3):593–599
Greenhouse SW, Geisser S (1959) On methods in the analysis of profile data. Psychometrika 24(2):95–112
Higgins JJ, Blair RC, Tashtoush S (1990) The aligned rank transform procedure. In: Proceedings of the conference on applied statistics in agriculture. Kansas State University, Manhattan, Kansas, pp 185–195
Higgins JJ, Tashtoush S (1994) An aligned rank transform test for interaction. Nonlinear World 1(2):201–211
Higgins JJ (2004) Introduction to modern nonparametric statistics. Duxbury Press, Pacific Grove
Hodges JL, Lehmann EL (1962) Rank methods for combination of independent experiments in the analysis of variance. Ann Math Stat 33(2):482–497
Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6(2):65–70
Kolmogorov A (1933) Sulla determinazione empirica di una legge di distributione. Giornale dell'Istituto Italiano degli Attuari 4:83–91
Kramer CY (1956) Extension of multiple range tests to group means with unequal numbers of replications. Biometrics 12(3):307–310
Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. J Amer Stat Assoc 47(260):583–621
Lehmann EL (2006) Nonparametrics: statistical methods based on ranks. Springer, New York
Levene H (1960) Robust tests for equality of variances. In: Olkin I, Ghurye SG, Hoeffding H, Madow WG, Mann HB (eds) Contributions to probability and statistics. Stanford University Press, Palo Alto, pp 278–292
Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73(1):13–22
Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18(1):50–60
Mansouri H (1999a) Aligned rank transform tests in linear models. J Stat Plann Inference 79(1):141–155
Mansouri H (1999b) Multifactor analysis of variance based on the aligned rank transform technique. Comput Stat Data Anal 29(2):177–189
Mansouri H, Paige RL, Surles JG (2004) Aligned rank transform techniques for analysis of variance and multiple comparisons. Commun Stat Theory Methods 33(9):2217–2232
Massey FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc 46(253):68–78
Mauchly JW (1940) Significance test for sphericity of a normal n-variate distribution. Ann Math Stat 11(2):204–209
McCullagh P (1980) Regression models for ordinal data. J R Stat Soc Ser B 42(2):109–142
Mehta CR, Patel NR (1983) A network algorithm for performing Fisher's exact test in r × c contingency tables. J Am Stat Assoc 78(382):427–434
Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc Ser A 135(3):370–384
Pearson K (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos Mag Ser 5 50(302):157–175
Razali NM, Wah YB (2011) Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J Stat Model Anal 2(1):21–33

Richter SJ (1999) Nearly exact tests in factorial experiments using the aligned rank transform. J Appl Stat 26(2):203–217
Salter KC, Fawcett RF (1985) A robust and powerful rank test of treatment effects in balanced incomplete block designs. Commun Stat Simul Comput 14(4):807–828
Salter KC, Fawcett RF (1993) The ART test of interaction: a robust and powerful rank test of interaction in factorial models. Commun Stat Simul Comput 22(1):137–153
Sawilowsky SS (1990) Nonparametric tests of interaction in experimental design. Rev Educ Res 60(1):91–126
Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52(3, 4):591–611
Smirnov H (1939) Sur les écarts de la courbe de distribution empirique. Recueil Mathématique (Matematiceskii Sbornik) 6:3–26
Sokal RR, Rohlf FJ (1981) Biometry: the principles and practice of statistics in biological research. W. H. Freeman, Oxford
Stewart WM (1941) A note on the power of the sign test. Ann Math Stat 12(2):236–239
Stiratelli R, Laird N, Ware JH (1984) Random-effects models for serial observations with binary response. Biometrics 40(4):961–971
Student (1908) The probable error of a mean. Biometrika 6(1):1–25
Tukey JW (1949) Comparing individual means in the analysis of variance. Biometrics 5(2):99–114
Tukey JW (1953) The problem of multiple comparisons. Princeton University, Princeton
von Bortkiewicz L (1898) Das Gesetz der kleinen Zahlen (The law of small numbers). Druck und Verlag von B.G. Teubner, Leipzig
Wald A (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Amer Math Soc 54(3):426–482
Welch BL (1951) On the comparison of several mean values: an alternative approach. Biometrika 38(3/4):330–336
White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48(4):817–838
Wilcoxon F (1945) Individual comparisons by ranking methods. Biomet Bull 1(6):80–83
Wobbrock JO, Findlater L, Gergle D, Higgins JJ (2011) The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures. In: Proceedings of the ACM conference on human factors in computing systems (CHI '11), Vancouver, British Columbia, 7–12 May 2011. ACM Press, New York, pp 143–146
Zeger SL, Liang K-Y, Albert PS (1988) Models for longitudinal data: a generalized estimating equation approach. Biometrics 44(4):1049–1060