CHAPTER

Statistical analysis

4

In Chapter 2, we discussed why we need to run statistical analysis on data collected through various methods. Appropriate selection of statistical analysis methods and accurate interpretation of the test results are essential for user studies. After weeks, months, or even years of arduous preparation and data collection, you finally have a heavy set of data on hand and may feel the need to lie back and enjoy a hard-earned break. Well, it is a little too early to relax and celebrate at this point. With many studies, the data analysis stage is equally or even more labor intensive than the data collection stage. Many critical decisions need to be made when analyzing the data, such as the type of statistical method to be used, the confidence threshold, as well as the interpretation of the significance test results. Incorrect selection of statistical methods or inappropriate interpretation of the results can lead to erroneous conclusions that let high-quality data go to waste.

This chapter discusses general data analysis procedures and commonly used statistical methods, including independent-samples *t* test, paired-samples *t* test, one-way analysis of variance (ANOVA), factorial ANOVA, repeated measures ANOVA, correlation, regression, chi-squared test, and four other nonparametric tests.[1] The focus of this chapter is not on the mathematical computation behind each method or how to use statistical software to conduct each analysis. Instead, we focus on the contexts of use and the assumptions of each method. We also discuss how to appropriately interpret the results of each significance test. Through this chapter, we hope that you will be able to choose appropriate statistical methods for data analysis, run the corresponding tests using statistical software, and accurately interpret the analysis results for your own studies. You will also learn how to assess the validity of the findings reported in academic articles based on the experimental design and the statistical analysis procedure.

4.1 PREPARING DATA FOR STATISTICAL ANALYSIS

In most cases, the original data collected from lab-based experiments, usability tests, field studies, surveys, and various other channels need to be carefully processed before any statistical analysis can be conducted. There are several reasons for the need for preprocessing. First, the original data collected, especially if they are entered

[1] Tests to be used when the assumptions of the parametric tests are not met. More details will be discussed in Sections 4.6 and 4.8.

Research Methods in Human-Computer Interaction. http://dx.doi.org/10.1016/B978-0-12-805390-4.00004-2

manually by participants, may contain errors or may be presented in inconsistent formats. If those errors or inconsistencies are not filtered out or fixed, they may contaminate the entire data set. Second, the original data collected may be too primitive and higher level coding may be necessary to help identify the underlying themes. Third, the specific statistical analysis method or software may require the data to be organized in a predefined layout or format so that they can be processed (Delwiche and Slaughter, 2008).

4.1.1 CLEANING UP DATA

The first thing that you need to do after data collection is to screen the data for possible errors. This step is necessary for any type of data collected, but is particularly important for data entered manually by participants. To err is human. All people make mistakes (Norman, 1988). Although it is not possible to identify all the errors, you want to trace as many errors as possible to minimize the negative impact of human errors. There are various ways to identify errors depending on the nature of the data collected.

Sometimes you can identify errors by conducting a reasonableness check. For instance, if the age of a participant is entered as “223,” you can easily conclude that there is something wrong. Your participant might have accidentally pushed the number “2” button twice, in which case the correct age should be 23, or he might have accidentally hit the number “3” button after the correct age, 22, has been entered. Sometimes you need to check multiple data fields in order to identify possible errors. For example, you may compare the participant's “age” and “years of computing experience” to check whether there is an unreasonable entry.

For automatically collected data, error checking usually boils down to time consistency issues or whether the performance is within a reasonable range. Something is obviously wrong if the logged start time of an event is later than the logged end time of the same event. You should also be on alert if any unreasonably high or low performance levels are documented.

In many studies, data about the same participant are collected from multiple channels. For example, in a study investigating multiple data-entry techniques, the performance data (such as time and number of keystrokes) might be automatically logged by data-logging software. The participants' subjective preference and satisfaction data might be manually collected via paper-based questionnaires. In this case, you need to make sure that all the data about the same participant are correctly grouped together. The result will be invalid if the performance data of one participant is grouped with the subjective data of another participant.

After errors are identified, how shall we deal with them? It is obvious that you always want to fix errors and replace them with accurate data. This is possible in some cases. If the age of a participant is incorrect, you can contact that participant and find out the accurate information. In many cases, fixing errors in the preprocessing stage is impossible. In many online studies or studies in which the participant remains anonymous, you may have no means of reaching participants after the data is collected. Under those circumstances, you need to remove the problematic data items and treat them as missing values in the statistical data analysis.

Sometimes, the data collected need to be cleaned up due to inappropriate formatting. Using age as an example, participants may enter age in various formats. In an online survey, most respondents used numeric values such as “9” to report their age (Feng et al., 2008). Some used text such as “nine” or “nine and a half.” A number of participants even entered detailed text descriptions such as “He will turn nine in January.” The entries in text formats were all transformed to numeric values before the data was analyzed by statistical software.

4.1.2 CODING DATA

In many studies, the original data collected need to be coded before any statistical analysis can be conducted. A typical example is the data about the demographic information of your participants. Table 4.1 shows the original demographic data of three participants. The information on age is numerical and does not need to be coded. The information on gender, highest degree earned, and previous software experience needs to be coded so that statistical software can interpret the input. In Table 4.2, gender information is coded using 1 to represent “male” and 0 to represent “female.” Highest degree earned has more categories, with 1 representing a high school degree, 2 representing a college degree, and 3 representing a graduate degree. Previous software experience is also coded, with 1 representing “Yes” and 0 representing “No.” Usually we use codes “0” and “1” for dichotomous variables (categorical variables with exactly two possible values). When coding variables with three or more possible values, the codes used may vary depending on the specific context. For

Table 4.1 Sample Demographic Data in Its Original Form

	Age	Gender	Highest Degree	Previous Experience In Software A
Participant 1	34	Male	College	Yes
Participant 2	28	Female	Graduate	No
Participant 3	21	Female	High school	No

Table 4.2 Sample Demographic Data in Coded Form

	Age	Gender	Highest Degree	Previous Experience In Software A
Participant 1	34	1	2	1
Participant 2	28	0	3	0
Participant 3	21	0	1	0

example, in Table 4.2, I used “1” to represent “high school degree” rather than “0.” However, when the data is processed by a statistics software, a coding scheme of “0, 1, 2” is exactly the same as a scheme of “1, 2, 3.”

In various studies such as surveys, interviews, and focus groups, content analysis needs to be conducted in which text reflecting different themes or critical events is coded and counted (Stemler, 2001). Detailed discussion on content analysis is provided in Chapter 11. Event coding is also quite common in usability tests or lab-based studies. For example, Hu and Feng (2015) used extensive coding schemes to analyze the causes for failed browsing or search tasks in an online environment. The coding scheme allowed the authors to further understand the difficulties that users experience when finding information online.

When coding your data, it is critical to ensure the coding is consistent. This is particularly challenging when the coding is completed by more than one person. If the coding is inconsistent, the validity of the analysis results will be greatly affected. Various statistical methods, such as Cronbach's alpha, can be used to assess the reliability of coding completed by multiple coders (Weber, 1990). Please see Chapter 11 for more details on this topic.

4.1.3 ORGANIZING DATA

Statistical and other data-processing software normally has predefined requirements for how data should be laid out for specific statistical analysis. In SPSS, for example, when running an independent-samples *t* test to compare two groups of data, the data of the two groups need to be listed in the same column. In contrast, when running a paired-samples *t* test to compare two means, the two groups of data need to be laid out parallel to each other in two separate columns. Similarly, other statistical methods such as ANOVA, repeated measures, and correlation all have different data organization requirements that need to be followed closely.

4.2 DESCRIPTIVE STATISTICS

After the collected data is cleaned up, you may want to run a number of basic descriptive statistical tests to understand the nature of your data set. For instance, you may want to know the range into which most of your data points fall; you may also want to know how your data points are distributed. The most commonly used descriptive measures include means, medians, modes, variances, standard deviations, and ranges.

4.2.1 MEASURES OF CENTRAL TENDENCY

When we study a data set, we often want to find out where the bulk of the data is located. In statistical terms, this characteristic is called the “central tendency.” Various measures can be used to describe the central tendency of a data set, including the mean, the median, and the mode (Rosenthal and Rosnow, 2008).

