Hypothesis Testing & ANOVA 6

Keywords

α-Inflation • α error • Analysis of Variance (ANOVA) • β error • Bonferroni correction • Degrees of freedom • eta-squared • F-test • F-test of sample variance • Factor variable • Familywise error rate • Independent samples *t*-test • Interaction effect • Levene's test • Mann-Whitney U test • Main effect • Nonparametric tests • Null and alternative hypothesis • Omega-squared • One-sample *t*-test • One-tailed test • One-way ANOVA • *p*-value • Paired samples *t*-test • Parametric test • Practical significance • Post hoc tests • Power of a test • Shapiro-Wilk test • Significance level • Sampling error • Statistical significance • *t*-test • Test statistic • Tukey's honestly significant difference test • Two-sample *t*-test • Two-tailed test • Two-way ANOVA • Type I and type II error • Welch correction • Wilcoxon signed-rank test • *z*-test

Learning Objectives

After reading this chapter, you should understand:

- The logic of hypothesis testing.
- The steps involved in hypothesis testing.
- What a test statistic is.
- Types of error in hypothesis testing.
- Common types of *t*-tests, one-way, and two-way ANOVA.
- How to interpret Stata output.

6.1 Introduction

Do men or women spend more money on the Internet? Assume that the mean amount that a sample of men spends online is \$200 per year against a women sample's mean of \$250. When we compare mean values such as these, we always

E. Mooi et al., *Market Research*, Springer Texts in Business and Economics,
DOI 10.1007/978-981-10-5218-7_6

expect some difference. But, how can we determine if such differences are statistically significant? Establishing statistical significance requires ascertaining whether such differences are attributable to chance or not. In this chapter, we will introduce hypothesis testing and how this helps determine statistical significance.

6.2 Understanding Hypothesis Testing

A **hypothesis** is a statement about a certain effect or parameter (such as a mean or correlation) that can be tested using a sample drawn from the population. A hypothesis may comprise a claim about the difference between two sample parameters (e.g., there is a difference between males' and females' mean spending). It can also be a test of a judgment (e.g., teenagers spend an average of 4 h per day on the Internet). Data from the sample are used to obtain evidence against, or in favor of, the statement.

Hypothesis testing is performed to infer whether or not a certain effect is statistically significant. **Statistical significance** means that the effect is so large that it is unlikely to have occured by chance. Whether results are statistically significant depends on several factors, including the size of the effect, the variation in the sample data, and the sample size (Agresti and Finlay 2014). When drawing a sample from the population, there is always some probability that we might reach the wrong conclusion due to a sampling error, which is the difference between the sample and the population characteristics. To determine whether the claim is true, we start by setting an acceptable probability (called the **significance level**) that we could incorrectly conclude there is an effect when, in fact, there is none. This significance level is typically set at 0.05, which corresponds to a 5% error probability. Next, subject to the claim made in the hypothesis, we should decide on the correct type of test to perform. This involves making decisions regarding four aspects.

First, we should understand the testing situation. What exactly are we testing: Are we comparing one value against a fixed value, or are we comparing groups, and, if so, how many?

Second, we need to specify the nature of the samples: Is our comparison based on *paired samples* or *independent samples* (the difference is discussed later in this chapter)?

Third, we should check assumptions about the distribution of our data to determine whether parametric or nonparametric tests are appropriate. **Parametric tests** make assumptions about the properties of the population distributions from which the data are drawn, while **nonparametric tests** are not based on any distributional assumptions.

Fourth, we need to decide on the region where we can reject our hypothesis; that is, whether the region of rejection will be on one side or both sides of the sampling distribution.

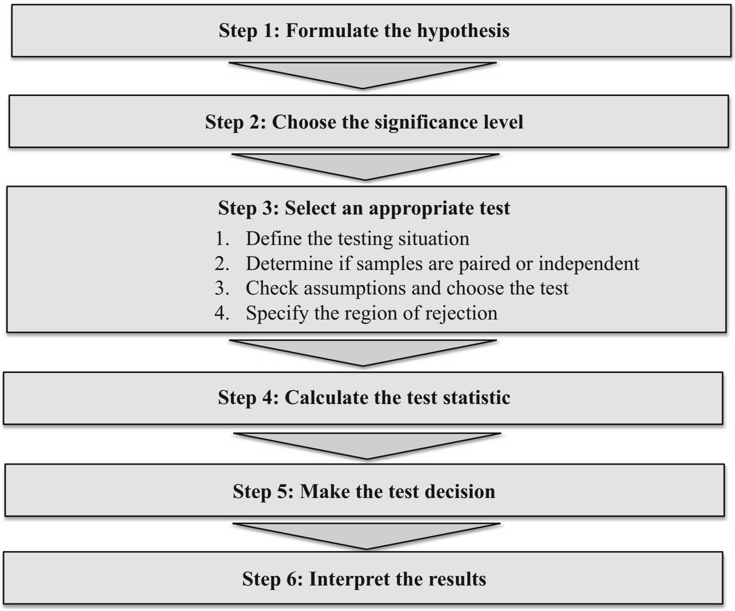

Step 1: Formulate the hypothesis
Step 2: Choose the significance level
Step 3: Select an appropriate test
1. Define the testing situation
2. Determine if samples are paired or independent
3. Check assumptions and choose the test
4. Specify the region of rejection
Step 4: Calculate the test statistic
Step 5: Make the test decision
Step 6: Interpret the results

Fig. 6.1 Steps involved in hypothesis testing

Once these four aspects are sorted, we calculate the *test statistic*, which identifies whether the sample supports or rejects the claim stated in the hypothesis. We can then decide to either reject or support the hypothesis. This decision enables us to draw market research conclusions in the final step. Figure 6.1 illustrates the six steps involved in hypothesis testing.

To illustrate the process of hypothesis testing, consider the following example: A department store chain wants to evaluate the effectiveness of three different in-store promotion campaigns that drive the sales of a specific product. These campaigns comprise: (1) a point of sale display, (2) a free tasting stand, and (3) in-store announcements. To help with the evaluation, the management decides to conduct a one-week experiment during which 30 stores are randomly assigned to each campaign type. This random assignment is important to obtain reliable and generalizable results, because randomization should equalize the effect of systematic factors not accounted for in the experimental design (see Chap. 4). Table 6.1 shows the sales of the three different in-store promotion campaigns. The table also contains information on the service type (personal or self-service) in the first column and the *marginal means* representing the means of sales within stores in the last column. The very last row also shows the marginal mean of the type of campaign, while the very last cell shows the grand mean across all the service types and campaigns.

Table 6.1 Sales data

Service type	Sales (units) Point of sale display (stores 1–10)	 Free tasting stand (stores 11–20)	 In-store announcements (stores 21–30)	Marginal mean
Personal	50	55	45	50.00
Personal	52	55	50	52.33
Personal	43	49	45	45.67
Personal	48	57	46	50.33
Personal	47	55	42	48.00
Self-service	45	49	43	45.67
Self-service	44	48	42	44.67
Self-service	49	54	45	49.33
Self-service	51	54	47	50.67
Self-service	44	44	42	43.33
Marginal mean	47.30	52.00	44.7	48.00 Grand mean

We will use these data to carry out tests to compare the different in-store promotion campaigns' mean sales separately, or in comparison to each other. We first discuss each test theoretically (including the formulas), followed by an empirical illustration. You will realize that the formulas are not as complicated as you might have thought! These formulas contain Greek characters and we have therefore included a table describing each Greek character in the ⭳ Web Appendix (→ Downloads).

6.3 Testing Hypotheses on One Mean

6.3.1 Step 1: Formulate the Hypothesis

Hypothesis testing starts with the formulation of a null and alternative hypothesis. A **null hypothesis** (indicated as H_0) is a statement expecting no difference or effect. Conversely, an **alternative hypothesis** (indicated as H_1) is the hypothesis against which the null hypothesis is tested (Everitt and Skrondal 2010). Examples of potential null and alternative hypotheses on the campaign types are:

1. H_0: The mean sales in stores that installed a point of sale display are equal to or lower than 45 units.
   H_1: The mean sales in stores that installed a point of sale display are higher than 45 units.

2. H_0: There's no difference in the mean sales of stores that installed a point of sale display and those that installed a free tasting stand (statistically, the average sales of the point of sale display = the average sales of the free tasting stand).
   H_1: There's a difference in the mean sales of stores that installed a point of sale display and those that installed a free tasting stand (statistically, the average sales of the point of sale display ≠ the average sales of the free tasting stand).

Hypothesis testing can have two outcomes: First, we do not reject the null hypothesis. This suggests there is no difference and that the null hypothesis can be retained. However, it would be incorrect to subsequently conclude that the null hypothesis is true, as it is not possible to "prove" the non-existence of a certain effect or condition. For example, one can examine any number of crows and find that they are all black, yet that would not make the statement "There are no white crows" true. Only sighting one white crow will prove its existence. Second, we could reject the null hypothesis, thus finding support for the alternative hypothesis in which some effect is expected. This outcome is, of course, desirable in most analyses, as we generally want to show that something (such as a promotion campaign) is related to a certain outcome (e.g., sales). Therefore, we frame the effect that we want to investigate as the alternative hypothesis.

Inevitably, each hypothesis test has a certain degree of uncertainty so that even if we reject a null hypothesis, we can never be totally certain that this was the correct decision. Consequently, market researchers should use terms such as "find support for the alternative hypothesis" when they discuss their findings. Terms like "prove" should never be part of hypotheses testing.

Returning to our initial example, the management only considers a campaign effective if the sales it generates are higher than the 45 units normally sold (you can choose any other value, the idea is to test the sample mean against a given standard). One way of formulating the null and alternative hypotheses in respect of this expectation is:

$$H_0\colon \mu \leq 45$$

$$H_1\colon \mu > 45$$

In words, the null hypothesis H_0 states that the population mean, indicated by μ (pronounced as *mu*), is equal to or smaller than 45, whereas the alternative hypothesis H_1 states that the population mean is larger than 45. It is important to note that the hypothesis always refers to a population parameter, in this case, the population mean, represented by μ. It is practice for Greek characters to represent population parameters and for Latin characters to indicate sample statistics (e.g., the Latin $\bar{x}$). In this example, we state a *directional hypothesis* as the alternative hypothesis, which is expressed in a direction (higher) relative to the standard of 45 units. Since we presume that during a campaign, the product sales are higher, we posit a *right-tailed hypothesis* (as opposed to a *left-tailed hypothesis*) for the alternative hypothesis H_1.

Alternatively, presume we are interested in determining whether the mean sales of the point of sale display (μ_1) are equal to the mean sales of the free tasting stand (μ_2). This implies a *non-directional hypothesis*, which can be written as:

$$H_0\colon \mu_1 = \mu_2$$

$$H_1\colon \mu_1 \neq \mu_2$$

The difference between the two general types of hypotheses is that a directional hypothesis looks for an increase or a decrease in a parameter (such as a population mean) relative to a specific standard. A non-directional hypothesis tests for *any* difference in the parameter, whether positive or negative.

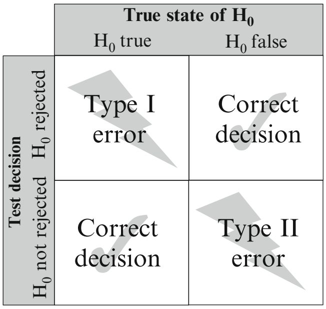
6.3.2 Step 2: Choose the Significance Level

No type of hypothesis testing can evaluate the validity of a hypothesis with absolute certainty. In any study that involves drawing a sample from the population, there is always some probability that we will erroneously retain or reject the null hypothesis due to **sampling error**, which is a difference between the sample and the population. In statistical testing, two types of errors can occur (Fig. 6.2):

1. a true null hypothesis is incorrectly rejected (**type I** or ***α*** **error**), and
2. a false null hypothesis is not rejected (**type II** or ***β*** **error**).

In our example, a type I error occurs if we conclude that the point of sale displays increased the sales beyond 45 units, when in fact it did not increase the sales, or may have even decreased them. A type II error occurs if we do not reject the null hypothesis, which suggests there was no increase in sales, even though the sales increased significantly.

Fig. 6.2 Type I and type II errors



		True state of H_0	
		H_0 true	H_0 false
Test decision	H_0 rejected	Type I error	Correct decision
	H_0 not rejected	Correct decision	Type II error

A problem with hypothesis testing is that we don't know the true state of the null hypothesis. Fortunately, we can establish a level of confidence that a true null hypothesis will not be erroneously rejected. This is the maximum probability of a type I error that we want to allow. The Greek character α (pronounced as *alpha*) represents this probability and is called the *significance level*. In market research reports, this is indicated by phrases such as "this test result is significant at a 5% level." This means that the researcher allowed for a maximum chance of 5% of mistakenly rejecting a true null hypothesis.

The selection of an α level depends on the research setting and the costs associated with a type I error. Usually, α is set to 0.05, which corresponds to a 5% error probability. However, when researchers want to be conservative or strict in their testing, such as when conducting experiments, α is set to 0.01 (i.e., 1%). In exploratory studies, an α of 0.10 (i.e., 10%) is commonly used. An α-level of 0.10 means that if you carry out ten tests and reject the null hypothesis every time, your decision in favor of the alternative hypothesis was, on average, wrong once. This might not sound too high a probability, but when much is at stake (e.g., withdrawing a product because of low satisfaction ratings) then 10% may be too high.

Why don't we simply set α to 0.0001% to really minimize the probability of a type I error? Setting α to such a low level would obviously make the erroneous rejection of the null hypothesis very unlikely. Unfortunately, this approach introduces another problem. The probability of a type I error is inversely related to that of a type II error, so that the smaller the risk of a type I error, the higher the risk of a type II error! However, since a type I error is considered more severe than a type II error, we control the former directly by setting α to a desired level (Lehmann 1993).

Sometimes statistical significance can be established even when differences are very small and have little or no managerial implications. Practitioners, usually refer to "significant" as being practically significant rather than statistically significant. **Practical significance** refers to differences or effects that are large enough to influence the decision-making process. An analysis may disclose that a type of packaging increases sales by 10%, which could be practically significant. Whether results are practically significant depends on the management's perception of the difference or effect and whether this warrants action. It is important to separate statistical significance from practical significance. Statistical significance does not imply practical significance.

Another important concept related to this is the **power of a statistical test** (defined by $1 - \beta$, where β is the probability of a type II error), which represents the probability of rejecting a null hypothesis when it is, in fact, false. In other words, the power of a statistical test is the probability of rendering an effect significant when it is indeed significant. Researchers want the power of a test to be as high as

Box 6.1 Statistical Power of a Test

Market researchers encounter the common problem that they, given a predetermined level of α and some fixed parameters in the sample, have to calculate the sample size required to yield an effect of a specific size. Computing the required sample size (called a *power analysis*) can be complicated, depending on the test or procedure used. Fortunately, Stata includes a power and sample size module that allows you to determine sample size under different conditions. In the ⭳ Web Appendix (→ Downloads), we use data from our example to illustrate how to run a power analysis using Stata. An alternative is G * Power 3.0, which is sophisticated and yet easy-to-use. It can be downloaded freely from http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/.