The mean is also called the "arithmetic average" of a data set. When multiple groups are involved in a study, comparing their means can provide preliminary insights on how the groups relate to each other. If you find that the mean of one group is notably higher than the other group, you may conduct significance tests, such as a *t* test, to examine whether that difference is statistically significant. The median is the middle score in a data set. Consider the following data set of typing speeds collected from seven users:

$$\{15,\ 19,\ 22,\ 29,\ 33,\ 45,\ 50\}$$

The mean of this data set is 30.4 while the median of the data set is 29.

The mode is the value that occurs with the greatest frequency in a data set. Suppose we collected the following data from seven participants about the number of hours they spend on the Internet every week:

$$\{12,\ 15,\ 22,\ 22,\ 22,\ 34,\ 34\}$$

The mode of the data set is 22.

4.2.2 MEASURES OF SPREAD

Another important group of descriptive measures that we usually want to know is how much the data points deviate from the center of the data set. In other words, we want to know how spread out our data set is. Measures in this group include range, variances, and standard deviations.

The range measures the distance between the highest and lowest scores in the data set. In the typing-speed data set of Section 4.2.1, the range is $50-15=35$. The larger the range, the more distributed the data set is.

The variance of a data set is the mean of the squared distances of all the scores from the mean of the data set. The square root of the variance is called the standard deviation. As with range, higher variances or standard deviations indicate that the data set is more distributed.

A commonly used method for describing the distribution of a data set is the normal distribution, a special bell-shaped distribution that can be defined by the mean and the standard deviation (see Figure 4.1). The pattern of normal distribution is very

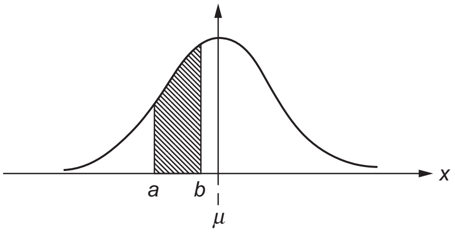


FIGURE 4.1

Normal distribution curve.

important and useful for data analysis since many attributes from various fields of study are distributed normally, such as the heights of a population, student grades, and various performance measures.

Testing a data set to determine whether it is normally distributed is a necessary step when selecting the type of significance tests to conduct. Parametric tests assume that the data set is normally distributed or approximately normally distributed. If you find that the data collected is not normally distributed, you may need to consider transforming the data so that they are normally distributed or you may adopt nonparametric tests for the analysis.

For detailed calculation of each of the measures, please refer to statistical textbooks, such as Hinkle et al. (2002), Newton and Rudestam (1999), Rosenthal and Rosnow (2008), and Albert and Tullis (2013). Microsoft Excel offers built-in functions that allow you to conveniently calculate or count various descriptive measures.

4.3 COMPARING MEANS

In user studies involving multiple conditions or groups, the ultimate objective of the researcher is to find out whether there is any difference between the conditions or groups. Suppose you are evaluating the effectiveness of two search engines; you may adopt a between-group design, in which case you will recruit two groups of participants and have each group use one of the two search engines to complete a number of search tasks. If you choose a within-group design, you will recruit one group of participants and have each participant complete a series of tasks using both search engines. In either case, you want to compare the performance measures of the two groups or conditions to find out whether there is any difference between the two search engines.

Many studies involve three or more conditions that need to be compared. Due to variances in the data, you should not directly compare the means of the multiple conditions and claim that a difference exists as long as the means are different. Instead, you have to use statistical significance tests to evaluate the variances that can be explained by the independent variables and the variances that cannot be explained by them. The significance test will suggest the probability of the observed difference occurring by chance. If the probability that the difference occurs by chance is fairly low (e.g., less than 5%), we can claim with high confidence that the observed difference is due to the difference in the controlled independent variables.

Various significance tests are available to compare the means of multiple groups. Commonly used tests include *t* tests and the ANOVA. A *t* test is a simplified ANOVA involving only two groups or conditions. Two commonly used *t* tests are the independent-samples *t* test and the paired-samples *t* test. When a study involves more than two conditions, an ANOVA test has to be used. Various ANOVA methods are available to fit the needs of different experimental designs. Commonly used ANOVA tests include one-way ANOVA, factorial ANOVA, repeated measures ANOVA, and ANOVA for split-plot design.

Table 4.3 summarizes the major types of empirical study regarding design methodology and the appropriate significance test for each design. For studies with between-group design that only investigate one independent variable with two conditions, an independent-samples *t* test can be used. When the independent variable has three or more conditions, a one-way ANOVA can be used. When a between-group study investigates two or more independent variables, a factorial ANOVA test should be considered. For studies that adopt a within-group design, if the study investigates only one independent variable with two conditions, a paired-samples *t* test can be used. If the study's independent variables have three or more conditions, a repeated measures ANOVA test can be used. Finally, a study may adopt a split-plot design that involves both a between-group component and a within-group component. In this case, a split-plot ANOVA test can be used.

Table 4.3 Commonly Used Significance Tests for Comparing Means and Their Application Context

Experiment Design	Independent Variables (IV)	Conditions for each IV	Types of Test
Between-group	1	2	Independent-samples *t* test
	1	3 or more	One-way ANOVA
	2 or more	2 or more	Factorial ANOVA
Within-group	1	2	Paired-samples *t* test
	1	3 or more	Repeated measures ANOVA
	2 or more	2 or more	Repeated measures ANOVA
Between- and within-group	2 or more	2 or more	Split-plot ANOVA

4.4 *T* TESTS

The most widely adopted statistical procedure for comparing two means is the *t* test (Rosenthal and Rosnow, 2008). Different types of *t* test should be adopted according to the specific design of the study. When the two groups being compared are presumably unrelated, an independent-samples *t* test can be used. When the two means are contributed by the same group, a paired-samples *t* test can be considered.

Suppose you want to investigate whether the use of specific word-prediction software has an impact on typing speed. The hypothesis of the test is:

> *There is no significant difference in the task completion time between individuals who use the word-prediction software and those who do not use the software.*

The following two sections will demonstrate how we investigate this hypothesis through two different designs that lead to the use of the independent-samples *t* test and the paired-samples *t* test.

4.4.1 INDEPENDENT-SAMPLES *T* TEST

You can test the hypothesis by recruiting two groups of participants and have one group type some text using standard word-processing software only and another group using the word-processing software with word-prediction functions. If a random-sampling method is used, the two groups are presumably independent from each other. In this case, the independent-samples *t* test is appropriate for data analysis.

If you use SPSS to run an independent-samples *t* test, the data points of the two groups should be listed in the same column. You need to create an additional column to mark the group to which each data point belongs. In Table 4.4, each condition has eight participants. The Coding column marks the group information, with 0 representing the participants who completed the tasks without word prediction and 1 representing the participants who completed the tasks with word prediction. When using SPSS, only the third and the fourth columns need to be entered.

Table 4.4 Sample Data for Independent-Samples *t* Test

Group	Participants	Task Completion Time	Coding
No prediction	Participant 1	245	0
No prediction	Participant 2	236	0
No prediction	Participant 3	321	0
No prediction	Participant 4	212	0
No prediction	Participant 5	267	0
No prediction	Participant 6	334	0
No prediction	Participant 7	287	0
No prediction	Participant 8	259	0
With prediction	Participant 9	246	1
With prediction	Participant 10	213	1
With prediction	Participant 11	265	1
With prediction	Participant 12	189	1
With prediction	Participant 13	201	1
With prediction	Participant 14	197	1
With prediction	Participant 15	289	1
With prediction	Participant 16	224	1

4.4.2 PAIRED-SAMPLES *T* TEST

An alternative strategy for the word-prediction software study is to recruit one group of participants and have each participant complete comparable typing tasks under both conditions. Since the data points contributed by the same participant are related, a paired-samples *t* test should be used.

If you use SPSS to run a paired-samples *t* test, the two data points contributed by the same participant should be listed parallel to each other in the same row. In Table 4.5, the two numeric values in each row were contributed by the same participant. When using SPSS to run the test, only the second and third columns need to be entered.

Table 4.5 Sample Data for Paired-Samples *t* Test

Participants	No Prediction	With Prediction
Participant 1	245	246
Participant 2	236	213
Participant 3	321	265
Participant 4	212	189
Participant 5	267	201
Participant 6	334	197
Participant 7	287	289
Participant 8	259	224

4.4.3 INTERPRETATION OF *T* TEST RESULTS

The *t* tests return a value, *t*, with larger *t* values suggesting higher probability of the null hypothesis being false. In other words, the higher the *t* value, the more likely the two means are different. As stated in Chapter 2, we normally use a 95% confidence interval in significance tests. So any *t* value that is higher than the corresponding *t* value at the 95% confidence interval suggests that there is a significant difference between participants (e.g., between users who use word-prediction software and those who do not).

SPSS generates a summary table for the results, containing both the *t* test results and additional test results that examine the data distribution. If we run an independent-samples *t* test using the data set provided in Table 4.4, the returned *t* value is 2.169, which is higher than the *t* value for the specific degree of freedom ($df = 15$) at the 95% confidence interval ($t = 2.131$).[2] In statistical terms, the result can be reported as:

> *An independent-samples t test suggests that there is significant difference in the task completion time between the group who used the standard word-processing software and the group who used word-processing software with word prediction functions ($t(15) = 2.169$, $p < 0.05$).*

Note that the *t* value needs to be reported together with the degree of freedom and the level of significance. Presenting the degree of freedom helps readers evaluate whether the data analysis is done correctly and interpret the results appropriately.

[2] The *t* value can be found in a summary table of *t*, which is available in many statistics books.

4.4.4 TWO-TAILED *T* TESTS AND ONE-TAILED *T* TESTS

In some empirical studies, the hypothesis indicates the direction of the difference. For example, you may expect the use of word-prediction software to improve typing speed. In this case, the hypothesis of the study will be:

> *Individuals who use word-prediction software can type faster than those who do not use word-prediction software.*

How does this hypothesis differ from the original hypothesis? In the original hypothesis, the direction of the difference is not specified, implying that the use of word-prediction software may improve typing speed, reduce typing speed, or have no impact on typing speed. In the hypothesis specified in this section, we expect the use of the word-prediction software to either improve typing speed, or have no impact at all. In this case, a "one-tailed *t* test" is appropriate. A *t* value that is larger than the 90% confidence interval suggests that the null hypothesis is false and that the difference between the two means is significant.

4.5 ANALYSIS OF VARIANCE

ANOVA is a widely used statistical method to compare the means of two or more groups. When there are only two means to be compared, the calculation of ANOVA is simplified to *t* tests. ANOVA tests normally return a value called the omnibus *F*. Therefore, ANOVA tests are also called "*F* tests."

4.5.1 ONE-WAY ANOVA

One-way ANOVA is appropriate for empirical studies that adopt a between-group design and investigate only one independent variable with three or more conditions. Let us revisit the word-prediction software study from Section 4.4.

Suppose you are also interested in a speech-based data-entry method and would like to compare three conditions: text entry using standard word-processing software, text entry using word-prediction software, and text entry using speech-based dictation software. The independent variable of the study has three conditions. With a between-group design, you need to recruit three groups of participants and have each group complete the text entry task using one of the three methods.

The data layout for running one-way ANOVA using SPSS is similar to that for the independent-samples *t* test. Table 4.6 presents a data set for the one-way ANOVA test. The Coding column marks the group that each data point belongs to. Normally we use 0 to mark the control group (those who used the basic word-processing software); 1 and 2 are used to mark the group who used the word-prediction software and the group who used the speech-based dictation software. When using SPSS, only the third and the fourth columns need to be entered.

Table 4.6 Sample Data for One-Way ANOVA Test

Group	Participants	Task Completion Time	Coding
Standard	Participant 1	245	0
Standard	Participant 2	236	0
Standard	Participant 3	321	0
Standard	Participant 4	212	0
Standard	Participant 5	267	0
Standard	Participant 6	334	0
Standard	Participant 7	287	0
Standard	Participant 8	259	0
Prediction	Participant 9	246	1
Prediction	Participant 10	213	1
Prediction	Participant 11	265	1
Prediction	Participant 12	189	1
Prediction	Participant 13	201	1
Prediction	Participant 14	197	1
Prediction	Participant 15	289	1
Prediction	Participant 16	224	1
Speech-based dictation	Participant 17	178	2
Speech-based dictation	Participant 18	289	2
Speech-based dictation	Participant 19	222	2
Speech-based dictation	Participant 20	189	2
Speech-based dictation	Participant 21	245	2
Speech-based dictation	Participant 22	311	2
Speech-based dictation	Participant 23	267	2
Speech-based dictation	Participant 24	197	2

Table 4.7 presents a simplified summary report provided by SPSS for the one-way ANOVA test. The between-group's sum of squares represents the amount of variances in the data that can be explained by the use of text entry methods. The within-group's sum of squares represents the amount of variances in the data that cannot be explained by the text entry methods. The mean square is calculated by dividing the sum of squares by the degree of freedom. The returned *F* value of

Table 4.7 Result of the One-Way ANOVA Test

Source	Sum of Squares	*df*	Mean Square	*F*	Significance
Between-group	7842.250	2	3921.125	2.174	0.139
Within-group	37,880.375	21	1803.827	—	—

2.174 is lower than the value at the 95% confidence interval, suggesting that there is no significant difference among the three conditions. The results can be reported as follows:

> *A one-way ANOVA test using task completion time as the dependent variable and group as the independent variable suggests that there is no significant difference among the three conditions ($F(2, 21) = 2.174$, n.s.).*

4.5.2 FACTORIAL ANOVA

Factorial ANOVA is appropriate for empirical studies that adopt a between-group design and investigate two or more independent variables.

Let us continue with the data-entry evaluation study. You may also want to know whether different types of task, such as composition or transcription, have any impact on performance. In this case, you can introduce two independent variables to your study: data-entry method and task type. There are three conditions for the data-entry method variable: standard word-processing software, word-prediction software, and speech-based dictation software. There are two conditions for the task type variable: transcription and composition. Accordingly, the empirical study has a total of $3 \times 2 = 6$ conditions. With a between-group design (see Table 4.8), you need to recruit six groups of participants and have each group complete the text entry task under one of the six conditions.

Table 4.8 A Between-Group Factorial Design With Two Independent Variables

	Standard	Prediction	Speech
Transcription	Group 1	Group 2	Group 3
Composition	Group 4	Group 5	Group 6

If you use SPSS to run the analysis, the data layout for running the factorial ANOVA test is more complicated than that of a one-way ANOVA test. Table 4.9 shows part of the data table for the factorial ANOVA test of the text entry study. The task completion time for all participants is listed in a single column. A separate coding column is created for each independent variable involved in the study. In Table 4.9, the fifth column shows whether a participant completed the transcription task or the composition task. The sixth column shows whether the participants completed the task using standard word-processing software, word-prediction software, or speech-based dictation software. When using SPSS to run the test, only columns 4, 5, and 6 need to be entered.

The SPSS procedure for a factorial ANOVA test is the univariate analysis. Table 4.10 presents the summary of the analysis results, with the first and second rows listing the information for the two independent variables, respectively. The third row lists the information for the interaction effect between the two independent variables. The analysis result suggests that there is no significant difference between participants who completed the transcription tasks and those who completed the composition tasks ($F(1, 42) = 1.41$, n.s.). There is significant difference among participants who used different text entry methods ($F(2, 42) = 4.51$, $p < 0.05$).