If these tools are too advanced, Cohen (1992) suggests required sample sizes for different types of tests. For example, detecting the presence of differences between two independent sample means for $\alpha = 0.05$ and a power of $\beta = 0.80$ requires a sample size (n) of $n = 26$ for large differences, $n = 64$ for medium differences, and $n = 393$ for small differences. This demonstrates that sample size requirements increase disproportionally when the effect that needs to be detected becomes smaller.

possible, but when maximizing the power and, therefore, reducing the probability of a type II error, the occurrence of a type I error increases (Everitt and Skrondal 2010). Researchers generally view a statistical power of 0.80 (i.e., 80%) as satisfactory, because this level is assumed to achieve a balance between acceptable type I and II errors. A test's statistical power depends on many factors, such as the significance level, the strength of the effect, and the sample size. In Box 6.1 we discuss the statistical power concept in greater detail.

6.3.3 Step 3: Select an Appropriate Test

Selecting an appropriate statistical test is based on four aspects. First, we need to assess the testing situation: What are we comparing? Second, we need to assess the nature of the samples that are being compared: Do we have one sample with observations from the same object, firm or individual (paired), or do we have two different sets of samples (i.e., independent)? Third, we need to check the assumptions for normality to decide which type of test to use: Parametric (if we meet the test conditions) or non-parametric (if we fail to meet the test conditions)? This step may involve further analysis, such as testing the homogeneity of group variances. Fourth, we should decide on the region of rejection: Do we want to test one side or both sides of the sampling distribution? Table 6.2 summarizes these four aspects with the recommended choice of test indicated in the grey shaded boxes. In the following we will discuss each of these four aspects.

Table 6.2 Selecting an appropriate test

	Testing situation	Nature of samples	Choice of Test[a]			Region of rejection
Test #	*What do we compare*	*Paired vs. Independent*	*Assumptions*	*Parametric*	*Non-parametric*	*One or two-sided test*
1	One group against a fixed value	Not applicable	Shapiro-Wilk test = normal	One sample t-test		One or two-sided
			Shapiro-Wilk test ≠ normal		Wilcoxon signed-rank test	One or two-sided
2	Outcome variable across two groups	Paired sample	Shapiro-Wilk test = normal & Levene's test: $\sigma_1^2 = \sigma_2^2$	Paired t-test		One or two-sided
			Shapiro-Wilk test ≠ normal & Levene's test: $\sigma_1^2 = \sigma_2^2$	Paired t-test[b]		One or two-sided
			Shapiro-Wilk test = normal & Levene's test: $\sigma_1^2 \neq \sigma_2^2$	Paired t-test with Welch's correction		One or two-sided
			Shapiro-Wilk test ≠ normal & Levene's test: $\sigma_1^2 \neq \sigma_2^2$		Wilcoxon matched-pairs signed-rank test	One or two-sided
3	Outcome variable across two groups	Independent samples	Shapiro-Wilk test = normal & Levene's test: $\sigma_1^2 = \sigma_2^2$	Two-sample t-test		One or two-sided
			Shapiro-Wilk test ≠ normal & Levene's test: $\sigma_1^2 = \sigma_2^2$	Two-sample t-test[b]		One or two-sided
			Shapiro-Wilk test = normal & Levene's test: $\sigma_1^2 \neq \sigma_2^2$	Two-sample t-test with Welch's correction		One or two-sided
			Shapiro-Wilk test ≠ normal & Levene's test: $\sigma_1^2 \neq \sigma_2^2$		Wilcoxon rank-sum test	One or two-sided
4	Outcome variable across three or more groups	One factor variable, independent samples	Shapiro-Wilk test = normal & Levene's test: $\sigma_1^2 = \sigma_2^2$	One-way ANOVA: F-test		Two-sided*[c]

(continued)

Table 6.2 (continued)

Test #	Testing situation	Nature of samples	Choice of Test[a]			Region of rejection
	What do we compare	*Paired vs. Independent*	*Assumptions*	*Parametric*	*Non-parametric*	*One or two-sided test*
			Shapiro-Wilk test $\neq$ normal & Levene's test: $\sigma_1^2 = \sigma_2^2$	One-way ANOVA: *F*-test		Two-sided*
			Shapiro-Wilk test $=$ normal & Levene's test: $\sigma_1^2 \neq \sigma_2^2$	One-way ANOVA: *F*-test with Welch's correction		Two-sided*
			Shapiro-Wilk test $\neq$ normal & Levene's test: $\sigma_1^2 \neq \sigma_2^2$		Kruskal-Wallis rank test	Two-sided*
5	Outcome variable across three or more groups	Two factor variables, independent samples	Shapiro-Wilk test $=$ normal & Levene's test: $\sigma_1^2 = \sigma_2^2$	Two-way ANOVA: *F*-test		Two-sided*
			Shapiro-Wilk test $\neq$ normal & Levene's test: $\sigma_1^2 = \sigma_2^2$	Two-way ANOVA:*F*-test		Two-sided*
			Shapiro-Wilk test $=$ normal & Levene's test: $\sigma_1^2 \neq \sigma_2^2$	Two-way ANOVA: *F*-test with Welch's correction		Two-sided*
			Shapiro-Wilk test $\neq$ normal & Levene's test: $\sigma_1^2 \neq \sigma_2^2$		Kruskal-Wallis rank test	Two-sided*

[a] Selection applies to sample sizes > 30

[b] If the sample size is less than 30, you should transform your outcome variable—through logarithms, square root or power transformations—and re-run the normality test. Alternatively, if normality remains an issue, you should use a non-parametric test that does not rely on the normality assumption

[c] * = Note that although the underlying alternative hypothesis in ANOVA is two-sided, its *F*-statistic is based on the *F*-distribution, which is right-skewed with extreme values only in the right tail of the distribution

6.3.3.1 Define the Testing Situation

When we test hypotheses, we may find ourselves in one of three situations. First, we can test if we want to compare a group to a hypothetical value (test 1). In our example, this can be a pre-determined target of 45 units to establish whether a promotion campaign has been effective or not. Second, we may want to compare the outcome variable (e.g., sales) across two groups (tests 2 or 3). Third, we may wish to compare whether the outcome variable differs between three or more levels of a categorical variable (also called a **factor variable**) with three or more sub-groups (tests 4 or 5). The factor variable is the categorical variable that we use to define the groups (e.g., three types of promotion campaigns). Similarly, we may have situations in which we have two or more factor variables (e.g., three types of promotion campaigns for two types of service). Each of these situations leads to different tests. When assessing the testing situation, we also need to establish the nature of the dependent variable and whether it is measured on an interval or ratio scale. This is important, because parametric tests are based on the assumption that the dependent variable is measured on an interval or ratio scale. Note that we only discuss situations when the test variable is interval or ratio-scaled (see Chap. 3).

6.3.3.2 Determine If Samples Are Paired or Independent

Next, we need to establish whether we compare *paired samples* or *independent samples*. The rule of thumb for determining the samples' nature is to ask if a respondent (or object) was sampled once or multiple times. If a respondent was sampled only once, this means that the values of one sample reveal no information about the values of the other sample. If we sample the same respondent or object twice, it means that the reported values in one period may affect the values of the sample in the next period.[1] Ignoring the "nested" nature of the data increases the probability of type I errors. We therefore need to understand the nature of our samples in order to select a test that takes the dependency between observations (i.e., paired versus independent samples tests) into account. In Table 6.2, test 2 deals with paired samples, whereas tests 3, 4 and, 5 deal with independent samples.

6.3.3.3 Check Assumptions and Choose the Test

Subsequently, we need to check the distributional properties and variation of our data before deciding whether to select a parametric or a non-parametric test.

[1] In experimental studies, if respondents were paired with others (as in a matched case control sample), each person would be sampled once, but it still would be a paired sample.

Normality Test

To test whether the data are normally distributed, we conduct the **Shapiro-Wilk test** (Shapiro and Wilk 1965) that formally tests for normality. Without going into too much detail, the Shapiro-Wilk test compares the correlation between the observed sample scores (which take the covariance between the sample scores into account) with the scores expected under a standard normal distribution. The resulting ratio is called the W-statistic and its scaled version is known as the V-statistic. When the distribution is close to normal, the V statistic will be close to 1 and the associated *p*-value will be larger than 0.05. Large deviations from a V of 1 will therefore be coupled with *p*-values that are smaller than 0.05, suggesting that the sample scores are not normally distributed. The Kolmogorov-Smirnov test, which we discuss in Box 6.2, is another popular test to check for normality. An alternative strategy to check for normality is by means of visual inspection, which we discuss in Box 6.3.

Equality of Variances Test

We use **Levene's test** (Levene 1960), also known as the ***F*-test of sample variance**, to test for the equality of the variances between two or more groups of data. The null hypothesis is that population variances across the sub-samples are the same, whereas the alternative hypothesis is that they differ. If the *p*-value associated with Levene's statistic (referred to as W0 in Stata) is lower than 0.05, we reject the null hypothesis, which implies that the variances are heterogeneous. Conversely, a *p*-value larger than 0.05 indicates homogeneous variances. In Levene's original paper, the formula for the test statistic is based on the sample mean (Levene

Box 6.2 The Kolmogorov-Smirnov Test

An important (nonparametric) test for normality is the *one-sample Kolmogorov–Smirnov (KS) test*. We can use it to test whether or not a variable is normally distributed. Technically, when assuming a normal distribution, the KS test compares the sample scores with an artificial set of normally distributed scores that has the same mean and standard deviation as the sample data. However, this approach is known to yield biased results, which are modified by means of the Lilliefors correction (1967). The Lilliefors correction takes into consideration that we do not know the true mean and standard deviation of the population. An issue with the KS test with the Lilliefors correction is that it is very sensitive when used on large samples and often rejects the null hypothesis if very small deviations are present. This also holds for Stata's version of the KS test, which only works well for very large sample sizes (i.e., at least 10,000 observations). Consequently, Stata does not recommend the use of a one-sample KS test (for more, read the information in Stata's help file on the KS test: https://www.stata.com/manuals14/rksmirnov.pdf).

Box 6.3 Visual Check for Normality

You can also use plots to visually check for normality. The *normal probability plot* visually contrasts the probability distribution of the test variable's ordered sample values with the cumulative probabilities of a standard normal distribution. The probability plots are structured such that the cumulative frequency distribution of a set of normally distributed data falls in a straight line. This straight line serves as a reference line, meaning that sample values' deviations from this straight line indicate departures from normality (Everitt and Skrondal 2010). The *quantile plot* is another type of probability plot, which differs from the former by comparing the quantiles (Chap. 5) of the sorted sample values with the quantiles of a standard normal distribution. Here, again, the plotted data that do not follow the straight line reveal departures from normality. The normal probability plot assesses the normality of the data in the middle of the distribution well. The quantile plot is better equipped to spot non-normality in the tails. Of the two, the quantile plots are most frequently used. Note that visual checks are fairly subjective and should always be used in combination with more formal checks for normality.

1960), which performs well when the variances are symmetric and have moderate tailed distributions. For skewed data, Brown and Forsythe (1974) proposed a modification of Levene's test whereby the group sample's median (referred to as W50 in Stata) replaces the group sample mean. Alternatively, the group sample's trimmed mean can also replace the group sample mean, whereby 10% of the smallest and largest values of the sample are removed before calculating the mean (referred to as W10 in Stata). If the data are normally distributed, the *p*-values associated with W0, W10, and W50 will all align, thus all be higher than 0.05. If the data are not normally distributed, this might not be the case. In this situation, we should focus on the *p*-values associated with the sample's median (W50), because this measure is robust for samples that are not normally distributed.

Parametric Tests

It is clear-cut that when the normality assumption is met, we should choose a *parametric test*. The most popular parametric test for examining one or two means is the ***t*-test**, which can be used for different purposes. For example, the *t-test* can be used to compare one mean with a given value (e.g., do males spend more than $150 a year online?). The **one-sample *t*-test** is an appropriate test. Alternatively, we can use a *t*-test to test the mean difference between two samples (e.g., do males spend more time online than females?). In this case, a **two-sample *t*-test** is appropriate. **The independent samples *t*-tests** considers two distinct groups, such as males versus females, or users versus non-users. Conversely, **the paired samples *t*-test** relates to the same set of twice observed objects (usually respondents), as in a before-after experimental design discussed in Chap. 4. We are, however, often interested in examining the differences between the means of more than two groups of respondents. Regarding our introductory example, we might be interested in evaluating the differences between the point of sale display, the free tasting stand, and the in-store announcements' mean sales. Instead of making several paired comparisons by means of separate *t*-tests, we should use the **Analysis of Variance (ANOVA)**. The ANOVA is useful when three or more means are compared and, depending on how many variables define the groups to be compared (will be discussed later in this chapter), can come in different forms.

The parametric tests introduced in this chapter are very robust against normality assumption violations, especially when the data are distributed symmetrically. That is, small departures from normality usually translate into marginal differences in the *p*-values, particularly when using sample sizes greater than 30 (Boneau 1960). Thus, even if the Shapiro–Wilk test suggests the data are not normally distributed, we don't have to be concerned that the parametric test results are far off, provided we have sample sizes greater than 30. The same holds for the ANOVA in cases where the sample sizes per group exceed 30.

In sum, with sample sizes greater than 30, we choose a parametric test even when the Shapiro-Wilk test suggests that the data are not normally distributed. When sample sizes are less than 30, we can transform the distribution of the outcome variable—through logarithms, square root, or power transformations (Chap. 5)—so that it approaches normality and re-run a normality test. If violations of normality remain an issue, you should use a non-parametric test that is not based on the normality assumption.

Next, we can also have a situation in which the data are normally distributed, but the variances between two or more groups of data are unequal. This issue is generally unproblematic as long as the group-specific sample sizes are (nearly) equal. If group-specific sample sizes are different, we recommend using parametric tests, such as the two-sample *t*-tests and the ANOVA, in combination with tests that withstand or correct the lack of equal group variances, such as **Welch's correction**. Welch's modified test statistic (Welch 1951) adjusts the underlying parametric tests if the variances are not homogenous in order to control for a type I error. This is particularly valuable when population variances differ and groups comprise very unequal sample sizes.[2] In sum, when samples are normally distributed, but the equality of the variance assumption is violated (i.e., the outcome variable is not distributed equally across three or more groups), we choose a parametric test with Welch's correction. Depending on the testing situation this can be: a paired *t*-test with Welch's correction, a one-way ANOVA *F*-test with Welch's correction, or a two-way ANOVA *F*-test with Welch's correction.

[2] Stata does not directly support Welch's correction for an ANOVA, but a user-written package called `wtest` is readily available and can be installed (see Chap. 5 on how to install user-written packages in Stata). This allows you to perform a test similar to the standard ANOVA test with Welch's correction. For more information see Stata's help file: http://www.ats.ucla.edu/stat/stata/ado/analysis/wtest.hlp

Finally, where both the normality and equality of variance assumptions are violated, non-parametric tests can be chosen directly. In the following, we briefly discuss these non-parametric tests.

Non-parametric Tests

As indicated in Table 6.2, there is a non-parametric equivalent for each parametric test. This would be important if the distributions are not symmetric. For single samples, the **Wilcoxon signed-rank test** is the equivalent of one sample *t*-test, which is used to test the hypothesis that the population median is equal to a fixed value. For two-group comparisons with independent samples, the **Mann-Whitney U test** (also called the *Wilcoxon rank-sum test*, or *Wilcoxon–Mann–Whitney test*) is the equivalent of the independent *t*-test, while, for paired samples, this is the *Wilcoxon matched-pairs signed-rank test*. The Mann-Whitney U test uses the null hypothesis that the distributions of the two independent groups being considered (e.g., randomly assigned high and low performing stores) have the same shape (Mann and Whitney 1947). In contrast to an independent sample *t*-test, the Mann-Whitney U test does not compare the means, but the two groups' median scores. Although we will not delve into the statistics behind the test, it is important to understand its logic. The Mann-Whitney U test is based on ranks and measures the differences in location (Liao 2002). The test works by first combining the separate groups into a single group. Subsequently, each outcome variable score (e.g., sales) is sorted and ranked in respect of each condition based on the values, with the lowest rank assigned to the smallest value. The ranks are then averaged based on the conditions (e.g., high versus low performing stores) and the test statistic U calculated. The test statistic represents the difference between the two rank totals. That is, if the distribution of the two groups is identical, then the sum of the ranks in one group will be the same as in the other group. The smaller the *p*-value (which will be discussed later in this chapter), the lower the likelihood that the two distributions' similarities have occurred by chance; the opposite holds if otherwise.