Table 4.9 Sample Data for the Factorial ANOVA Test

Task type	Entry method	Participant Number	Task time	Task Type coding	Entry Method coding
Transcription	Standard	Participant 1	245	0	0
Transcription	Standard	Participant 2	236	0	0
…	…	…	…	…	…
Transcription	Prediction	Participant 9	246	0	1
Transcription	Prediction	Participant 10	213	0	1
…	…	…	…	…	…
Transcription	Speech-based dictation	Participant 17	178	0	2
Transcription	Speech-based dictation	Participant 18	289	0	2
…	…	…	…	…	…
Composition	Standard	Participant 25	256	1	0
Composition	Standard	Participant 26	269	1	0
…	…	…	…	…	…
Composition	Prediction	Participant 33	265	1	1
Composition	Prediction	Participant 34	232	1	1
…	…	…	…	…	…
Composition	Speech-based dictation	Participant 41	189	1	2
Composition	Speech-based dictation	Participant 42	321	1	2
…	…	…	…	…	…
Composition	Speech-based dictation	Participant 48	202	1	2

Table 4.10 Result of the Factorial ANOVA Test

Source	Sum of Square	*df*	Mean Square	*F*	Significance
Task type	2745.188	1	2745.188	1.410	0.242
Entry method	17,564.625	2	8782.313	4.512	0.017
Task*entry	114.875	2	57.437	0.030	0.971
Error	81,751.625	42	1946.467	—	—

4.5.3 REPEATED MEASURES ANOVA

Repeated measures ANOVA tests are appropriate for empirical studies that adopt a within-group design. As stated in Section 4.5.2, the investigation of the text entry method and task type variables requires six conditions. If you adopt a between-group design, you need to recruit six groups of participants. If 12 participants are needed for each group, you must recruit a total of 72 participants. It is quite difficult to recruit such a large sample size in many HCI studies, especially those that involve participants with disabilities or specific expertise. To address that problem, you may decide to use a within-group design, in which case you recruit just one group of participants and have each participant complete the tasks under all conditions.

Repeated measures ANOVA tests can involve just one level or multiple levels. A one-way, repeated measures ANOVA test can be used for within-group studies that investigate just one independent variable. For example, if you are interested only in the impact of the text entry method, a one-way, repeated measures ANOVA test would be appropriate for the data analysis. If you use SPSS to run the test, the three data points contributed by each participant should be listed in the same row. Table 4.11 demonstrates the sample data layout for the analysis.

Table 4.11 Sample Data for One-Way, Repeated Measures ANOVA

	Standard	Prediction	Speech
Participant 1	245	246	178
Participant 2	236	213	289
Participant 3	321	265	222
Participant 4	212	189	189
Participant 5	267	201	245
Participant 6	334	197	311
Participant 7	287	289	267
Participant 8	259	224	197

Table 4.12 is the simplified summary table for the one-way, repeated measures ANOVA test generated by SPSS. The returned *F* value with degree of freedom (2, 14) is 2.925. It is below the 95% confidence interval, suggesting that there is no significant difference between the three text entry methods.

Table 4.12 Result of the One-Way, Repeated Measures ANOVA Test

Source	Sum of Square	*df*	Mean Square	*F*	Significance
Entry method	7842.25	2	3921.125	2.925	0.087
Error	18,767.083	14	1340.506	—	—

Multiple-level, repeated measures ANOVA tests are needed for within-group studies that investigate two or more independent variables. If you are interested in the impact of both the text entry method and the types of task, the study involves six conditions as illustrated in Table 4.13. A two-way, repeated measures ANOVA test can be used to analyze the data collected under this design.

Table 4.13 Experiment Design of a Two-Way, Repeated Measures ANOVA Test

	Standard	Prediction	Speech
Transcription	Group 1	Group 1	Group 1
Composition	Group 1	Group 1	Group 1

When using SPSS to run the analysis, the data need to be carefully arranged to avoid potential errors. The data points contributed by the same participant need to be listed in the same row. It is recommended that you repeat the same pattern when arranging the columns (see Table 4.14).

Table 4.14 Sample Data for Two-Way, Repeated Measures ANOVA Test

	Transcription			Composition		
	Standard	**Prediction**	**Speech**	**Standard**	**Prediction**	**Speech**
Participant 1	245	246	178	256	265	189
Participant 2	236	213	289	269	232	321
Participant 3	321	265	222	333	254	202
Participant 4	212	189	189	246	199	198
Participant 5	267	201	245	259	194	278
Participant 6	334	197	311	357	221	341
Participant 7	287	289	267	301	302	279
Participant 8	259	224	197	278	243	229

Table 4.15 presents the simplified summary table for the two-way, repeated measures ANOVA test. The task type has a significant impact on the time spent to complete the task ($F(1, 7) = 14.217$, $p < 0.01$). There is no significant difference among the three text entry methods ($F(2, 14) = 2.923$, n.s.). The interaction effect between the two independent variables is not significant either ($F(2, 14) = 0.759$, n.s.).

Table 4.15 Result of the Two-Way, Repeated Measures ANOVA Test

Source	Sum of Square	*df*	Mean Square	*F*	Significance
Task type	2745.187	1	2745.187	14.217	0.007
Error (task type)	1351.646	7	193.092	—	—
Entry method	17,564.625	2	8782.313	2.923	0.087
Error (entry method)	42,067.708	14	3004.836	—	—
Task type*entry method	114.875	2	57.438	0.759	0.486
Error (task type*entry method)	1058.792	14	75.628	—	—

4.5.4 ANOVA FOR SPLIT-PLOT DESIGN

Sometimes you may choose a study design that involves both between-group factors and within-group factors. In the text entry study, you may recruit two groups of participants. One group completes transcription tasks using all three data-entry methods. The other group completes composition tasks using all three data-entry methods (see Table 4.16). In this case, the type of task is a between-group factor and the text entry method is a within-group factor. There are two benefits of this design as compared to a pure within-group design. First, it greatly reduces the time of the study and the participants are less likely to feel tired or bored. Second, it controls the learning effect to some extent. Compared to a pure between-group study, the mixed design allows you to compare the same number of conditions with a fairly small sample size.

Table 4.16 Split-Plot Experiment Design

	Keyboard	Prediction	Speech
Transcription	Group 1	Group 1	Group 1
Composition	Group 2	Group 2	Group 2

Table 4.17 demonstrates the sample data table for the mixed design when running the test using SPSS. Note that one column needs to be added to specify the value of the between-group variable (types of task). Data points collected from the same participant need to be listed parallel to each other in the same row.

Table 4.17 Sample Data for the Split-Plot ANOVA Test

Task Type	Participant Number	Task Type Coding	Standard	Prediction	Speech
Transcription	Participant 1	0	245	246	178
Transcription	Participant 2	0	236	213	289
Transcription	Participant 3	0	321	265	222
Transcription	Participant 4	0	212	189	189
Transcription	Participant 5	0	267	201	245
Transcription	Participant 6	0	334	197	311
Transcription	Participant 7	0	287	289	267
Transcription	Participant 8	0	259	224	197
Composition	Participant 9	1	256	265	189
Composition	Participant 10	1	269	232	321
Composition	Participant 11	1	333	254	202
Composition	Participant 12	1	246	199	198
Composition	Participant 13	1	259	194	278
Composition	Participant 14	1	357	221	341
Composition	Participant 15	1	301	302	279
Composition	Participant 16	1	278	243	229

The results of a mixed design are presented in two tables in the outputs of SPSS. Table 4.18 provides the result for the between-group factor (task type). Table 4.19 provides the result for the within-group factor (text entry method). Table 4.18 suggests that there is no significant difference between participants who complete composition or transcription tasks ($F(1, 14) = 0.995$, n.s.). Table 4.19 suggests that there is a significant difference among the three text entry methods ($F(2, 28) = 5.702$, $p < 0.01$). The interaction effect between task types and text entry methods is not significant ($F(2, 28) = 0.037$, n.s.).

Table 4.18 Results of the Split-Plot Test for the Between-Group Variable

Source	Sum of Square	*df*	Mean Square	*F*	Significance
Task type	2745.187	1	2745.187	0.995	0.335
Error	38,625.125	14	2758.937	—	—

Table 4.19 Results of the Split-Plot Test for the Within-Group Variable

Source	Sum of Square	*df*	Mean Square	*F*	Significance
Entry method	17,564.625	2	8782.313	5.702	0.008
Entry method*task type	114.875	2	57.437	0.037	0.963
Error (entry method)	43,126.5	28	1540.232	—	—

4.6 ASSUMPTIONS OF *T* TESTS AND *F* TESTS

Before running a *t* test or an *F* test, it is important to examine whether your data meet the assumptions of the two tests. If the assumptions are not met, you may make incorrect inferences from those tests. Both *t* tests and *F* tests typically require three assumptions for the data:

First, the errors of all data points should be independent of each other. If they are not independent of each other, the result of the test can be misleading (Snedecor and Cochran, 1989). For example, in the text-entry method study, if two investigators conducted the study and one investigator consistently gave the participants more detailed instructions than the other investigator, the participants who completed the study with more detailed instructions might perform consistently better than those who received less detailed instructions. In this case, the errors of the participants who were instructed by the same investigator are no longer independent and the test results would be spurious.