The *Kruskal-Wallis rank test* is the non-parametric equivalent of the ANOVA. The null hypothesis of the *Kruskal-Wallis* rank test is that the distribution of the test variable across group sub-samples is identical (Schuyler 2011). Given that the emphasis is on the distribution rather than on a point estimate, rejecting the null hypothesis implies that such distributions vary in their dispersion, central tendency and/or variability. According to Schuyler (2011) and Liao (2002), the following are the steps when conducting this test: First, single group categories are combined into one group with various categories. Next, objects in this variable (e.g., stores/campaigns) are sorted and ranked based on their associations with the dependent variable (e.g., sales), with the lowest rank assigned to the smallest value. Subsequently, the categorical variable is subdivided to reestablish the original single comparison groups. Finally, each group's sum of its ranks is entered into a formula that yields the calculated test statistic. If this calculated statistic is higher than the critical value, the null hypothesis is rejected. The test statistic of the Kruskal-Wallis rank follows a χ^2 distribution with $k - 1$ degrees of freedom. Use the *Kruskal-Wallis*

H test in situations where the group variances are not equal, as it corrects group variances' heterogeneity.

6.3.3.4 Specify the Region of Rejection

Finally, depending on the formulated hypothesis (i.e., directional versus non-directional), we should decide on whether the region of rejection is on one side (i.e., **one-tailed test**) or on both sides (i.e., **two-tailed test**) of the sampling distribution. In statistical significance testing, a one-tailed test and a two-tailed test are alternative ways of computing the statistical significance of a test statistic, depending on whether the hypothesis is expressed directionally (i.e., $<$ or $>$ in case of a one-tailed test) or not (i.e., $\neq$ in case of a two-tailed test). The word tail is used, because the extremes of distributions are often small, as in the normal distribution or bell curve shown in Fig. 6.3 later in this chapter. Instead of the word tail, the word "sided" is sometimes used.

We need to use two-tailed tests for non-directional hypotheses. Even when directional hypotheses are used, two-tailed tests are used for 75% of directional hypotheses (van Belle 2008). This is because two-tailed tests have strong advantages; they are stricter (and therefore generally considered more appropriate) and can also reject a hypothesis when the effect is in an unexpected direction. The use of two-tailed testing for a directional hypothesis is also valuable, as it identifies significant effects that occur in the opposite direction from the one anticipated. Imagine that you have developed an advertising campaign that you believe is an improvement on an existing campaign. You wish to maximize your ability to detect the improvement and opt for a one-tailed test. In doing so, you do not test for the possibility that the new campaign is significantly less effective than the old campaign. As discussed in various studies (van Belle 2008; Ruxton and Neuhaeuser 2010), one-tailed tests should only be used when the opposite direction is theoretically meaningless or impossible (Kimmel 1957; Field 2013). Such an example would apply to controlled experiments where the intervention (i.e., the drug) can only have a positive and no negative outcome, because such differences are removed beforehand and have no possible meaning (e.g., Lichters et al. 2016). The use of two-tailed tests can seem counter to the idea of hypothesis testing, because two-tailed tests, by their very nature, do not reflect any directionality in a hypothesis. However, in many situations when we have clear expectations (e.g., sales are likely to increase), the opposite is also a possibility.

6.3.4 Step 4: Calculate the Test Statistic

Having formulated the study's main hypothesis, the significance level, and the type of test, we can now proceed with calculating the test statistic by using the sample data at hand. The **test statistic** is a statistic, calculated by using the sample data, to assess the strength of evidence in support of the null hypothesis (Agresti and Finlay 2014). In our example, we want to compare the mean with a given standard of

45 units. Hence, we make use of a *one-sample t-test*, whose test statistic is computed as follows:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

Here $\bar{x}$ is the sample mean, μ is the hypothesized population mean, and $s_{\bar{x}}$ the standard error (i.e., the standard deviation of the sampling distribution). Let's first look at the formula's numerator, which describes the difference between the sample mean $\bar{x}$ and the hypothesized population mean μ. If the point of sale display was highly successful, we would expect $\bar{x}$ to be higher than μ, leading to a positive difference between the two in the formula's numerator. Alternatively, if the point of sale display was not effective, we would expect the opposite to be true. This means that the difference between the hypothesized population mean and the sample mean can go either way, implying a two-sided test. Using the data from the second column of Table 6.1, we can compute the marginal mean of the point of sales display campaign as follows:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(50 + 52 + \ldots + 51 + 44) = 47.30$$

When comparing the calculated sample mean (47.30) with the hypothesized value of 45, we obtain a difference of 2.30:

$$\bar{x} - \mu = 47.30 - 45 = 2.30$$

At first sight, it appears as if the campaign was effective as sales during the time of the campaign were higher than those that the store normally experiences. However, as discussed before, we have not yet considered the variation in the sample. This variation is accounted for by the *standard error* of $\bar{x}$ (indicated as $s_{\bar{x}}$), which represents the uncertainty of the sample estimate.

This sounds very abstract, so what does it mean? The sample mean is used as an estimator of the population mean; that is, we assume that the sample mean can be a substitute for the population mean. However, when drawing different samples from the same population, we are likely to obtain different sample means. The standard error tells us how much variance there probably is in the mean across different samples from the same population.

Why do we have to divide the difference $\bar{x} - \mu$ by the standard error $s_{\bar{x}}$? We do so, because when the standard error is very low (there is a low level of variation or uncertainty in the data), the value in the test statistic's denominator is also small, which results in a higher value for the *t*-test statistic. Higher *t*-values favor the rejection of the null hypothesis. In other words, the lower the standard error $s_{\bar{x}}$, the greater the probability that the population represented by the sample truly differs from the hypothesized value of 45.

But how do we compute the standard error? We do so by dividing the sample standard deviation (s) by the square root of the number of observations (n), as follows:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}}{\sqrt{n}}$$

As we can see, a low standard deviation s decreases the standard error (which means less ambiguity when making inferences from these data). That is, less variation in the data decreases the standard error, thus favoring the rejection of the null hypothesis. Note that the standard error also depends on the sample size n. By increasing the number of observations, we have more information available, thus reducing the standard error.

If you understand this basic principle, you will have no problems understanding most other statistical tests. Let's go back to the example and compute the standard error as follows:

$$s_{\bar{x}} = \frac{\sqrt{\frac{1}{10-1}\left[(50-47.30)^2 + \ldots + (44-47.30)^2\right]}}{\sqrt{10}} = \frac{3.199}{\sqrt{10}} \approx 1.012$$

Thus, the result of the test statistic is:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{2.30}{1.012} \approx 2.274$$

This test statistic applies when we compute a sample's standard deviation. In some situations, however, we might know the population's standard deviation, which requires the use of a different test, the **z-test**, (see Box 6.4).

Box 6.4 The *z*-Test

In the previous example, we used sample data to calculate the standard error $s_{\bar{x}}$. If we know the population's standard deviation beforehand, we should use the z-test. The z-test follows a normal (instead of a t-distribution).[3] The z-test is also used in situations when the sample size exceeds 30, because the t-distribution and normal distribution are similar for $n > 30$. As the t-test is slightly more accurate (also when the sample size is greater than 30), Stata uses the t-test, which can be accessed by going to ► Statistics ► Summaries, tables, and tests ► Classical tests of hypotheses ► t test (mean-comparison test). We do not show the formulas associated with the z-test here, but have included these in the ⭳ Web Appendix (→ Downloads).

[3] The fundamental difference between the z- and t-distributions is that the t-distribution is dependent on sample size n (which the z-distribution is not). The distributions become more similar with larger values of n.

6.3.5 Step 5: Make the Test Decision

Once we have calculated the test statistic, we can decide how likely it is that the claim stated in the hypothesis is correct. This is done by comparing the test statistic with the critical value that it must exceed (*Option 1*). Alternatively, we can calculate the actual probability of making a mistake when rejecting the null hypothesis and compare this value with the significance level (*Option 2*). In the following, we discuss both options.

6.3.5.1 Option 1: Compare the Test Statistic with the Critical Value

To make a test decision, we must first determine the critical value, which the test statistic must exceed for the null hypothesis to be rejected. In our case, the critical value comes from a *t*-distribution and depends on three parameters:

1. the significance level,
2. the degrees of freedom, and
3. one-tailed versus two-tailed testing.

We have already discussed the first point, so let's focus on the second. The **degrees of freedom** (usually abbreviated as *df*) represent the amount of information available to estimate the test statistic. In general terms, an estimate's degrees of freedom are equal to the amount of independent information used (i.e., the number of observations) minus the number of parameters estimated. Field (2013) provides a great explanation, which we adapted and present in Box 6.5.

In our example, we count $n - 1$ or $10 - 1 = 9$ degrees of freedom for the *t*-statistic to test a two-sided hypothesis of one mean. Remember that for a two-tailed test,

Box 6.5 Degrees of Freedom

Suppose you have a soccer team and 11 slots on the playing field. When the first player arrives, you have the choice of 11 positions in which you can place him or her. By allocating the player to a position, this occupies one position. When the next player arrives, you can choose from 10 positions. With every additional player who arrives, you have fewer choices where to position him or her. With the very last player, you no longer have the freedom to choose where to put him or her—there is only one spot left. Thus, there are 10 degrees of freedom. You have some degree of choice with 10 players, but for 1 player you don't. The degrees of freedom are the number of players minus 1.

when α is 0.05, the cumulative probability distribution is $1 - \alpha/2$ or $1 - 0.05/2 = 0.975$. We divide the significance level by two, because half of our alpha tests the statistical significance in the lower tail of the distribution (bottom 2.5%) and half in the upper tail of the distribution (top 2.5%). If the value of the test statistic is greater than the critical value, we can reject the $H_{0.}$

We can find critical values for combinations of significance levels and degrees of freedom in the *t*-distribution table, shown in Table A1 in the ⭳ Web Appendix (→ Downloads). For 9 degrees of freedom and using a significance level of, for example, 5%, the critical value of the *t*-statistic is 2.262. Remember that we have to look at the $\alpha = 0.05/2 = 0.025$ column, because we use a two-tailed test. This means that for the probability of a type I error (i.e., falsely rejecting the null hypothesis) to be less than or equal to 5%, the value of the test statistic must be 2.262 or greater. In our case, the test statistic (2.274) exceeds the critical value (2.262), which suggests that we should reject the null hypothesis.[4] Even though the difference between the values is very small, bear in mind that hypothesis testing is binary—we either reject or don't reject the null hypothesis. This is also the reason why a statement such as "the result is highly significant" is inappropriate.

Figure 6.3 summarizes this concept graphically. In this figure, you can see that the critical value $t_{critical}$ for an α-level of 5% with 9 degrees of freedoms equals $\pm$ 2.262 on both sides of the distribution. This is indicated by the two dark-shaded rejection regions in the upper 2.5% and bottom 2.5% and the remaining white 95% non-rejection region in the middle. Since the test statistic t_{test} (indicated by the dotted line) falls in the dark-shaded area, we reject the null hypothesis.

Table 6.3 summarizes the decision rules for rejecting the null hypothesis for different types of *t*-tests, where t_{test} describes the test statistic and $t_{critical}$ the critical value for a specific significance level α. Depending on the test's formulation, test values may well be negative (e.g., −2.262). However, due to the symmetry of the *t*-distribution, only positive critical values are displayed.

6.3.5.2 Option 2: Compare the *p*-Value with the Significance Level

The above might make you remember your introductory statistics course with horror. The good news is that we do not have to bother with statistical tables when working with Stata. Stata automatically calculates the probability of obtaining a test statistic that is at least as extreme as the actually observed one if the null hypothesis is supported. This probability is also referred to as the ***p*-value** or the probability of observing a more extreme departure from the null hypothesis (Everitt and Skrondal 2010).

[4] To obtain the critical value, write `display invt(9,1-0.05/2)` in the command window.

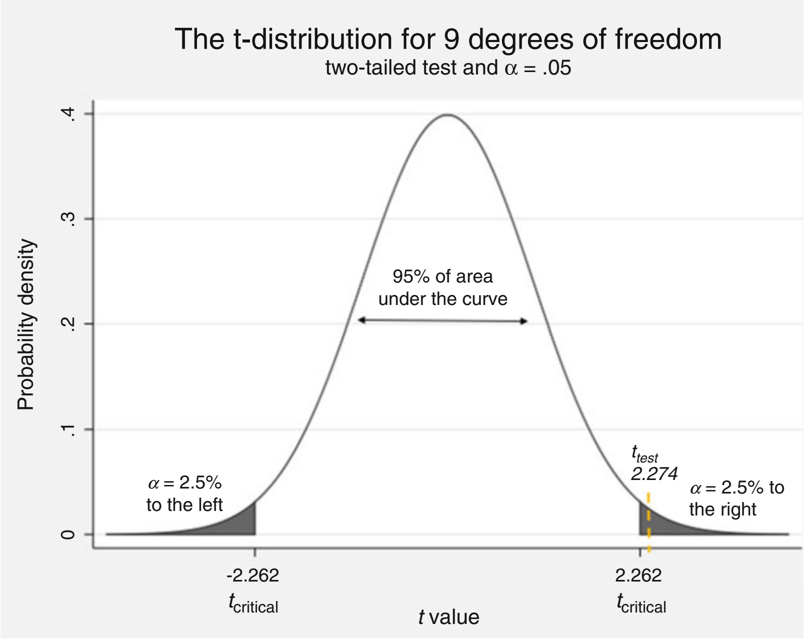


Fig. 6.3 Relationship between test value, critical value, and *p*-value

Table 6.3 Decision rules for testing decisions

Type of test	Null hypothesis (H_0)	Alternative hypothesis (H_1)	Reject H_0 if
Right-tailed test	$\mu \leq$ value	$\mu >$ value	$\lvert t_{test} \rvert > t_{critical}\ (\alpha)$
Left-tailed test	$\mu \geq$ value	$\mu <$ value	$\lvert t_{test} \rvert > t_{critical}\ (\alpha)$
Two-tailed test	$\mu =$ value	$\mu \neq$ value	$\lvert t_{test} \rvert > t_{critical}\ \left(\frac{\alpha}{2}\right)$

In the previous example, the *p*-value is the answer to the following question: If the population mean is not equal to 45 (i.e., therefore, the null hypothesis holds), what is the probability that random sampling could lead to a test statistic value of at least ± 2.274? This description shows that there is a relationship between the *p*-value and the test statistic. More precisely, these two measures are inversely related; the higher the absolute value of the test statistic, the lower the *p*-value and vice versa (see Fig. 6.3).

The description of the p-value is similar to the significance level α, which describes the acceptable probability of rejecting a true null hypothesis. However, the difference is that the p-value is calculated using the sample, and that α is set by the researcher before the test outcome is observed.[5] The p-value is not the probability of the null hypothesis being supported! Instead, we should interpret it as evidence against the null hypothesis. The α-level is an arbitrary and subjective value that the researcher assigns to the level of risk of making a type I error; the p-value is calculated from the available data. Related to this subjectivity, there has been a revived discussion in the literature on the usefulness of p-values (e.g., Nuzzo 2014; Wasserstein and Lazar 2016).

The comparison of the p-value and the significance level allows the researcher to decide whether or not to reject the null hypothesis. Specifically, if the p-value is smaller than or equal to the significance level, we reject the null hypothesis. Thus, when examining test results, we should make use of the following decision rule—this should become second nature![6]

– p-value $\leq \alpha \rightarrow$ reject H_0
– p-value $> \alpha \rightarrow$ do not reject H_0

Note that this decision rule applies to two-tailed tests. If you apply a one-tailed test, you need to divide the p-value in half before comparing it to α, leading to the following decision rule[7]:

– p-value/2 $\leq \alpha \rightarrow$ reject H_0
– p-value/2 $> \alpha \rightarrow$ do not reject H_0

In our example, the actual two-tailed p-value is 0.049 for a test statistic of ± 2.274, just at the significance level of 0.05. We can therefore reject the null hypothesis and find support for the alternative hypothesis.[8]

[5] Unfortunately, there is some confusion about the difference between the α and p-value. See Hubbard and Bayarri (2003) for a discussion.

[6] Note that this is convention and most textbooks discuss hypothesis testing in this way. Originally, two testing procedures were developed, one by Neyman and Pearson and another by Fisher (for more details, see Lehmann 1993). Agresti and Finlay (2014) explain the differences between the convention and the two original procedures.

[7] Note that this doesn't apply, for instance, to exact tests for probabilities.

[8] We don't have to conduct manual calculations and tables when working with Stata. However, we can easily compute the p-value ourselves by using the TDIST function in Microsoft Excel. The function has the general form "TDIST(t, df, tails)", where t describes the test value, df the degrees of freedom, and *tails* specifies whether it's a one-tailed test (tails = 1) or two-tailed test (tails = 2). Just open a new spreadsheet for our example and type in "=TDIST(2.274,9,1)". Likewise, there are several webpages with Java-based modules (e.g., http://graphpad.com/quickcalcs/pvalue1.cfm) that calculate p-values and test statistic values.

6.3.6 Step 6: Interpret the Results

The conclusion reached by hypothesis testing must be expressed in terms of the market research problem and the relevant managerial action that should be taken. In our example, we conclude that there is evidence that the point of sale display influenced the number of sales significantly during the week it was installed.

6.4 Two-Samples *t*-Test

6.4.1 Comparing Two Independent Samples

Testing the relationship between two independent samples is very common in market research. Some common research questions are:

- Do heavy and light users' satisfaction with a product differ?
- Do male customers spend more money online than female customers?
- Do US teenagers spend more time on Facebook than Australian teenagers?

Each of these hypotheses aim at evaluating whether two populations (e.g., heavy and light users), represented by samples, are significantly different in terms of certain key variables (e.g., satisfaction ratings).

To understand the principles of the *two independent samples t-test*, let's reconsider the previous example of a promotion campaign in a department store. Specifically, we want to test whether the population mean of the point of sale display's sales (μ_1) differs in any (positive or negative) way from that of the free tasting stand (μ_2). The resulting null and alternative hypotheses are now:

$$\mathrm{H}_0\colon \mu_1 = \mu_2$$

$$\mathrm{H}_1\colon \mu_1 \neq \mu_2$$

The test statistic of the two independent samples *t*-test—which is distributed with $n_1 + n_2 - 2$ degrees of freedom—is similar to the one-sample *t*-test:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}},$$

where $\bar{x}_1$ is the mean of the first sample (with n_1 numbers of observations) and $\bar{x}_2$ is the mean of the second sample (with n_2 numbers of observations). The term $\mu_1 - \mu_2$ describes the hypothesized difference between the population means. In this case, $\mu_1 - \mu_2$ is zero, as we assume that the means are equal, but we could use any other value to hypothesize a specific difference in population means. Lastly, $s_{\bar{x}_1 - \bar{x}_2}$ describes the standard error, which comes in two forms:

– If we assume that the two populations have the same variance (i.e., $\sigma_1^2 = \sigma_2^2$), we compute the standard error based on the so called *pooled* variance estimate:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\left[(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2\right]}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{\mathrm{n}_1} + \frac{1}{\mathrm{n}_2}}$$

– Alternatively, if we assume that the population variances differ (i.e., $\sigma_1^2 \neq \sigma_2^2$), we compute the standard error as follows:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\mathrm{s}_1^2}{\mathrm{n}_1} + \frac{\mathrm{s}_2^2}{\mathrm{n}_2}}$$

How do we determine whether the two populations have the same variance? As discussed previously, this is done using Levene's test, which tests the following hypotheses:

$$\mathrm{H}_0: \sigma_1^2 = \sigma_2^2$$

$$\mathrm{H}_1: \sigma_1^2 \neq \sigma_2^2$$

The null hypothesis is that the two population variances are the same and the alternative hypothesis is that they differ. In this example, Levene's test provides support for the assumption that the variances in the population are equal, which allows us to proceed with the pooled variance estimate. First, we estimate the variances of the first and second group as follows:

$$s_1^2 = \frac{1}{n_1 - 1} \sum\nolimits_{i=1}^{10} (x_1 - \bar{x}_1)^2 = \frac{1}{10 - 1}\left[(50 - 47.30)^2 + \cdots + (44 - 47.30)^2\right] \approx 10.233$$

$$s_2^2 = \frac{1}{n_2 - 1} \sum\nolimits_{i=1}^{10} (x_2 - \bar{x}_2)^2 = \frac{1}{10 - 1}\left[(55 - 52)^2 + \ldots + (44 - 52)^2\right] \approx 17.556.$$

Using the variances as input, we can compute the standard error:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\left[(10 - 1) \cdot 10.233 + (10 - 1) \cdot 17.556\right]}{10 + 10 - 2}} \cdot \sqrt{\frac{1}{10} + \frac{1}{10}} \approx 1.667$$

Inserting the estimated standard error into the test statistic results in:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{(47.30 - 52) - 0}{1.667} \approx -2.819$$

The test statistic follows a *t*-distribution with $n_1 - n_2$ degrees of freedom. In our case, we have 10 + 10 – 2 = 18 degrees of freedom. Looking at the statistical Table A1 in the ⭳ Web Appendix (→ Downloads), we can see that the critical value of a two-sided test with a significance level of 5% is 2.101 (note that we should look at the column labeled $\alpha = 0.025$ and the line labeled $df = 18$). The absolute value of the test statistic (i.e., 2.819) is greater than the critical value of 2.101 and, thus, falls within the bottom 2.5% of the distribution. We can therefore reject the null hypothesis at a significance level of 5% and conclude that the absolute difference between means of the point of sale display's sales (μ_1) and those of the free tasting stand (μ_2) is significantly different from 0.

6.4.2 Comparing Two Paired Samples

In the previous example, we compared the mean sales of two independent samples. If the management wants to compare the difference in the units sold before and after they started the point of sale display campaign. The difference can be either way; that is, it can be higher or lower, indicating a two-sided test. We have sales data for the week before the point of sale display was installed, as well as for the following week when the campaign was in full swing (i.e., the point of sale display had been installed). Table 6.4 shows the sale figures of the 10 stores in respect of both experimental conditions. You can again assume that the data are normally distributed.

At first sight, it appears that the point of sale display generated higher sales: The marginal mean of the sales in the week during which the point of sale display was

Table 6.4 Sales data (extended)

	Sales (units)	
Store	No point of sale display	Point of sale display
1	46	50
2	51	53
3	40	43
4	48	50
5	46	47
6	45	45
7	42	44
8	51	53
9	49	51
10	43	44
Marginal mean	46.10	48

installed (48) is slightly higher than in the week when it was not (46.10). However, the question is whether this difference is statistically significant.

We cannot assume that we are comparing two independent samples, as each set of two samples originates from the same set of stores, but at different points in time and under different conditions. Hence, we should use a *paired samples t-test*. In this example, we want to test whether the sales differ significantly with or without the installation of the point of sale display. We can express this by using the following hypotheses, where μ_d describes the population difference in sales; the null hypothesis assumes that the point of sale display made no difference, while the alternative hypothesis assumes a difference in sales:

$$\mathrm{H}_0\colon \mu_d = 0$$

$$\mathrm{H}_1\colon \mu_d \neq 0$$

To carry out this test, we define a new variable d_i, which captures the differences in sales between the two conditions (point of sale display – no point of sale display) in each of the stores. Thus:

$$\mathrm{d}_1 = 50 - 46 = 4$$

$$\mathrm{d}_2 = 53 - 51 = 2$$

$$\ldots$$

$$\mathrm{d}_9 = 51 - 49 = 2$$

$$\mathrm{d}_{10} = 44 - 43 = 1$$

Based on these results, we calculate the mean difference:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{10} d_i = \frac{1}{10}(4 + 2 + \ldots + 2 + 1) = 1.9$$

as well as the standard error of this difference:

$$s_{\bar{d}} = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{10}\left(d_i - \bar{d}\right)^2}}{\sqrt{n}}$$

$$= \frac{\sqrt{\frac{1}{9}\left[(4-1.9)^2 + (2-1.9)^2 + \ldots + (2-1.9)^2 + (1-1.9)^2\right]}}{\sqrt{10}} \approx 0.383$$

Next, we compare the mean difference $\bar{d}$ in our sample with the difference expected under the null hypothesis μ_d and divide this difference by the standard error $s_{\bar{d}}$. Thus, the test statistic is:

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}} = \frac{1.9 - 0}{0.383} \approx 4.960.$$

The test statistic follows a *t*-distribution with $n - 1$ degrees of freedom, where n is the number of pairs that we compare. Recall that for a two-tailed test, when α is 0.05, we need to look at the column labeled $\alpha = 0.025$ and the line labeled $df = 9$. Looking at Table A1 in the ⭳ Web Appendix (→ Downloads), we can see that the critical value of a two-sided test with a significance level of 5% is 2.262 for 9 degrees of freedom. Since the test value (4.960) is larger than the critical value, we can reject the null hypothesis and presume that the point of sale display makes a difference.

6.5 Comparing More Than Two Means: Analysis of Variance (ANOVA)

Researchers are often interested in examining differences in the means between more than two groups. For example:

- Do light, medium, and heavy internet users differ in respect of their monthly disposable income?
- Do customers across four different types of demographic segments differ in their attitude towards a certain brand?
- Is there a significant difference in hours spent on Facebook between US, UK, and Australian teenagers?

Continuing with our previous example on promotion campaigns, we might be interested in whether there are significant sales differences between the stores in which the three different types of campaigns were launched. One way to tackle this research question would be to carry out multiple pairwise comparisons of all the groups under consideration. In this example, doing so would require the following comparisons:

1. the point of sale display versus the free tasting stand,
2. the point of sale display versus the in-store announcements, and
3. the free tasting stand versus the in-store announcements.

While three comparisons seem manageable, you can imagine the difficulties when a greater number of groups are compared. For example, with ten groups, we would have to carry out 45 group comparisons.[9]

[9] The number of pairwise comparisons is calculated as follows: $k \cdot (k - 1)/2$, with k the number of groups to compare.

Making large numbers of comparisons induces the severe problem of **α-inflation**. This inflation refers to the more tests that you conduct at a certain significance level, the more likely you are to claim a significant result when this is not so (i.e., an increase or inflation in the type I error). Using a significance level of $\alpha = 0.05$ and making all possible pairwise comparisons of ten groups (i.e., 45 comparisons), the increase in the overall probability of a type I error (also referred to as **familywise error rate**) is:

$$\alpha^* = 1 - (1 - \alpha)^{45} = 1 - (1 - 0.05)^{45} = 0.901$$

That is, there is a 90.1% probability of erroneously rejecting your null hypothesis in at least some of your 45 *t*-tests—far greater than the 5% for a single comparison! The problem is that you can never tell which of the comparisons' results are wrong and which are correct.

Instead of carrying out many pairwise tests, market researchers use ANOVA, which allows a comparison of three or more groups' averages. In ANOVA, the variable that differentiates the groups is referred to as the *factor variable* (don't confuse this with the factors of factor analysis discussed in Chap. 8!). The values of a factor (i.e., as found in respect of the different groups under consideration) are also referred to as *factor levels*.

In the previous example of promotion campaigns, we considered only one factor variable with three levels, indicating the type of campaign. This is the simplest form of an ANOVA and is called a one-way ANOVA. However, ANOVA allows us to consider more than one factor variable. For example, we might be interested in adding another grouping variable (e.g., the type of service offered), thus increasing the number of treatment conditions in our experiment. In this case, we should use a two-way ANOVA to analyze both factor variables' effect on the sales (in isolation and jointly). ANOVA is even more flexible, because you can also integrate interval or ratio-scaled independent variables and even multiple dependent variables. We first introduce the one-way ANOVA, followed by a brief discussion of the two-way ANOVA.[10] For a more detailed discussion of the latter, you can turn to the ⭳ Web Appendix (→ Downloads).

6.6 Understanding One-Way ANOVA

We now know ANOVA is used to examine the mean differences between more than two groups. In more formal terms, the objective of the **one-way ANOVA** is to test the null hypothesis that the population means of the groups (defined by the factor variable and its levels) are equal. If we compare three groups, as in the promotion campaign example, the null hypothesis is:

[10] Mitchell (2015) provides a detailed introduction to other ANOVA types, such as the analysis of covariance (ANCOVA).

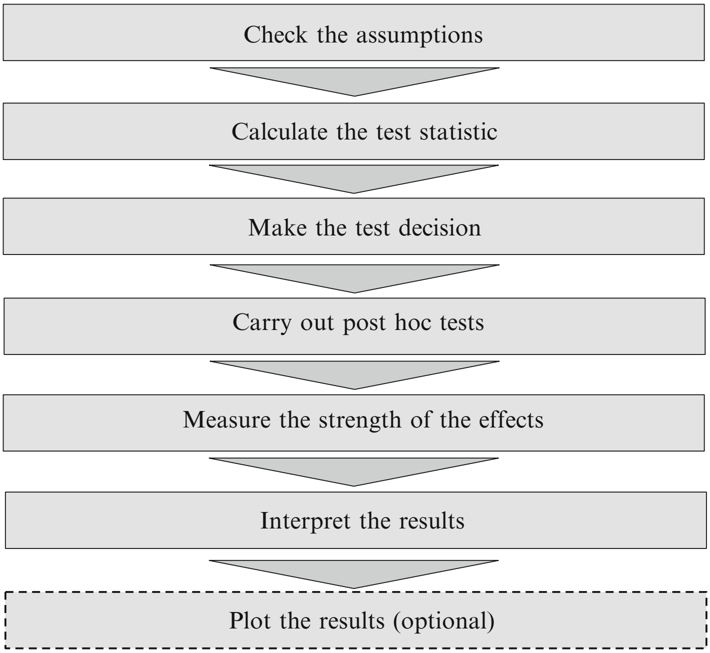

Check the assumptions
Calculate the test statistic
Make the test decision
Carry out post hoc tests
Measure the strength of the effects
Interpret the results
Plot the results (optional)

Fig. 6.4 Steps in conducting an ANOVA

$$H_0\colon \mu_1 = \mu_2 = \mu_3$$

This hypothesis implies that the population means of all three promotion campaigns are identical (which is the same as saying, that the campaigns have the same effect on the mean sales). The alternative hypothesis is:

$$H_1\colon \text{At least two of } \mu_1, \mu_2, \text{and } \mu_3 \text{ are unequal.}$$

Before we even think of running an ANOVA, we should, of course, produce a problem formulation, which requires us to identify the dependent variable and the factor variable, as well as its levels. Once this task is done, we can dig deeper into ANOVA by following the steps described in Fig. 6.4. We next discuss each step in more detail.