Second, the errors in the data need to be identically distributed. This assumption is also called "homogeneity of variance." When multiple group means are compared, the *t* test or the *F* test is more accurate if the variances of the sample population are nearly equal. This assumption does not mean that we can only run *t* tests or *F* tests when the variances in the populations are exactly the same. Actually, we only become concerned when the population variances are very different or when the two sample sizes are very different (Rosenthal and Rosnow, 2008). In cases when this assumption is violated, you can use transformation techniques, such as square roots, logs, and the reciprocals of the original data (Hamilton, 1990), to make the variances in the sample population nearly equal.

Third, the errors in the data should be normally distributed. Similar to the assumption of "homogeneity of variance," this assumption is only considered to be violated when the sample data is highly skewed. When the errors are not normally distributed, nonparametric tests (discussed in Section 4.8) should be used to analyze the data.

4.7 IDENTIFYING RELATIONSHIPS

One of the most common objectives for HCI-related studies is to identify relationships between various factors. For example, you may want to know whether there is a relationship between age, computing experience, and target selection speed. In statistical terms, two factors are correlated if there is a significant relationship between them.

4.7.1 CORRELATION

The most widely used statistical method for testing correlation is the Pearson's product moment correlation coefficient test (Rosenthal and Rosnow, 2008). This test returns a correlation coefficient called Pearson's *r*. The value of Pearson's *r* ranges from −1.00 to 1.00. When the Pearson's *r* value between two variables is −1.00, it suggests a perfect negative linear relationship between the two variables. In other words, any specific increase in the scores of one variable will perfectly predict a specific amount of decrease in the scores of the other variable. When the Pearson's *r* value between two variables is 1.00, it suggests a perfect positive linear relationship between the two variables. That is, any specific increase in the scores of one variable will perfectly predict a specific amount of increase in the scores of the other variable. When the Pearson's *r* value is 0, it means that there is no linear relationship between the two variables. In other words, the increase or decrease in one variable does not predict any changes in the other variable.

In the data-entry method example, suppose the eight participants each complete two tasks, one using standard word-processing software, the other using word-prediction software. Table 4.20 lists the number of years that each participant had used computers and the time they spent on each task. We can run three Pearson's correlation tests based on this data set to examine the correlation between:

Table 4.20 Sample Data for Correlation Tests

	Computer Experience	Standard	Prediction
Participant 1	12	245	246
Participant 2	6	236	213
Participant 3	3	321	265
Participant 4	19	212	189
Participant 5	16	267	201
Participant 6	5	334	197
Participant 7	8	287	289
Participant 8	11	259	224

- computer experience and task time under the standard word-processing software condition;
- computer experience and task time under the prediction software condition; and
- task times under the standard word-processing software condition and those under the prediction software condition.

Table 4.21 illustrates the correlation matrix between the three variables generated by SPSS. The three variables are listed in the top row and the left column in the same order. The correlation between the same variable is always 1, as indicated by the three *r* values on the diagonal line of the table. The correlation between computer experience and the time using the standard software is significant, with *r* value equal to −0.723. The negative *r* value suggests that as computer experience increases, the time spent on completing the task using the standard software decreases. The correlation between computer experience and time spent using prediction software is not significant ($r = -0.468$). The correlation between the completion times using the standard software and using the prediction software is not significant either ($r = 0.325$).

Table 4.21 Results of the Correlation Tests

		Experience	Time Keyboard	Time Prediction
Experience	**Pearson correlation**	1	–0.723[a]	–0.468
	Significance	—	0.043	0.243
	N	8	8	8
Time keyboard	**Pearson correlation**	–0.723[a]	1	0.325
	Significance	0.043	—	0.432
	N	8	8	8
Time prediction	**Pearson correlation**	–0.468	0.325	1
	Significance	0.243	0.432	—
	N	8	8	8

[a] *Correlation is significant at the 0.05 level (two-tailed).*

In practice, the Pearson's r^2 is reported more often than the Pearson's *r*. The r^2 represents the proportion of the variance shared by the two variables. In other words, suppose we have two variables *X* and *Y*, the r^2 represents the percentage of variance in variable *X* that can be explained by variable *Y*. It also represents the percentage of variance in variable *Y* that can be explained by variable *X*. For many researchers, the r^2 is a more direct measure of the degree of correlation than the Pearson's *r*.

The most important thing to keep in mind about correlation is that it does not imply a causal relationship. That is, the fact that two variables are significantly correlated does not necessarily mean that the changes in one variable cause the changes in the other variable. In some cases, there is causal relationship between the two variables. In other cases, there is a hidden variable (also called the "intervening" variable, which is one type of confounding variable) that serves as the underlying cause of the change.

For example, in an experiment that studies how users interact with an e-commerce website, you may find a significant correlation between income and performance. More specifically, participants with higher income spend longer time finding a specific item and make more errors during the navigation process. Can you claim that earning a higher income causes people to spend longer time retrieving online items and make more errors? The answer is obviously no. The truth might be that people who earn a higher income tend to be older than those who earn a lower income. People in the older age group do not use computers as intensively as in the younger age group, especially when it comes to activities such as online shopping. Consequently, they may spend longer time to find items and make more errors. In this case, age is the intervening variable that is hidden behind the two variables examined in the correlation. Although income and performance are significantly correlated, there is no causal relationship between them. A correct interpretation of the relationship between the variables is listed in Figure 4.2.

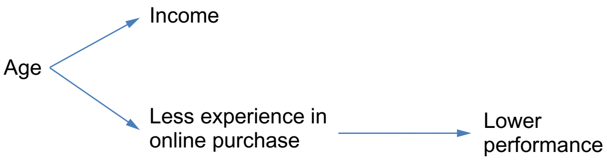


FIGURE 4.2

Relationship between correlated variables and an intervening variable.

This example demonstrates the danger of claiming causal relationship based on significant correlation. In data analysis, it is not uncommon for researchers to conduct pairwise correlation tests on all variables involved and then claim that "variable A has a significant impact on variable B" or "the changes in variable A cause variable B to change," which can be spurious in many cases. To avoid this mistake, you should keep in mind that empirical studies should be driven by hypothesis, not data. That is, your analysis should be based on a predefined hypothesis, not the other way around. In the earlier example, you are unlikely to develop a hypothesis that "income has a significant impact on online purchasing performance" since it does not make much sense. If your study is hypothesis driven, you will not be fooled by correlation analysis results. On the other hand, if you do not have a clearly defined hypothesis before the study, you will derive hypotheses driven by the data analysis, making it more likely that you will draw false conclusions.

4.7.2 REGRESSION

Unlike correlation analysis, which allows the study of only two variables, regression analysis allows you to investigate the relationship among one dependent variable and a number of independent variables. In HCI-related studies, regression analysis is used for two main purposes: model construction and prediction. In cases of model construction, we are interested in identifying the quantitative relationship between one dependent variable and a number of independent variables. That is, we want to find a mathematical equation based on the independent variables that best explains the variances in the dependent variable. In cases of prediction, we are interested in using a number of known factors, also called "predictor variables," to predict the value of the dependent variable, also called the "criterion variable" (Share, 1984). The two objectives are closely related. You need to build a robust model in order to predict the values of the criterion factor in which you are interested.

Depending on the specific research objective, you need to choose different regression procedures to construct the model. If the objective of the study is to find the relationship between the dependent variable and the independent variables as a group, you can enter all the independent variables simultaneously. This is the most commonly adopted regression procedure (Darlington, 1968). Using this approach, you will find out the percentage of variances in the dependent variable that can be explained by the independent variables as a group. This percentage is presented in the form of R^2. If the procedure returns a significant R^2, it suggests that the independent variables as a group have significant impact on the dependent variable. This procedure is useful but is insufficient if you are interested in the impact of each individual independent variable.

If you want to create a model that explains the relationship between the dependent variable and each individual independent variable, the hierarchical regression procedure is appropriate. Using this procedure, you will enter the independent variables one at a time into the regression equation. The order of the entry of the independent variables is determined by the predefined theoretical model. The independent variables that are entered into the equation first usually fall into two categories. One category includes variables that are considered to be important according to previous literature or observation; in this case, you want to evaluate the overall impact of this variable on the dependent variable. The second category includes the variables that are of no interest to you but have significant impact on the dependent variable (also called covariates); in this case, you want to exclude the variable's impact on the dependent variable before you study the variables that you are interested in. In other words, entering the covariates first allows you to remove the variances in the dependent variable that can be explained by the covariates, making it easier to identify significant relationships for the variables in which you are interested.