6.6.1 Check the Assumptions

ANOVA is a parametric test that relies on the same distributional assumptions as discussed in Sect. 6.3.3.3. We may use ANOVA in situations when the dependent variable is measured on an ordinal scale and is not normally distributed, but we should then ensure that the group-specific sample sizes are similar. Thus, if possible, it is useful to collect samples of a similar size for each group. As discussed previously, ANOVA is robust to departures from normality with sample sizes greater than 30, meaning that it can be performed even when the data are not normally distributed. Even though ANOVA is rather robust in this respect, violations of the assumption of the equality of variances can bias the results significantly, especially when the groups are of very unequal sample size.[11] Consequently, we should always test for the equality of variances by using Levene's test. We already touched upon Levene's test and you can learn more about it in ⭳ Web Appendix (→ Chap. 6).

Finally, like any data analysis technique, the sample size must be sufficiently high to have sufficient statistical power. There is general agreement that the bare minimum sample size per group is 20. However, 30 or more observations per group are desirable. For more detail, see Box 6.1.

6.6.2 Calculate the Test Statistic

ANOVA examines the dependent variable's variation across groups and, based on this variation, determines whether there is reason to believe that the population means of the groups differ. Returning to our example, each store's sales are likely to deviate from the overall sales mean, as there will always be some variation. The question is therefore whether a specific promotion campaign is likely to cause the difference between each store's sales and the overall sales mean, or whether this is due to a natural variation in sales. To disentangle the effect of the treatment (i.e., the promotion campaign type) and the natural variation, ANOVA separates the total variation in the data (indicated by SS_T) into two parts:

1. the between-group variation (SS_B), and
2. the within-group variation (SS_W).[12]

These three types of variation are estimates of the population variation. Conceptually, the relationship between the three types of variation is expressed as:

$$SS_T = SS_B + SS_W$$

However, before we get into the math, let's see what SS_B and SS_W are all about.

6.6.2.1 The Between-Group Variation (SS_B)

SS_B refers to the variation in the dependent variable as expressed in the variation in the group means. In our example, it describes the variation in the mean values of sales across the three treatment conditions (i.e., point of sale display, free tasting

[11] In fact, these two assumptions are interrelated, since unequal group sample sizes result in a greater probability that we will violate the homogeneity assumption.

[12] SS is an abbreviation of "sum of squares," because the variation is calculated using the squared differences between different types of values.

Box 6.6 Types of Sums of Squares in Stata

Stata allows two options to represent a model's sums of squares. The first, and also the default option, is the *partial sums of squares*. To illustrate what this measure represents, presume we have a model with one independent variable var_1 and we want to know what additional portion of our model's variation is explained if we add independent variable var_2, to the model. The partial sums of squares indicates the portion of the variation that is explained by var_2, given var_1. The second option is the *sequential sums of squares*, which adds variables one at a time to the model in order to assess the model's incremental improvement with each newly added variable. Of the two options, the partial sums of squares is the simpler one and does not rely on the ordering of the variables in the model, but is not suitable for full factorial designs that include interactions between two variables.

stand, and in-store announcements) in relation to the overall mean. What does SS_B tell us? Imagine a situation in which all mean values across the treatment conditions are the same. In other words, regardless of which campaign we choose, the sales are the same, we cannot claim that the promotion campaigns had differing effects. This is what SS_B expresses: it tells us how much variation the differences in observations that truly stem from different groups can explain (for more, see Box 6.6). Since SS_B is the *explained variation* (explained by the grouping of data) and thus reflects different effects, we would want it to be as high as possible. However, there is no given standard of how high SS_B should be, as its magnitude depends on the scale level used (e.g., are we looking at 7-point Likert scales or income in US$?). Consequently, we can only interpret the explained variation expressed by SS_B in relation to the variation that the grouping of data does not explain. This is where SS_W comes into play.

6.6.2.2 The Within-Group Variation (SS_W)

As the name already suggests, SS_W describes the variation in the dependent variable within each of the groups. In our example, SS_W simply represents the variation in sales in each of the three treatment conditions. The smaller the variation within the groups, the greater the probability that the grouping of data can explain all the observed variation. It is obviously the ideal for this variation to be as small as possible. If there is much variation within some or all the groups, then some extraneous factor, not accounted for in the experiment, seems to cause this variation instead of the grouping of data. For this reason, SS_W is also referred to as *unexplained variation*.

Unexplained variation can occur if we fail to account for important factors in our experimental design. For example, in some of the stores, the product might have been sold through self-service, while personal service was available in others. This is a factor that we have not yet considered in our analysis, but which will be used when we look at the two-way ANOVA later in the chapter. Nevertheless, some unexplained variation will always be present, regardless of how sophisticated our experimental design is and how many factors we consider. If the unexplained variation cannot be explained, it is called *random noise* or simply *noise*.

6.6.2.3 Combining SS_B and SS_W into an Overall Picture

The comparison of SS_B and SS_W tells us whether the variation in the data is attributable to the grouping, which is desirable, or due to sources of variation not captured by the grouping, which is not desirable. Figure 6.5 shows this relationship across the stores featuring our three different campaign types:

– point of sale display (•),
– free tasting stand (▪), and
– in-store announcements (▲).

We indicate the group mean of each level by dashed lines. If the group means are all the same, the three dashed lines are horizontally aligned and we then conclude that the campaigns have identical sales. Alternatively, if the dashed lines are very different, we conclude that the campaigns differ in their sales.

At the same time, we would like the variation within each of the groups to be as small as possible; that is, the vertical lines connecting the observations and the dashed lines should be short. In the most extreme case, all observations would lie on

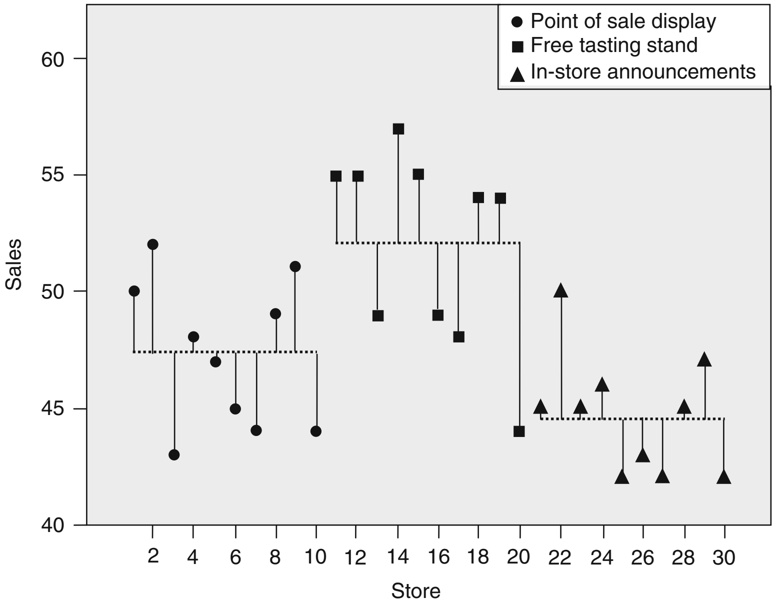


Fig. 6.5 Scatter plot of stores with different campaigns vs. sales

their group-specific dashed lines, implying that the grouping explains the variation in sales perfectly. This, however, hardly ever occurs.

If the vertical bars were all, say, twice as long, it would be difficult to draw any conclusions about the effects of the different campaigns. Too great a variation within the groups then swamps the variation between the groups. Based on the discussion above, we can calculate the three types of variation.

1. The total variation, computed by comparing each store's sales with the overall mean, which is equal to 48 in our example:[13]

$$\begin{aligned} SS_T &= \sum\nolimits_{i=1}^{n} (x_i - \bar{x})^2 \\ &= (50-48)^2 + (52-48)^2 \ + \ldots + (47-48)^2 \ + (42-48)^2 \ = 584 \end{aligned}$$

2. The between-group variation, computed by comparing each group's mean sales with the overall mean, is:

$$SS_B = \sum\nolimits_{j=1}^{k} n_j \left(\bar{x}_j - \bar{x}\right)^2$$

As you can see, besides index i, as previously discussed, we also have index j to represent the group sales means. Thus, $\bar{x}_j$ describes the mean in the j-th group and n_j the number of observations in that group. The overall number of groups is denoted with k. The term n_j is used as a weighting factor: Groups that have many observations should be accounted for to a higher degree relative to groups with fewer observations. Returning to our example, the between-group variation is then given by:

$$SS_B = 10 \cdot (47.30 - 48)^2 + 10 \cdot (52 - 48)^2 + 10 \cdot (44.70 - 48)^2 = 273.80$$

3. The within-group variation, computed by comparing each store's sales with its group sales mean is:

$$SS_w = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

[13] Note that the group-specific sample size in this example is too small to draw conclusions and is only used to show the calculation of the statistics.

Here, we should use two summation signs, because we want to compute the squared differences between each store's sales and its group sales' mean for all k groups in our set-up. In our example, this yields the following:

$$\begin{aligned} SS_W = & \left[(50-47.30)^2+\ldots+(44-47.30)^2\right] \\ & +\left[(55-52)^2+\ldots+(44-52)^2\right] \\ & +\left[(45-44.70)^2+\ldots+(42-44.70)^2\right] \\ = & \ 310.20 \end{aligned}$$

In the previous steps, we discussed the comparison of the between-group and within-group variation. The higher the between-group variation is in relation to the within-group variation, the more likely it is that the grouping of the data is responsible for the different levels in the stores' sales instead of the natural variation in all the sales.

A suitable way to describe this relation is by forming an index with SS_B in the numerator and SS_W in the denominator. However, we do not use SS_B and SS_W directly, because they are based on summed values and the scaling of the variables used therefore influence them. Therefore, we divide the values of SS_B and SS_W by their degrees of freedom to obtain the true mean square values MS_B (called *between-group mean squares*) and MS_W (called *within-group mean squares*). The resulting mean square values are:

$$MS_B=\frac{SS_B}{k-1}, \text{and } MS_W=\frac{SS_W}{n-k}$$

We use these mean squares to compute the following test statistic, which we then compare with the critical value:

$$F=\frac{MS_B}{MS_W}$$

Turning back to our example, we calculate the test statistic as follows:

$$F=\frac{MS_B}{MS_W}=\frac{SS_B/_{k-1}}{SS_W/_{n-k}}=\frac{273.80/_{3-1}}{310.20/_{30-3}}\approx 11.916$$

6.6.3 Make the Test Decision

Making the test decision in ANOVA is like the t-tests discussed earlier, with the difference that the test statistic follows an F-distribution (as opposed to a t-distribution). Different from before, however, we don't have to divide α by 2 when looking up the critical value, even though the underlying alternative hypothesis in ANOVA is two-sided! The reason for this is that an F-test statistic is the ratio of the variation explained by systematic variance (i.e., between-group mean squares) to the unsystematic variance (i.e., within-group mean squares), which is always equal to or greater than 0, but never lower than 0. For this reason, and given that no negative values can be taken, it makes no sense to split the significance level in half, although you can always choose a more restrictive alpha (van Belle 2008).

Unlike the t-distribution, the F-distribution depends on two degrees of freedom: One corresponding to the between-group mean squares ($k - 1$) and the other referring to the within-group mean squares ($n - k$). The degrees of freedom of the promotion campaign example are 2 and 27; therefore, on examining Table A2 in the ⭳ Web Appendix (→ Downloads), we see a critical value of 3.354 for $\alpha = 0.05$. In our example, we reject the null hypothesis, because the F-test statistic of 11.916 is greater than the critical value of 3.354. Consequently, we can conclude that at least two of the population sales means of the three types of promotion campaigns differ significantly.

At first sight, it appears that the free tasting stand is most successful, as it exhibits the highest mean sales ($\bar{x}_2 = 52$) compared to the point of sale display ($\bar{x}_1 = 47.30$) and the in-store announcements ($\bar{x}_3 = 44.70$). However, rejecting the null hypothesis does not mean that all the population means differ—it only means that at least two of the population means differ significantly! Market researchers often assume that all means differ significantly when interpreting ANOVA results, but this is wrong. How then do we determine which of the mean values differ significantly from the others? We deal with this problem by using post hoc tests, which is done in the next step of the analysis.

6.6.4 Carry Out Post Hoc Tests

Post hoc tests perform multiple comparison tests on each pair of groups and tests which of the groups differ significantly from each other. The basic idea underlying post hoc tests is to perform tests on each pair of groups and to correct the level of significance of each test. This way, the overall type I error rate across all the comparisons (i.e., the *familywise error rate*) remains constant at a certain level, such as at $\alpha = 0.05$ (i.e., α-inflation is avoided).

There are several post hoc tests, the easiest of which is the **Bonferroni correction.** This correction maintains the familywise error rate by calculating a new pairwise alpha that divides the statistical significance level of α by the number of comparisons made. How does this correction work? In our example, we can compare three groups pairwise: (1) Point of sale display vs. free tasting stand, (2) point of sale display vs. in-store announcements, and (3) free tasting stand vs. in-store announcements. Hence, we would use $0.05/3 \approx 0.017$ as our criterion for significance. Thus, to reject the null hypothesis that the two population means are equal, the p-value would have to be smaller than 0.017 (instead of 0.05!). The Bonferroni adjustment is a very strict way of maintaining the familywise error rate.

However, there is a trade-off between controlling the familywise error rate and increasing the type II error. By reducing the type I error rate, the type II error increases. Hence the statistical power decreases, potentially causing us to miss significant effects in the population.

The good news is that there are alternatives to the Bonferroni method. The bad news is that there are numerous types of post hoc tests—Stata provides no less than nine such methods! All these post hoc tests are based on different assumptions and designed for different purposes, whose details are clearly beyond the scope of this book.[14]

The most widely used post hoc test in market research is **Tukey's honestly significant difference test**, often simply referred to as *Tukey's method*. Tukey's method is a very versatile test controlling for type I error, but is limited in terms of statistical power (Everitt and Skrondal 2010). The test divides the difference between the largest and smallest pairs of means by the data's standard error that combines all possible pairwise differences and produces a value called *Tukey's statistic*. Where Tukey's statistic is larger than the critical value obtained from a normal distribution, the pairwise differences are rendered statistically significant. Tukey's method relies on two important requirements:

1. they require an equal number of observations for each group (differences of only a few observations are not problematic, though), and
2. they assume that the population variances are equal.

Alternative post hoc tests are available if these requirements are not met. When sample sizes clearly differ, we can draw on *Scheffé's method*, which is conservative by nature and thus has low statistical power. Alternatively, we can use *Dunnett's method*, which is useful when multiple pairwise comparisons (i.e., multiple treatment groups) are made with reference to a single control group. This is standard in experiments that distinguish between control and treatment groups, as is often encountered in marketing research.

Post hoc tests thus facilitate pairwise comparisons between groups while maintaining the familywise error rate. However, they do not allow for making statements regarding the strength of a factor variable's effects on the dependent variable. We can only do this after calculating the effect sizes, which we will do next.

6.6.5 Measure the Strength of the Effects

We can compute the η^2 (the **eta-squared**) coefficient to determine the strength of the effect (also referred to as the *effect size*) that the factor variable exerts on the dependent variable. The eta squared is the ratio of the between-group variation

[14] The Stata `help contrast` function provides an overview and references.

(SS_B) to the total variation (SS_T) and therefore indicates the variance accounted for by the sample data. Since η^2 is equal to the *coefficient of determination* (R^2), known from regression analysis (Chap. 7), Stata refers to it as R-squared in the output.

η^2 can take on values between 0 and 1. If all groups have the same mean value, and we can thus assume that the factor has no influence on the dependent variable, η^2 is 0. Conversely, a high value implies that the factor exerts a strong influence on the dependent variable. In our example, η^2 is:

$$\eta^2 = \frac{SS_B}{SS_T} = \frac{273.80}{584} \approx 0.469$$

The outcome indicates that 46.9% of the total variation in sales is explained by the promotion campaigns. The η^2 is often criticized as being too high for small sample sizes of 50 or less. We can compute ω^2 (pronounced **omega-squared**), which corresponds to the *Adjusted* $\mathbf{R^2}$ from regression analysis (Chap. 7), to compensate for small sample sizes:

$$\omega^2 = \frac{SS_B - (k-1) \cdot MS_W}{SS_T + MS_W} = \frac{273.80 - (3-1) \cdot 11.916}{584 + 11.916} \approx 0.421$$

This result indicates that 42.1% of the total variation in sales is accounted for by the promotion campaigns. Generally, you should use ω^2 for n ≤ 50 and η^2 for $n > 50$.

It is difficult to provide firm rules of thumb regarding which values are appropriate for η^2 or ω^2, as this varies from research area to research area. However, since the η^2 resembles Pearson's correlation coefficient (Chap. 7), we follow the suggestions provided in Chap. 7. Thus, we can consider values below 0.30 weak, values from 0.31 to 0.49 moderate, and values of 0.50 and higher strong.

6.6.6 Interpret the Results and Conclude

Just as in any other type of analysis, the final step is to interpret the results. Based on our results, we conclude that not all promotional activities have the same effect on sales. An analysis of the strength of the effects revealed that this association is moderate.

6.6.7 Plotting the Results (Optional)

In a final step, we can plot the estimated means of the dependent variable across the different samples. We could, for example, plot the mean of the sales across the stores with different types of promotion campaigns (i.e., point of sales display, free tasting stand or in-store display). When plotting the estimated group means, it is common to show the *confidence interval*, which is the interval within which the mean estimate falls with a certain probability (e.g., 95%). In this way, we can see whether the mean of the outcome variable across the different groups differ significantly or not without examining the numbers in the table. We will illustrate this optional step in the case study later in this chapter.

6.7 Going Beyond One-Way ANOVA: The Two-Way ANOVA

A logical extension of the one-way ANOVA is to add a second factor variable to the analysis. For example, we could assume that, in addition to the different promotion campaigns, management also varied the type of service provided by offering either self-service or personal service (see column "Service type" in Table 6.1). The two-way ANOVA is similar to the one-way ANOVA, except that the inclusion of a second factor variable creates additional types of variation. Specifically, we need to account for two types of between-group variations:

1. the between-group variation in factor variable 1 (i.e., promotion campaigns), and
2. the between-group variation in factor variable 2 (i.e., service type).

In its simplest form, the two-way ANOVA assumes that these factor variables are unrelated. However, in market research applications, this is rarely so, thereby requiring us to consider *related factors*. When we take two related factor variables into account, we not only have to consider each factor variable's direct effect (also called the **main effect**) on the dependent variable, but also their **interaction effect**. Conceptually, an interaction effect is an additional effect due to combining two (or more) factors. Most importantly, this extra effect cannot be observed when considering each of the factor variables separately and thus reflects a concept known as synergy. There are many examples in everyday life where the whole is more than simply the sum of its parts as we know from cocktail drinks, music, and paintings. For an entertaining example of interaction, see Box 6.7.

In our example, the free tasting stand might be the best promotion campaign when studied separately, but it could well be that when combined with personal service, the point of sale display produces higher sales. A significant interaction effect indicates that the combination of the two factor variables results in higher (or lower) sales than when each factor variable is considered separately. The computation of these effects, as well as a discussion of other technical aspects, lies beyond the scope of this book, but are discussed in the ⭳ Web Appendix (→ Downloads).

Table 6.5 provides an overview of the steps involved when carrying out the following tests in Stata: One-sample *t*-test, paired samples *t*-test, independent samples *t*-test, and the one-way ANOVA. Owing to data limitations to accommodate all types of parametric tests in this chapter by means of Oddjobs Airways, we will use data from the case study in the theory section to illustrate the Stata commands. This data restriction applies only to this chapter.

Box 6.7 A Different Type of Interaction

http://tinyurl.com/interact-coke

Table 6.5 Steps involved in carrying out one, two, or more group comparisons with Stata

Theory	Action
One-sample t-test[a]	
Formulate the hypothesis:	
Formulate the study's hypothesis:	*For example:*
	H_0: $\mu = \#$[b]
	H_1: $\mu \neq \#$
Choose the significance level:	
	Usually, α is set to 0.05, but: if you want to be conservative, α is set to 0.01, and: in exploratory studies, α is set to 0.10. We choose a significance level of 0.05.
Select an appropriate test:	
What is the testing situation?	Determine the fixed value again which that you are comparing.
Is the test variable measured on an interval or ratio scale?	Check Chap. 3 to determine the measurement level of the variables.
Are the observations independent?	Consult Chap. 3 to determine whether the observations are independent.
Is the test variable normally distributed or is n>30 and are the group variances the same?	*Check for normality*
	Carry out the Shapiro-Wilk normality test. Go to ► Statistics ► Summaries, tables, and tests ►Distributional plots and tests ►Shapiro-Wilk normality test. Select the test variable *outcome1* under **Variables:** and click on **OK**. A p-value below 0.05 indicates non-normality.
	`swilk outcome1`

(continued)

Table 6.5 (continued)

Theory	Action
Specify the type of t-test	Select the one-sample *t*-test.
Is the test one or two-sided?	Determine the region of rejection.
Calculate the test statistic:	
Specify the test variable and the fixed value	Go to ► Statistics ► Summaries, tables, and tests ► Classical tests of hypotheses ► t test (mean-comparison test). Select the first option **One-sample** from the dialog box and specify the test variable outcome1 under the **Variable name**:. Next, enter the hypothetical mean under **Hypothesized mean**:, choose the confidence interval **95**, which equates to a statistical significance at < 0.05 and click **OK**.
	`ttest outcome1==#`
Interpret the results:	
Look at the test results	For two-sided tests: compare the *p*-value under **Ha: mean(diff) != 0** with 0.05 and decide whether the null hypothesis is supported. The *p*-value under **Pr(|T| > |t|)** should be lower than 0.05 to reject the null hypothesis.
	For one-sided tests, look at either **Ha: mean (diff) < 0** (left-sided) or **Ha: mean(diff) > 0** (right-sided).
What is your conclusion?	Reject the null hypothesis that the population mean of the outcome variable *outcome1* is equal to the hypothetical known parameter against which you compare (i.e., #) if the *p*-value is lower than 0.05.
Paired samples t-test	
Formulate the hypothesis:	
Formulate the study's hypothesis:	*For example:*
	$H_0: \mu_1 = \mu_2$
	$H_1: \mu_1 \neq \mu_2$
Choose the significance level:	
	Usually, α is set to 0.05, but: if you want to be conservative, α is set to 0.01, and: in exploratory studies, α is set to 0.10. We choose a significance level of 0.05.
Select an appropriate test:	
What is the testing situation?	Determine the number of groups you are comparing.
Are the test variables measured on an interval or ratio scale?	Check Chap. 3 to determine the measurement level of the variables.
Are the observations dependent?	Next, consult Chap. 3 to determine whether the observations are independent.
	Check for normality

(continued)

Table 6.5 (continued)

Theory	Action
Are the test variables normally distributed or is $n > 30$ in each of the groups and are the group variances the same?	Run the Shapiro-Wilk normality test. Go to ► Statistics ► Summaries, tables, and tests ►Distributional plots and tests ►Shapiro-Wilk normality test. Select the test variable *outcome1* under **Variables**. Next, under the tab **by/if/in** tick the box **Repeat command by groups**, specify the grouping variable *groupvar* under **Variables that define groups:** and click on **OK**. A *p*-value below 0.05 indicates non-normality.
	`by groupvar,`[c] `sort: swilk outcome1`
	Check for equality of variances assumptions
	To perform Levene's test, go to ► Statistics ► Classical tests of hypotheses ► Robust equal-variance test. Specify the dependent variable under **Variable** *outcome1*, and the grouping variable *groupvar* under **Variable defining comparison groups** and click **OK**. To validate the equality of variances, the assumption *p*-values should lie above 0.05 for **W0**, **W50**, and **W10**.
	`robvar outcome1, by(groupvar)`
Specify the type of t-test	The data appear to be normally distributed with equal group variances. We can now proceed with the paired sample *t*-test.
Is the test one or two-sided?	Determine the region of rejection.
Calculate the test statistic:	
Select the paired test variables	Go to ► Statistics ► Summaries, tables, and tests ► Classical tests of hypotheses ► t test (mean-comparison test). Select the fourth option **Paired** from the dialog box and specify the test variable "*variable1*" under the First Variable. Next, enter the second comparison group "*variable2*" under **Second variable**, choose the confidence interval **95**, which equates to a statistical significance of $\alpha = 0.05$ and click **OK.**
	`ttest variable1==variables2`[d]
Interpret the results:	
Look at the test results	For two-sided tests: compare the *p*-value under **Ha: mean(diff) != 0** with 0.05 and decide whether the null hypothesis is supported. The *p*-value under **Pr(|T| > |t|)** should be lower than 0.05 to reject the null hypothesis.
	For one-sided tests, look at either **Ha: mean (diff) < 0** (left-sided) or **Ha: mean(diff) > 0** (right-sided).
What is your conclusion?	Reject the null hypothesis if the *p*-value is lower than 0.05.

(continued)

Table 6.5 (continued)

Theory	Action
Independent sample t-test	
Formulate the hypothesis:	
Formulate the study's hypothesis:	*For example:*
	H_0: $\mu_1 = \mu_2$
	H_1: $\mu_1 \neq \mu_2$
Choose the significance level:	
	Usually, α is set to 0.05, but: if you want to be conservative, α is set to 0.01, and: in exploratory studies, α is set to 0.10. We choose a significance level of 0.05.
Select an appropriate test:	
What is the testing situation?	Determine the number of groups you are comparing.
Are the test variables measured on an interval or ratio scale?	Check Chap. 3 to determine the measurement level of the variables.
Are the observations dependent?	Next, consult Chap. 3 to determine whether the observations are independent.
Are the test variables normally distributed or is $n > 30$ in each of the groups and are the group variances the same?	*Check for normality*
	Run the Shapiro-Wilk normality test. Go to ► Statistics ► Summaries, tables, and tests ► Distributional plots and tests ► Shapiro-Wilk normality test. Select the test variable *overall_sat* under **Variables**. Next, under the tab **by/if/in** tick the box **Repeat command by groups**, specify the grouping variable *gender* under **Variables that define groups** and click on **OK**. A *p*-value below 0.05 indicates non-normality.
	`by gender, sort: swilk overall_sat`
	Check for equality of variances assumptions
	Next, to perform Levene's test, go to ► Statistics ► Classical tests of hypotheses ►Robust equal-variance test. Specify the dependent variable *overall_sat* under **Variable:** and the variable *gender* under **Variable defining comparison groups:** aand click **OK**. The *p*-values of **W0**, **W50** and **W10** should be above 0.05 to validate the equality of the variances assumption.
	`robvar overall_sat, by(gender)`
Specify the type of *t*-test	The data appear to be normally distributed with equal group variances. We can now proceed with the two-sample *t*-test.
Is the test one or two-sided?	Determine the region of rejection.

(continued)

Table 6.5 (continued)

Theory	Action
Calculate the test statistic:	
Select the test variable and the grouping variable	Go to ▶ Statistics ▶ Summaries, tables, and tests ▶ Classical tests of hypotheses ▶ t test (mean-comparison test). Select the second option **Two-sample using groups** from the dialog box and specify the test variable *overall_sat* under the **Variable name**. Next, enter the variable *gender* under **Group variable name**. Select the confidence interval **95**, which equates to a statistical significance of $\alpha = 0.05$ and then click **OK**.
	`ttest overall_sat, by(gender)`
Interpret the results:	
Look at the test results	For two-sided tests: Compare the *p*-value under **Ha: mean(diff)!** = **0** with 0.05 and decide whether the null hypothesis is supported. The *p*-value under **Pr(|T| > |t|)** should be lower than 0.05 to reject the null hypothesis.
	For one-sided tests, look either at **Ha: mean (diff)** < **0** (left-sided) or **Ha: mean(diff)** > **0** (right-sided).
What is your conclusion?	Reject the null hypothesis that the population mean of the overall satisfaction score among female travelers is equal to the population overall mean satisfaction score of male traverlers if the *p*-value is lower than *0.05*.
One-way ANOVA	
Formulate the hypothesis:	
Formulate the study's hypothesis:	*For example:*
	H_0: $\mu_1 = \mu_2 = \mu_3$
	H_1: At least two of the population means are different.
Choose the significance level:	
	Usually, α is set to 0.05, but: if you want to be conservative, α is set to 0.01, and: in exploratory studies, α is set to 0.10. We choose a significance level of 0.05.
Select an appropriate test:	
What is the testing situation?	Determine the number of groups you are comparing.
Are there at least 20 observations per group?	Check Chap. 5 to determine the sample size in each group.
Is the dependent variable measured on an interval or ratio scale?	Determine the type of test that you need to use for your analyses by checking the underlying assumptions first. Check Chap. 3 to determine the measurement level of the variables.

(continued)

Table 6.5 (continued)

Theory	Action
Are the observations independent?	Next, consult Chap. 3 to determine whether the observations are independent.
Is the test variable normally distributed or is *n* larger than 30 per group and are the group variances the same?	*Check for normality*
	Carry out the Shapiro-Wilk normality test. Go to ► Statistics ► Summaries, tables, and tests ►Distributional plots and tests ►Shapiro-Wilk normality test. Select the test variable *overall_sat* under **Variables**. Next, under the tab **by/if/in** tick the box **Repeat command by groups**, specify the grouping variable *status* under **Variables that define groups** and click on **OK**. Values (**V**) larger than **1** with *p*-values below 0.05 indicate non-normality.
	`by status, sort: swilk overall_sat`
	Check for Equality of Variances Assumption
	Perform Levene's test. Go to ► Statistics ► Classical tests of hypotheses ►Robust equal-variance test. Specify the dependent variable *overall_sat* under **Variable:** aand the grouping variable *status* under **Variable defining comparison groups:** aand then click on **OK**. The *p*-values of **W0**, **W50** and **W10** should be above 0.05 to validate the equality of the variances assumption.
	`robvar overall_sat, by (status)`
Select the type of the test	Now that the assumption of normality and equality of the variance are met, proceed with the one-way ANOVA analysis.
Calculate the test statistic:	
Specify the dependent variable and the factor (grouping variable)	Go to ►Statistics ►Linear models and related ANOVA/MANOVA ►Analysis of variance and covariance. Specify the dependent variable *overall_sat* under the **Dependent variable:** aand the variable *status* under the **Model:**. Next, select the **Partial Sums of squares** and then click on **OK**.
	`anova overall_sat status`
Interpret the results:	
Look at the test results	Compare the *p*-value under **Model** with the significance level. The *p*-value should be lower than 0.05 to reject the null hypothesis.
Carry out pairwise comparisons	You can only carry out post hoc tests *after* you have carried out the ANOVA analysis. After the ANOVA analysis, go to ► Statistics ► Postestimation. In the next window that follows, go to ►Tests, contrasts, and comparisons of parameter estimates ► Pairwise comparisons and click on **Launch**.