Suppose you conduct a user study that investigates target selection tasks using a standard mouse. One important dependent variable of interest is the task completion time and you want to know what factors have an impact on task completion time. There are a number of potential factors such as target size, distance, computer experience, age, etc. In order to find the relationships among the factors, you can conduct a regression analysis using task completion time as the dependent variable and the other factors as independent variables. Table 4.22 demonstrates a portion of the data from this study.

Table 4.22 Sample Data for the Regression Analysis

Age	Computer Experience	Target Size	Target Distance	Task Time
18	6	10	10	7
…	…	…	…	…
12	4	10	20	10
…	…	…	…	…
32	16	30	10	5
…	…	…	…	…
45	15	40	20	5
…	…	…	…	…

In this regression analysis, the dependent variable is the task completion time. The independent variables are age, computer experience (as represented by the number of years using computers), target size, and the distance between the current cursor location and the target. If you want to find out the relationship between task completion time and the independent variables as a group, simultaneous regression can be adopted. If you use SPSS to run the procedure, you enter task completion time into the dependent variable block and age, computer experience, target size, and distance into the same block for independent variables.

Table 4.23 shows the summary result of the simultaneous regression analysis. There is a significant relationship between task completion time and the independent variables as a group ($F(4, 59) = 41.147$, $p < 0.001$). The R^2 indicates the percentage of variance in the dependent variable that can be explained by the independent variables. Age, computer experience, target size, and navigation distance explain a total of 73.6% of the variance in task completion time. Please note that this percentage is unusually high since the data were made up by the authors.

Table 4.23 Result for Simultaneous Regression Procedure

Model	R	R^2	F	$df1$	$df2$	Significance
1	0.858	0.736	41.147	4	59	0.000

If you are interested in the impact that each independent variable has on task completion time, the hierarchical regression procedure can be adopted. Suppose target size and navigation distance are the most important factors that you want to examine; you can enter target size in the first block for independent variables and navigation distance, age, and computer experience into the subsequent blocks. Table 4.24 shows the summary result of this procedure. Since the four independent variables were entered separately, four regression models were constructed. Model 1 describes the relationship between task completion time and target size. It shows that target size explains a significant percentage of the variance (31.9%) in task completion time ($F(1, 62)=29.054$, $p<0.001$). The R^2 change column represents the additional variance in the dependent variable that can be explained by the newly entered independent variable. For example, Model 2 suggests that adding navigation distance to the regression model explains an additional 8.4% of the variance in task completion time. Navigation distance also has a significant impact on task completion time ($F(1, 61)=8.615$, $p<0.01$).

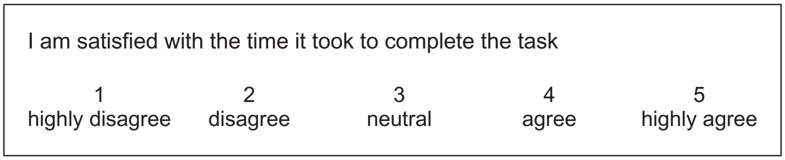
Table 4.24 Result for Hierarchical Regression Procedure

Model	*R*	*R*²	*R*² change	*F*	*df*1	*df*2	Significance
1	0.565	0.319	0.319	29.054	1	62	0.000
2	0.635	0.403	0.084	8.615	1	61	0.005
3	0.767	0.588	0.184	26.817	1	60	0.000
4	0.858	0.736	0.148	33.196	1	59	0.000

4.8 NONPARAMETRIC STATISTICAL TESTS

All the analysis methods discussed in the previous sections are parametric tests that require several general assumptions. First, the data needs to be collected from a population that is normally distributed. Usually we consider this assumption as being met if the population has an approximately normal distribution. Second, the variables should be at least scaled by intervals. That is, the distance between any two adjacent data units should be equal. For example, when examining the age variable, the distances between 1 and 2, 2 and 3, and 80 and 81 are all equal to each other. And third, for tests that compare means of different groups, the variance in the data collected from different groups should be approximately equal.

In reality, you may encounter situations where one or more of the three assumptions are not met. Some studies may yield data that poorly approximates to normal distribution. Some hypotheses may have to be measured through categorical variables (e.g., race or gender) or ordinal variables (e.g., ranking scales) where different items are compared directly with each other. In these cases, the intervals between the values are not equally spaced. For example, when collecting subjective satisfaction about an application, you may use a Likert scale question, as shown in Figure 4.3. In this case, the distance between the two adjacent data points can be unequal. The same problem exists for questions that require “yes” or “no” answers or ask participants to rank a number of options.



I am satisfied with the time it took to complete the task

1	2	3	4	5
highly disagree	disagree	neutral	agree	highly agree

FIGURE 4.3

Likert scale question.

When the assumptions of parametric tests are not met, you need to consider the use of nonparametric analysis methods. Compared to parametric tests, nonparametric methods make fewer assumptions about the data. Although nonparametric tests are also called "assumption-free" tests, it should be noted that they are not actually free of assumptions. For example, the Chi-squared test, one of the most commonly used nonparametric tests, has specific requirements on the sample size and independence of data points.

Another important message to note about nonparametric analysis is that information in the data can be lost when the data tested are actually interval or ratio. The reason is that the nonparametric analysis collapses the data into ranks so all that matters is the order of the data while the distance information between the data points is lost. Therefore, nonparametric analysis sacrifices the power to use all available information to reject a false null hypothesis in exchange for less strict assumptions about the data (Mackenzie, 2013).

4.8.1 CHI-SQUARED TEST

In user studies, we frequently encounter situations where categorical data (e.g., yes or no) are collected and we need to determine whether there is any relationship in the variables. Those data are normally presented in tables of counts (also called contingency tables) that can be as simple as a 2-by-2 table or as complicated as tables with more than 10 columns or rows. The Chi-squared test is probably the most popular significance test used to analyze frequency counts (Rosenthal and Rosnow, 2008).

Let us explore the Chi-squared test through an example. Suppose you are examining the impact of age on users' preferences toward two target selection devices: a mouse and a touchscreen. You recruit two groups of users. One group consists of 20 adult users who are younger than 65 and the other consists of 20 users who are 65 or older. After completing a series of target selection tasks using both the mouse and the touchscreen, participants specify the type of device that they prefer to use. You can generate a contingency table (see Table 4.25) that summarizes the frequency counts of the preferred device specified by the two groups of participants.

Table 4.25 A 2-by-2 Frequency Count Table

Age	Preferred Device	
	Mouse	Touchscreen
<65	14	6
≥65	4	16

As demonstrated in Table 4.25, more participants under the age of 65 prefer the mouse while more senior participants prefer the touchscreen. In order to examine whether this result is merely by chance or there is indeed a relationship between age and the preference for pointing devices, you can run a Chi-squared test. The test returns a Chi-squared value and a *P* value that helps you determine whether the result is significant. The result for the data in Table 4.25 is ($\chi^2(1) = 10.1$, $p < 0.005$). It suggests that the probability of the difference between the rows and columns occurring by chance is less than 0.005. Using the 95% confidence interval, you reject the null hypothesis and conclude that there is a relationship between age and preferred pointing device.

The degree of freedom of a Chi-squared test is calculated by the following equation:

$$\text{Degree of freedom} = (\text{Number of rows} - 1) \times (\text{Number of columns} - 1)$$

In the earlier example, the degree of freedom is $(2-1) \times (2-1) = 1$. If you have a contingency data with 3 rows and 3 columns, the degree of freedom of the Chi-squared test will be $(3-1) \times (3-1) = 4$.

If you expand the study to three pointing devices and include children in it, you have three task conditions and three participant groups. Suppose the data collected are as demonstrated in Table 4.26. In this case, the Chi-squared test result is ($\chi^2(4) = 16.8$, $p < 0.005$), suggesting that there is significant difference among the three age groups regarding preference for the pointing devices.

Table 4.26 A 3-by-3 Frequency Count Table

Age	Preferred Device		
	Mouse	Touchscreen	Stylus
<18	4	9	7
18–65	12	6	2
≥65	4	15	1

As we mentioned before, nonparametric tests are not assumption free. The Chi-squared test requires two assumptions that the data must satisfy in order to make a valid judgment. First, the data points in the contingency table must be independent from each other. In other words, one participant can only contribute one data point in the contingency table. To give a more specific example, you cannot have a participant that prefers both the mouse and the touchscreen. All the numbers presented in Tables 4.25 and 4.26 have to be contributed by independent samples. Second, the Chi-squared test does not work well when the sample is too small. It is generally suggested that, to acquire a robust Chi-square, the total sample size needs to be 20 or larger (Camilli and Hopkins, 1978).

4.8.2 OTHER NONPARAMETRIC TESTS

Many parametric tests have corresponding nonparametric alternatives. If you are comparing data collected from two independent samples (e.g., data collected using a between-group design), the independent-samples *t* test can be used when the parametric analysis assumptions are met. When the assumptions are not met, the Mann-Whitney *U* test or the Wald-Wolfowitz runs test may be considered. If you are comparing two sets of data collected from the same user group (e.g., data collected using a within-group design), the paired-samples *t* test is typically adopted when the assumptions are met. If not, the Wilcoxon signed-rank test can be used instead.

The following example illustrates the use of the Mann-Whitney *U* test. Suppose you are evaluating two authentication techniques: the traditional alphanumeric password and an image-based password that contain several images preselected by the user. You recruit two groups of participants. Each group uses one authentication technique to complete a number of login tasks. In addition to performance measures such as task completion time, failed login tasks, and keystroke level data, you also ask the participants to answer a questionnaire at the end of the study. Each participant rates the general level of frustration when using the authentication technique through a 7-point Likert scale question (1 = least frustrated, 7 = most frustrated). Sample data for the test is demonstrated in Table 4.27. The mean score for the alphanumeric password is 3.88. The mean score for the image-based password is 5.50. In order to determine whether the difference is statistically significant, you need to use nonparametric tests to compare the two groups of data. Since the data is collected from two independent groups of participants, you can use the Mann-Whitney *U* test for this analysis.

The result of the Mann-Whitney test includes a *U* value and a *z* score with the corresponding *P* value. The *z* score is a normalized score calculated based on the

Table 4.27 Sample Data for Mann-Whitney *U* test

Group	Participants	Rating	Coding
Alphanumeric	Participant 1	4	0
Alphanumeric	Participant 2	3	0
Alphanumeric	Participant 3	6	0

Table 4.27 Sample Data for Mann-Whitney *U* test *Continued*

Group	Participants	Rating	Coding
Alphanumeric	Participant 4	4	0
Alphanumeric	Participant 5	3	0
Alphanumeric	Participant 6	2	0
Alphanumeric	Participant 7	4	0
Alphanumeric	Participant 8	5	0
Image-based	Participant 9	4	1
Image-based	Participant 10	6	1
Image-based	Participant 11	6	1
Image-based	Participant 12	7	1
Image-based	Participant 13	5	1
Image-based	Participant 14	6	1
Image-based	Participant 15	4	1
Image-based	Participant 16	6	1

U value. For this example, $U=10.5$ and $p<0.05$. Therefore, the null hypothesis is rejected. The data suggests that there is significant difference in the level of perceived frustration between the two authentication techniques. Participants experienced significantly lower level of frustration when using the image-based password than the alphanumeric password.

Let us examine another scenario in which you are interested in the use of the two authentication techniques by people with Down syndrome. Since recruiting participant with Down syndrome from the local area is quite challenging, you only successfully recruit 10 participants for the study. The small participant size suggests that a within-group design will be more appropriate. So each participant completes the study using both authentication techniques and answers a questionnaire after the interaction with each technique. Sample data for the test is demonstrated in Table 4.28. The mean score for the alphanumeric password is 3.9. The mean score for the image-based password is 4.7. Since the data is collected from one group of participants, you can use the Wilcoxon signed-rank test to examine whether there is significant difference in the perceived level of frustration between the two techniques.

The result of the Wilcoxon signed-rank test is a normalized z score. For this example, $z=-1.31$ and $p=0.19$. There is no significant difference in the perceived level of frustration between the two techniques.

Table 4.28 Sample Data for Wilcoxon Signed-Rank Test

Participants	Alphanumeric	Image-Based
Participant 1	5	6
Participant 2	3	4
Participant 3	4	3

Continued

Table 4.28 Sample Data for Wilcoxon Signed-Rank Test *Continued*

Participants	Alphanumeric	Image-Based
Participant 4	5	5
Participant 5	2	7
Participant 6	3	4
Participant 7	4	6
Participant 8	6	5
Participant 9	4	3
Participant 10	3	4

In cases when three or more sets of data are compared and the parametric analysis assumptions are not met, the Kruskal-Wallis one-way ANOVA by ranks (an extension of the Mann-Whitney *U* test) may be considered when the samples are independent. When the data sets are dependent on each other, you can consider using Friedman's two-way ANOVA test.

In the authentication study discussed previously, suppose you would like to evaluate a drawing-based password technique in additional to the alphanumeric technique and the image-based technique, the study will include three conditions. If you recruit three groups of participants and let each group use one authentication technique during the study, the data will be collected from independent samples and the frustration rating can be analyzed through the Kruskal-Wallis one-way ANOVA by ranks. Sample data for the test is demonstrated in Table 4.29. The mean score for the alphanumeric password is 3.25. The mean score for the image-based password is 4.63. The mean score for the drawing-based password is 5.63.

The result of the Kruskal-Wallis test is an *H* value. In this example, $H(2)=11.897$, $p<0.05$. Therefore, the null hypothesis is rejected. The result suggests that there is significant difference in the perceived level of frustration between the three authentication techniques.

Table 4.29 Sample Data for Kruskal-Wallis One-Way Analysis of Variance by Ranks

Group	Participants	Rating	Coding
Alphanumeric	Participant 1	5	0
Alphanumeric	Participant 2	3	0
Alphanumeric	Participant 3	4	0
Alphanumeric	Participant 4	4	0
Alphanumeric	Participant 5	2	0
Alphanumeric	Participant 6	3	0
Alphanumeric	Participant 7	3	0
Alphanumeric	Participant 8	2	0
Image-based	Participant 9	3	1
Image-based	Participant 10	5	1
Image-based	Participant 11	6	1

Table 4.29 Sample Data for Kruskal-Wallis One-Way Analysis of Variance by Ranks *Continued*

Group	Participants	Rating	Coding
Image-based	Participant 12	4	1
Image-based	Participant 13	5	1
Image-based	Participant 14	4	1
Image-based	Participant 15	5	1
Image-based	Participant 16	5	1
Drawing-based	Participant 17	6	2
Drawing-based	Participant 18	4	2
Drawing-based	Participant 19	5	2
Drawing-based	Participant 20	5	2
Drawing-based	Participant 21	6	2
Drawing-based	Participant 22	7	2
Drawing-based	Participant 23	5	2
Drawing-based	Participant 24	7	2

Similarly, if you would like to evaluate the three authentication techniques when being used by people with Down syndrome, you may choose to adopt a within-group design that requires each participant to complete the tasks using all three authentication methods. In this case, the data can be analyzed through the Friedman's two-way ANOVA test. Sample data for the test is demonstrated in Table 4.30. The mean score for the alphanumeric password is 3.6. The mean score for the image-based password is 4. The mean score for the drawing-based password is 4.6.

The result of the Friedman's ANOVA test is an H value. In this example, $H(2)=2.722$, $p=0.256$. There is no significant difference in the perceived level of frustration between the three techniques.

All four nonparametric methods discussed earlier can only be used to analyze data that involves only one independent variable (factor). If you need to analyze

Table 4.30 Sample Data for Friedman's Two-Way Analysis of Variance Test

Participants	Alphanumeric	Image-Based	Drawing-Based
Participant 1	2	4	6
Participant 2	4	5	6
Participant 3	3	3	5
Participant 4	5	5	3
Participant 5	5	7	7
Participant 6	3	4	5
Participant 7	4	2	3
Participant 8	1	5	4
Participant 9	4	3	4
Participant 10	5	2	3

nonparametric data that involves two or more independent variables, you can consider using more recent approaches that extend nonparametric analysis to multifactor analysis (e.g., Kaptein et al., 2010; Wobbrock et al., 2011). For more information on this topic, please refer to sources that discuss the nonparametric analysis methods in depth, such as Conover (1999), Newton and Rudestam (1999), and Wasserman (2007).