(continued)

Table 6.5 (continued)

Theory	Action
	Select the variable *status* under **Factor terms to compute pairwise comparisons for:** aand select **Tukey's method** option in the **Multiple comparisons** box. Finally, go to the **Reporting** tab, tick the box **Show effects table with confidence intervals and p-values** and the box **Sort the margins/differences in each term** and click on **OK**. Now check whether the pairwise mean comparisons differ significantly if the *p*-values (under **P>|t|)** are lower than 0.05.
	`pwcompare status, effects sort mcompare (tukey)`
	If unequal variances are assumed, use Scheffé's method.
Look at the strength of the effects	Check for the strengths of the effects under **R-squared** and **Adjusted R-squared** in the output.
What is your conclusion?	Based on pairwise comparisons: Check which pairs differ significantly from each other. If the *p*-values tied to the pairwise mean comparisons are $<$ 0.05, reject the null hypothesis that the mean comparisons between the two groups are equal.
	Based on the output from the one-way ANOVA table, reject the null hypothesis that at least two population means are equal if the *p*-value is lower than 0.05.
Plotting the ANOVA results (optional)	To plot the results from the one-way ANOVA, go to ► Statistics ► Postestimation. In the next window that follows, go to ► Marginal analysis ► Marginal means and marginal effects, fundamental analyses and click on **Launch**. Enter the variable *status* under **Covariate**, tick the box **Draw profile plots of results** and then click on **OK**.

[a] Note that the Oddjob Airways dataset is not well suited to perform (1) the one-sample *t*-test and (2) the paired samples *t*-test. We therefore use hypothetical variables to illustrate the Stata commands

[b] # = refers to a hypothetical constant (number) against which you want to compare

[c] **Outcome1** refers to the dependent variable, with **groupvar** representing the two groups with and without treatment

[d] **Variable1** and **Variable2** represent the outcome variable with and without treatment

6.8 Example

Let's now turn to the Oddjob Airways case study and apply the materials discussed in this chapter. Our aim is to identify the factors that influence customers' overall price/performance satisfaction with the airline and explore the relevant target groups for future advertising campaigns. Based on discussions with the Oddjob Airways management, answering the following three research questions will help achieve this aim:

1. Does the overall price/performance satisfaction differ by gender?
2. Does the overall price/performance satisfaction differ according to the traveler's status?
3. Does the impact of the traveler's status on the overall price/performance satisfaction depend on the different levels of the variable gender?

The following variables (variable names in parentheses) from the Oddjob Airways dataset (⭳ Web Appendix → Downloads) are central to this example:

- overall price/performance satisfaction (*overall_sat*),
- respondent's gender (*gender)*, and
- traveler's status (*status*).

6.8.1 Independent Samples *t*-Test

6.8.1.1 Formulate Hypothesis

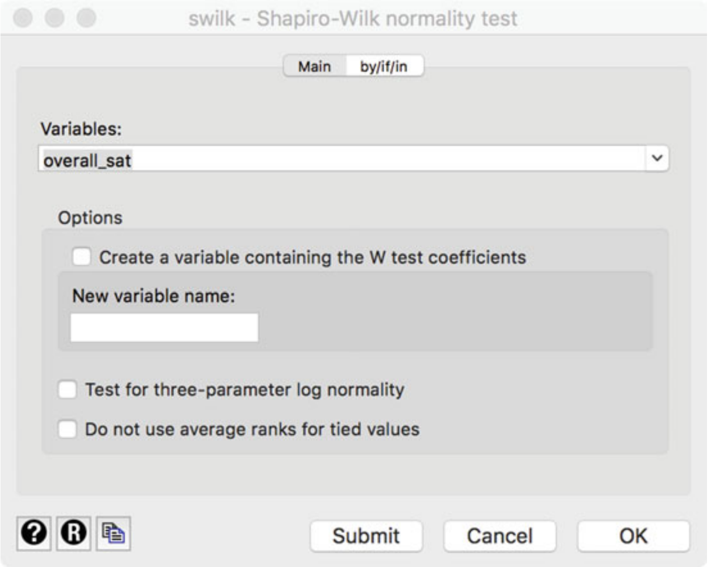
We start by formulating a non-directional hypothesis. The null hypothesis of the first research question is that the overall price/performance satisfaction means of male and female travelers are the same (H_0), while the alternative hypothesis (H_1) expects the opposite.

6.8.1.2 Choose the Significant Level

Next, we decide to use a significance level (α) of 0.05, which means that we allow a maximum chance of 5% of mistakenly rejecting a true null hypothesis.

6.8.1.3 Select an Appropriate Test

We move to the next step to determine the type of test, which involves assessing the testing situation, the nature of the measurements, checking the assumptions, and selecting the region of rejection. We start by defining the testing situation of our analysis, which concerns comparing the mean overall price/performance satisfaction scores (measured on a ratio scale) of male and female travelers. In our example, we know that the sample is a random subset of the population and we also know that other respondents' responses do not influence those of the respondents (i.e., they are independent). Next, we need to check if the dependent variable *overall_sat* is normally distributed between male and female travelers (i.e., normality assumption) and whether male travelers show the same variance in their overall price satisfaction as female travelers (i.e., equality of variance assumption). We use the Shapiro-Wilk



Variables: overall_sat

Options

□ Create a variable containing the W test coefficients

New variable name:

□ Test for three-parameter log normality

□ Do not use average ranks for tied values

Fig. 6.6 Shapiro-Wilk normality test dialog box

Table 6.6 Shapiro-Wilk normality test output

```
by gender, sort: swilk overall_sat

-------------------------------------------------------------------------------------
---
-> gender = female

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
 overall_sat |        280    0.98357      3.293     2.788    0.00265

-------------------------------------------------------------------------------------
---
-> gender = male

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
 overall_sat |        785    0.99050      4.805     3.848    0.00006
```

test for the normality test. Go to ► Statistics ► Summaries, tables, and tests ► Distributional plots and tests ► Shapiro-Wilk normality test. In the **Main** dialog box that follows (Fig. 6.6), select the variable *overall_sat* under **Variables**. Next, in the **by/if/in** tab, tick the box **Repeat command by groups** and enter the variable *gender* under **Variables that define groups** and then click on **OK**.

Table 6.6 displays the Stata output that follows. Stata reports the Shapiro-Wilk test statistic in its original version (**W**) and scaled version (**V**) with their corresponding *z*-values (*z*) and *p*-values under (**Prob** > **z**). Table 6.6 shows that

Table 6.7 Levene's test output

```
robvar overall_sat, by(gender)

            | Summary of Overall, I am satisfied
            | with the price performance ratio of
            |           Oddjob Airways.
     Gender |        Mean   Std. Dev.       Freq.
------------+------------------------------------
     female |         4.5   1.6461098         280
       male |   4.2369427   1.6130561         785
------------+------------------------------------
      Total |   4.3061033   1.6251693       1,065

W0  =  0.41763753   df(1, 1063)     Pr > F = 0.51825772
W50 =  0.23527779   df(1, 1063)     Pr > F = 0.62773769
W10 =  0.02419285   df(1, 1063)     Pr > F = 0.87642474
```

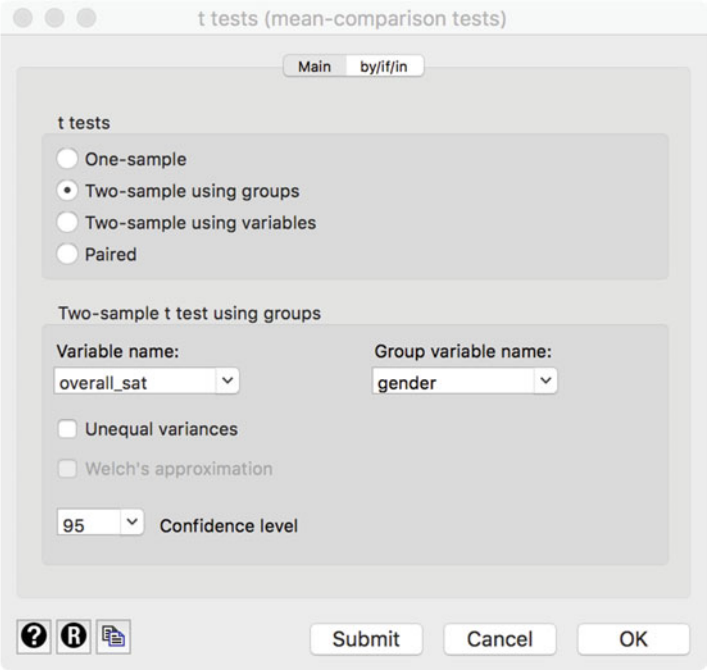
the p-values under (**Prob** $>$ **z**) of both female (**0.00265**) and male (**0.00006**) samples are smaller than 0.05, indicating that the normality assumption *is* violated.

Next, we need to check for the equality of the variances assumption. Go to ▶ Statistics ▶ Summaries, tables, and tests ▶ Classical tests of hypotheses ▶ Robust equal-variance test. Enter the dependent variable *overall_sat* into the **Variable** box and *gender* in the **Variable defining the comparison groups** box and click on **OK**.

As you can see in Table 6.7, Stata calculates Levene's test for the mean (**W0**), the median (**W50**), and for the 10% trimmed mean replacement (**W10**), with their corresponding p-values (**Pr** $>$ **F**) at the bottom of the table. We can see that the p-values of **W0**, **W50**, and **W10** are higher than 0.05 and, thus, not significant. This means that there is no reason to think that the variances for male and female travelers are different. Overall, we conclude that the data are not normally distributed, but that the variances across the male and the female groups are equal, allowing us to utilize a parametric test for group differences, because these are robust against violations of normality when sample sizes are larger than 30. Having checked the underlying assumptions, we can now decide on the region of rejection of our study's main research questions. This relates does, of course, relate to our study's main hypothesis, which was formulated as non-directional, implying a two-sided test.

6.8.1.4 Calculate the Test Statistic and Make the Test Decision

In the next step, and given that the equal variances assumption was tenable, we decide to use an independent samples t-test. To run this test, go to ▶ Statistics ▶ Summaries, tables, and tests ▶ Classical tests of hypotheses ▶ t test (mean-comparison test). In the **Main** dialog box that follows (Fig. 6.7), select the second option **Two-sample using groups** from the dialog box and specify the outcome variable (*overall_sat*) under **Variable name**. Next, enter the grouping variable *gender* under **Group variable name**. Select the confidence interval **95**, which equates to a significance level of 5% and then click **OK**.



t tests

○ One-sample
● Two-sample using groups
○ Two-sample using variables
○ Paired

Two-sample t test using groups

Variable name: overall_sat

Group variable name: gender

□ Unequal variances

□ Welch's approximation

95 Confidence level

Fig. 6.7 Dialog box, independent sample *t*-test

Table 6.8 Output of the independent sample *t*-test in Stata

```
ttest overall_sat, by(gender)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  female |     280         4.5    .0983739     1.64611    4.306351    4.693649
    male |     785    4.236943    .0575724    1.613056    4.123928    4.349957
---------+--------------------------------------------------------------------
combined |   1,065    4.306103    .0497994    1.625169    4.208387    4.403819
---------+--------------------------------------------------------------------
    diff |            .2630573    .1128905                .0415438    .4845708
------------------------------------------------------------------------------
    diff = mean(female) - mean(male)                              t =   2.3302
Ho: diff = 0                                     degrees of freedom =     1063

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9900         Pr(|T| > |t|) = 0.0200          Pr(T > t) = 0.0100
```

The output that follows (Table 6.8) provides diverse information on the male and female travelers, including their mean overall price/performance satisfaction, the standard error, standard deviation, and the 95% confidence intervals (see Chap. 5). At the bottom of the table, Stata shows the results of the *t*-test for both one-sided and two-sided tests, and leaves it to the researcher to decide which results to interpret. In our case, our main hypothesis was formulated non-directionally (i.e., two-sided) and we therefore focus on the test results under **Ha: diff!** = **0**, which are based on the two-tailed significance level. You can ignore the other test results, since these tests are for directional hypotheses (**Ha: diff** < **0** for a one-(left) sided hypothesis and **Ha: diff** > **0** for a one-(right) sided hypothesis).

6.8.1.5 Interpret the Results

When comparing the *p*-value under **Pr(|T|** > **|t|)** with the significance level, we learn that the *p*-value (**0.020**) is smaller than the significance level (0.05). Hence, we conclude that that the overall price satisfaction differs significantly between female and male travelers.

> If the normality and equality of the variance assumptions were violated, and the sample sizes were small (i.e., < 30 observations), the Mann-Whitney U test, which Stata refers to as the Wilcoxon signed rank-sum test, should have been performed. This is the non-parametric counterpart of the independent sample *t*-test. To obtain the Wilcoxon signed rank-sum test, go to ►Statistics ► Summaries, tables, and tests ► Non-parametric tests of hypotheses ► Wilcoxon rank-sum test. In the **Main** dialog box that follows, enter the dependent variable *overall_sat* under **Variable** and the variable *gender* under **Grouping variable**. Stata lists the number of observations for female and male travelers separately, followed by their corresponding observed and expected rank sums. The test statistic can be found at the bottom of the table under the **H0**. The *p*-value under **Prob** > **|z|** is **0.0125** and, thus, smaller than 0.05. This result indicates that the medians of male and female travelers differ significantly.

6.8.2 One-way ANOVA

In the second research question, we examine whether customers' membership status influences their overall price/performance satisfaction (i.e., *overall_sat)* with Oddjob Airways. The membership can have three forms: *Blue*, *Silver*, and *Gold*. Again, we start by formulating a null hypothesis that is again non-directional in nature, expecting that the mean of the overall price/performance satisfaction is the same between the status groups, while the alternative hypothesis states that at least two status groups differ. Next, we decide to use a significance level (α) of 0.05. We have already established that a comparison of three or more groups involves a one-way ANOVA and we therefore follow the steps as indicated in Fig. 6.4.

6.8.2.1 Check the Assumptions

In checking the assumptions, we already know that the sample is a random subset of the population and we also know that other respondents' responses do not influence those of the respondents (i.e., they are independent). Next, we check the normality and equality of the variance assumptions by focusing directly on Stata's output tables (see previous research question for the menu options). We start with the results of the Shapiro-Wilk test displayed in Table 6.9.

Table 6.9 Stata ouput of the Shapiro-Wilk test

```
by status, sort: swilk overall_sat
       -------------------------------------------------------------------------
       -> status = Blue
                         Shapiro-Wilk W test for normal data

           Variable |        Obs       W           V         z       Prob>z
       -------------+--------------------------------------------------
        overall_sat |        677    0.98403      7.067     4.764    0.00000

       -------------------------------------------------------------------------
       -> status = Silver
                          Shapiro-Wilk W test for normal data

           Variable |        Obs       W           V         z       Prob>z
       -------------+--------------------------------------------------
        overall_sat |        245    0.99353      1.153     0.331    0.37021

       -------------------------------------------------------------------------
       -> status = Gold
                          Shapiro-Wilk W test for normal data

           Variable |        Obs       W           V         z       Prob>z
       -------------+--------------------------------------------------
        overall_sat |        143    0.98998      1.119     0.255    0.39927
```

We can see that the Shapiro-Wilk test produces significant effects for *Blue* members (**Prob** > **z** = **0.00000**), but not for *Silver* members (**Prob** > **z** = **0.37021**) and *Gold* members (**Prob** > **z** = **0.39927**). This means that in the three different samples, the overall price satisfaction is normally distributed in the *Silver* and *Gold* groups, but *not* in the *Blue* group. As we mentioned previously, the ANOVA is robust to violations from normality when samples are greater than 30, meaning that we can move to the next step to test the equality of variance assumptions, even though one of our samples violates the normality assumption. The output of Levene's test is shown in Table 6.9.