4.9 SUMMARY

Statistical analysis is a powerful tool that helps us find interesting patterns and differences in the data as well as identify relationships between variables. Before running significance tests, the data needs to be cleaned up, coded, and appropriately organized to meet the needs of the specific statistical software package. The nature of the data collected and the design of the study determine the appropriate significance test that should be used. If the data are normally distributed and intervally scaled, parametric tests are appropriate. When the normal distribution and interval scale requirements are not met, nonparametric tests should be considered.

A number of statistical methods are available for comparing the means of multiple groups. A simple *t* test allows us to compare the means of two groups, with the independent-samples *t* test for the between-group design and the paired-samples *t* test for the within-group design. A one-way ANOVA test allows us to compare the means of three or more groups when a between-group design is adopted and there is only one independent variable involved. When two or more independent variables are involved in a between-group design, the factorial ANOVA test would be appropriate. If a study adopts a within-group design and involves one independent variable with more than two conditions, the one level repeated measures ANOVA test would be appropriate. When two or more independent variables are involved in a within-group design, the multiple-level repeated measures ANOVA test should be adopted. For studies that involve both a between-group factor and a within-group factor, the split-plot ANOVA test should be considered.

Correlation analysis allows us to identify significant relationships between two variables. When three or more variables are involved and a quantitative model is needed to describe the relationships between the dependent variables and the independent variables, regression analysis can be considered. Different regression procedures should be used based on the specific goals of the study.

Nonparametric statistical tests should be used when the data does not meet the required assumptions of parametric tests. The Chi-squared test is widely used to analyze frequency counts of categorical data. Other commonly used nonparametric tests include the Mann-Whitney *U* test, the Wilcoxon signed-rank test, the Kruskal-Wallis one-way ANOVA by ranks, and the Friedman's two-way ANOVA test. Although nonparametric tests have less strict requirements for the data, they are not assumption free and the data still need to be carefully examined before running any nonparametric tests.

DISCUSSION QUESTIONS

1. What are the major steps to prepare data for statistical analysis?
2. What are the measures of central tendency?
3. What are the measures of spread?
4. What is normal distribution? Why is it important to test whether a data sample is normally distributed?
5. What statistical methods are available for comparing group means?
6. What statistical method can be used to compare two group means contributed by two independent groups?
7. What statistical method can be used to compare two group means contributed by the same group?
8. When should a one-way ANOVA test be used? Describe a research study design that fits the one-way ANOVA test.
9. When should a factorial ANOVA test be used? Describe a research study design that fits the factorial ANOVA test.
10. When should a repeated measures ANOVA test be used? Describe a research study design that fits the repeated measures ANOVA test.
11. When should a split-plot ANOVA test be used? Describe a research study design that fits the split-plot ANOVA test.
12. When should correlation analysis be used? What does Pearson's r^2 represent?
13. When should regression analysis be used? Describe a research study that requires regression analysis.
14. Name two regression procedures and discuss when a specific procedure should be used.
15. What are the assumptions for parametric statistical tests?
16. When should nonparametric tests be considered?
17. Is the Chi-squared test “assumption free”? If not, what are the assumptions of a Chi-squared test?
18. What are the alternative nonparametric tests for the independent-samples *t* test, the paired-samples *t* test, the one-way ANOVA test, and the one-way repeated measures ANOVA test?

RESEARCH DESIGN EXERCISES

Read the following research questions and identify the appropriate statistical methods for each scenario.

1. Is there a difference in the time spent online per week for people who are single, people who are married without kids, and people who are married with kids?
2. Is there a difference between the weights of Americans and Canadians within the age ranges 20–40, 40–60, and above 60?
3. Is there a difference in the target selection speed between the mouse and the joystick for children who are 5–9 years old? (Each child uses both the mouse and the joystick during the study.)
4. Use the distance between the current cursor location and the target location to predict the amount of time needed to select a target.
5. Is there a difference between users in the United States and users in the United Kingdom when using three search engines? (Each user should use all three engines during the study.)
6. Do students in the English department have a higher GPA than students in the Education department?
7. Is there a relationship between the sales of Cheerios and the sales of milk in a grocery store?
8. Is there a difference between the blood pressures of people over 60 in the morning, at noon, and in the evening? (Each participant contributes three data points, each from a different time of the day.)

TEAM EXERCISES

Form a team of 4–5 members. Find a research topic that could be studied with existing resources available to your team. For example, comparing the time it takes to find a specific product on an e-Commerce website by two different user groups. Develop a research hypothesis and an appropriate experimental design to evaluate that hypothesis.

Recruit 10–20 participants and collect a set of data. The participants can be your classmates, friends, or relatives. Complete the following steps to analyze the data:

1. Clean up the data and code it if necessary.
2. Describe the data using descriptive statistics.
3. Select the appropriate statistical method for analyzing the data.

4. Run a significance test using statistical software.
5. Write a report to discuss the findings of the significant test. Include graphical presentations to help illustrate your findings.

Depending on how the data will be collected and used, IRB approval may or may not be needed for the study. Specific instructions should be provided by the instructors regarding the IRB requirement.

REFERENCES

Albert, W., Tullis, T., 2013. Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics, second ed. Morgan Kaufmann, Waltham, MA.

Camilli, G., Hopkins, K., 1978. Applicability of chi-square to 2×2 contingency tables with small expected cell frequencies. Psychological Bulletin 85 (1), 163–167.

Conover, W., 1999. Practical Nonparametric Statistics, third ed. John Wiley & Sons, Hoboken, NJ.

Darlington, R., 1968. Multiple regression in psychological research and practice. Psychological Bulletin 69 (3), 161–182.

Delwiche, L., Slaughter, S., 2008. The Little SAS Book: A Primer, fourth ed. SAS Institute Inc., Cary, NC.

Feng, J., Lazar, J., Kumin, L., Ozok, A., 2008. Computer usage by children with down syndrome: an exploratory study. In: Proceedings of the 10th ACM Conference on Computers and Accessibility (ASSETS). pp. 35–42.

Hamilton, L., 1990. Modern Data Analysis: A First Course in Applied Statistics. Wadsworth Publishing Company, Belmont, CA.

Hinkle, D., Wiersma, W., Jurs, S., 2002. Applied Statistics for the Behavioral Sciences, fifth ed. Houghton Mifflin Company, Boston, MA.

Hu, R., Feng, J., 2015. Investigating Information Search by People with Cognitive Disabilities. ACM Transactions on Accessible Computing 7 (1), 1–30.

Kaptein, M., Nass, C., Markopoulos, P., 2010. Powerful and consistent analysis of Likert type rating scales. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems. pp. 2391–2394.

MacKenzie, S., 2013. Human-Computer Interaction: An Empirical Research Perspective. Elsevier, Waltham, MA.

Newton, R., Rudestam, K., 1999. Your Statistical Consultant: Answers to Your Data Analysis Questions. Sage Publications, Thousand Oaks, CA.

Norman, D., 1988. The Design of Everyday Things. Basic Books, New York.

Rosenthal, R., Rosnow, R., 2008. Essentials of Behavioral Research: Methods and Data Analysis, third ed. McGraw Hill, Boston, MA.

Share, D., 1984. Interpreting the output of multivariate analyses: a discussion of current approaches. British Journal of Psychology 75 (3), 349–362.

Snedecor, G., Cochran, W., 1989. Statistical Methods, eighth ed. Iowa State University, Ames.

Stemler, S., 2001. An overview of content analysis. Practical Assessment, Research & Evaluation 7 (17) 137–146.

Wasserman, L., 2007. All of Nonparametric Statistics. Springer Science+Business Media, New York.

Weber, R.P., 1990. Basic Content Analysis: Quantitative Analysis in the Social Sciences, second ed. Sage Publications, Newbury Park, CA.

Wobbrock, J., Findlater, L., Gergle, D., Higgins, J., 2011. The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems. pp. 143–146.