As we can see in Table 6.10, the *p*-values of **W0**, **W50**, and **W10** under **PR** > **F** are higher than 0.05 (thus not significant), which means that the variances between travelers with different statuses are the same. Overall, we conclude that the equal variance assumption is tenable, we can therefore move ahead and test our study's second research question.

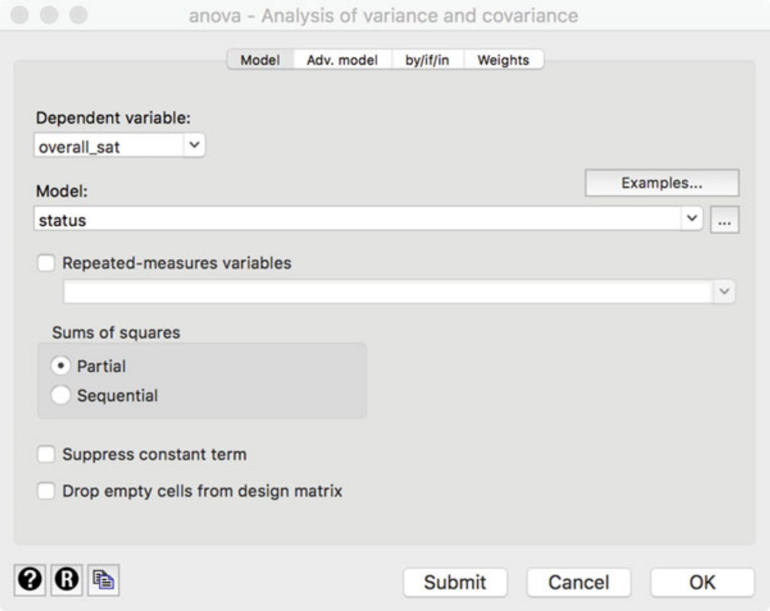
6.8.2.2 Calculate the Test Statistic

To run an ANOVA, go to ▶ Statistics ▶ Linear models and related ▶ ANOVA/MANOVA ▶ Analysis of variance and covariance. In the dialog box that follows (Fig. 6.8), select *overall_sat* from the drop-down menu under **Dependent variable** and *status* (**status**) from the drop-down menu under **Model**. Next, then click on **OK**.

Stata will produce the output as shown in Table 6.11. The top part of the table reports several measures of model fit, which we discuss in Box 6.6. Under **Partial SS**, Stata lists different types of variation. **Model** represents the between-group variation (SS_B), whereas **Residual** indicates the within-group variation (SS_W). Next

Table 6.10 Stata output of Levene's test

```
robvar overall_sat, by(status)

            | Summary of Overall, I am satisfied
            | with the price performance ratio of
   Traveler |           Oddjob Airways.
     status |        Mean   Std. Dev.       Freq.
------------+------------------------------------
       Blue |   4.4726736   1.6411609         677
     Silver |   4.0326531   1.5599217         245
       Gold |    3.986014   1.5563863         143
------------+------------------------------------
      Total |   4.3061033   1.6251693       1,065

W0  =  0.90696260   df(2, 1062)     Pr > F = 0.4040612
W50 =  0.06775398   df(2, 1062)     Pr > F = 0.93449438
W10 =  0.88470222   df(2, 1062)     Pr > F = 0.41314113
```



Dependent variable: overall_sat

Model: status

☐ Repeated-measures variables

Sums of squares

● Partial

○ Sequential

☐ Suppress constant term

☐ Drop empty cells from design matrix

Fig. 6.8 ANOVA dialog box

to **status**, Stata lists this variable's partial contribution to the total variance. Given that the model has only one variable at this stage, the partial sums of squares explained by *status* (**51.755064**) is exactly the same as the partial sums of squares of the model (**51.755064**).

Table 6.11 One-way ANOVA

```
anova overall_sat status

                         Number of obs =      1,065    R-squared     =  0.0184
                         Root MSE      =    1.61165    Adj R-squared =  0.0166

                  Source | Partial SS         df         MS        F    Prob>F
              -----------+----------------------------------------------------
                   Model |  51.755064          2   25.877532      9.96  0.0001
                  status |  51.755064          2   25.877532      9.96  0.0001
                Residual |  2758.4553      1,062   2.5974155
              -----------+----------------------------------------------------
                   Total |  2810.2103      1,064   2.6411751
```

6.8.2.3 Make the Test Decision

Let's now focus on the F-test result with respect to the overall model. The model has an F-value of **9.96**, which yields a p-value of **0.0001** (see **Prob** $>$ **F**), suggesting a statistically significant model.

6.8.2.4 Carry Out Post Hoc Tests

Next, we carry out pairwise group comparisons using Tukey's method. In Stata, this is a post estimation command, which means that comparisons can only be carried out after estimating the ANOVA. To run Tukey's method, go to ► Statistics ► Postestimation. In the window that follows, go to ► Tests, contrasts, and comparisons of parameter estimates ► Pairwise comparisons and click on **Launch**. In the dialog box that opens, select the variable *status* under **Factor terms to compute pairwise comparisons for** and select the **Tukey's method** option from the **Multiple comparisons** drop-down menu. Next, go to the **Reporting** tab and first tick **Specify additional tables (default is effects table with confidence intervals)** and then tick **Show effects table with confidence intervals and p-values**. Finally, in the same window, tick the box **Sort the margins/differences in each term**, and then click on **OK**. This produces the following output as in Table 6.12.

To check whether the means differ significantly, we need to inspect the p-values under **Tukey P** $>$ **|t|.** The results in Table 6.12 indicate that the overall price/performance satisfaction differs significantly between *Gold* and *Blue* members, as well as between *Silver* and *Blue* members. The p-values of these pairwise comparisons are respectively **0.003** and **0.001** and, thus, smaller than 0.05. In contrast, the pairwise mean difference between *Gold* and *Silver* members does not differ significantly, as the p-value of **0.959** is above 0.05, indicating that members from these two status groups share similar views about the overall price/performance satisfaction with Oddjob Airways.

6.8.2.5 Measure the Strength of the Effects

The upper part of Table 6.11 reports the model fit. Besides the **R-squared** Stata reports the effect size η^2 (**0.0184**). This means that differences in the travelers' status explain **1.841%** of the total variation in the overall price satisfaction. The ω^2 displayed under **Adj R-squared** is **0.0166**.

6.8.2.6 Interpret the Results

Overall, based on the outputs of the ANOVA in Tables 6.11 and 6.12, we conclude that:

1. *Gold* and *Blue* members, as well as *Silver* and *Blue* members, differ significantly in their mean overall price satisfaction.
2. Membership status explains only a minimal share of the customers' price satisfaction. Hence, other factors—presently not included in the model—explain the remaining variation in the outcome variable.

6.8.2.7 Plot the Results

Next, we plot the results, but we first should save the estimated parameters. Note that these estimated parameters capture the instantaneous change in the overall satisfaction level in respect of every unit change in the membership status, also termed the *marginal effect* (Greene 1997; Bartus 2005). Stata allows a fast and efficient way of saving the estimated parameters from the ANOVA. To do so, go to ► Statistics ► Postestimation. In the dialog box that follows, go to ► Marginal analysis ► Marginal means and marginal effects, fundamental analyses and click on **Launch**. Enter the variable *status* under **Covariate**, tick the box **Draw profile plots of results** and then click on **OK**. Stata will simultaneously produce Table 6.13 and the plot shown in Fig. 6.9.

The plot depicts the predicted group means (listed in Table 6.13) surrounded by their confidence intervals, which the vertical lines through the dots in Fig. 6.9 indicate. These intervals are very useful, because they immediately reveal whether the predicted group means differ significantly. More precisely, if the vertical bars do *not* overlap vertically with each other, we can say there is a significant difference. Overall, Fig. 6.9 indicates that there is a significant mean difference in the overall price satisfaction between *Blue* and *Silver* members and between *Blue* and *Gold* members, but no significant difference between *Silver* and *Gold* members. This is exactly what Tukey's pairwise comparison indicated in Table 6.12.

Table 6.12 Output of Tukey's method

```
pwcompare status, effects sort mcompare(tukey)

Pairwise comparisons of marginal linear predictions
Margins      : asbalanced
---------------------------
             |    Number of
             |  Comparisons
-------------+-------------
      status |            3
---------------------------
-------------------------------------------------------------------------------
                |                              Tukey                Tukey
                |   Contrast   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+--------------------------------------------------------------
         status |
  Gold vs Blue  |  -.4866596   .1483253    -3.28   0.003    -.8347787   -.1385404
Silver vs Blue  |  -.4400205   .1201597    -3.66   0.001     -.722035    -.158006
Gold vs Silver  |  -.0466391   .1696038    -0.27   0.959    -.4446987    .3514205
---------------------------------------------------------------------------------
```

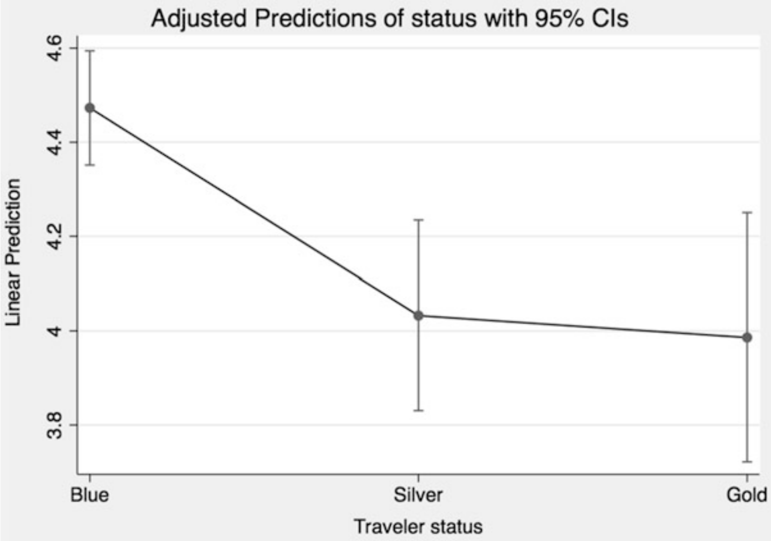


Fig. 6.9 Mean predicted level of overall satisfaction by flight frequency

Table 6.13 Average marginal effects by flight frequency

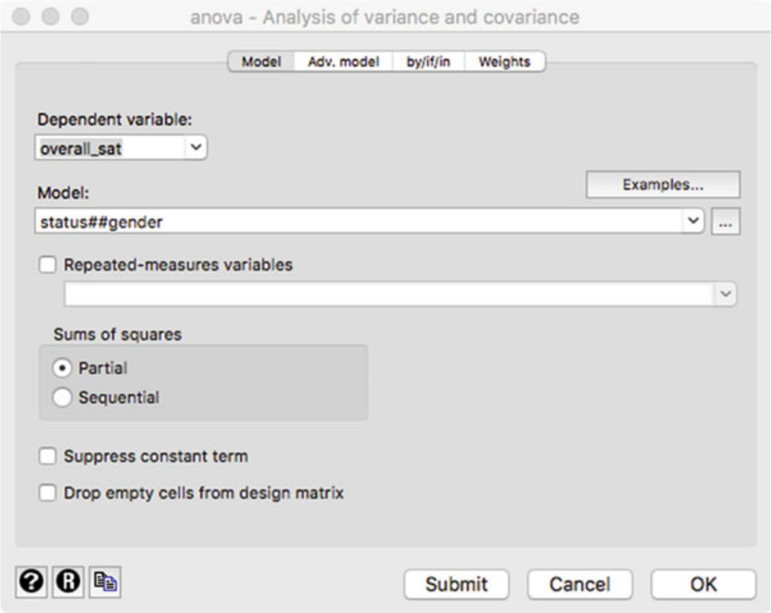
```
margins status

Adjusted predictions                              Number of obs     =      1,065

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      status |
        Blue |   4.472674   .0619407    72.21   0.000     4.351133    4.594214
      Silver |   4.032653   .1029645    39.17   0.000     3.830616     4.23469
        Gold |   3.986014   .1347729    29.58   0.000     3.721562    4.250465
------------------------------------------------------------------------------
```

6.8.3 Two-way ANOVA

The final research question asks whether the impact of *status* on *overall price satisfaction* depends on the different levels of the variable *gender* (i.e., male versus female). The two-way ANOVA allows answering this research question. The null hypothesis for this combined effect of status and gender (i.e., their interaction effect) is that the difference in the overall price satisfaction between *Blue*, *Silver*, and *Gold* members is the same regardless of the travelers' gender. We decide to use a significance level (α) of 0.05 and directly calculate the test statistic, given that we already checked the ANOVA assumptions in the second research question.

6.8.3.1 Calculate the Test Statistic

In Stata, interaction effects between two variables are indicated by a hashtag (#) between the variables that we interact. Now, let us test for interaction effects. Go to ► Statistics ► Linear models and related ► ANOVA/MANOVA ► Analysis of variance and covariance. The dialog box that opens is similar to that shown in Fig. 6.10, only this time we enter **status##gender** (instead of only **status**) in the **Model** box and click on **OK**. This produces the output shown in Table 6.14. Note that it is essential to have two hashtags (i.e., ##), as this tells Stata to include the two main variables **status** and **gender**, plus the interaction variable **status#gender**. The



Dependent variable: overall_sat

Model: status##gender

□ Repeated-measures variables

Sums of squares

● Partial

○ Sequential

□ Suppress constant term

□ Drop empty cells from design matrix

Fig. 6.10 ANOVA dialog box

Table 6.14 Output ANOVA

```
anova overall_sat status##gender

                        Number of obs =      1,065    R-squared     =  0.0255
                        Root MSE      =     1.6081    Adj R-squared =  0.0209

                 Source | Partial SS         df         MS        F    Prob>F
          --------------+----------------------------------------------------
                  Model |  71.656085          5   14.331217      5.54  0.0000
                        |
                 status |  53.277593          2   26.638796     10.30  0.0000
                 gender |   .2061234          1    .2061234      0.08  0.7777
          status#gender |  14.512817          2   7.2564083      2.81  0.0609
                        |
               Residual |  2738.5542      1,059   2.5859813
          --------------+----------------------------------------------------
                  Total |  2810.2103      1,064   2.6411751
```

reason for this is that *status* and *gender* are conditional main effects, whereby the effect of the one main variable is conditional on the effect of the other main variable with a value of 0. Thus, in this example, the conditional main effect of *status* represents the effect of *status* when *gender* is equal to 0, while the conditional effect of gender represents the effect of *gender* when *status* is equal to 0. When the included variables have no 0 category, such conditional effects have no interpretation. It is therefore important to set a meaningful 0 category.

6.8.3.2 Make the Test Decision

The output in Table 6.14 shows an *F*-value of **2.81** with a corresponding *p*-value of **0.0609** for the interaction term (**status#gender**). This *p*-value is higher than 0.05 and thereby not statistically significant. Note that had we decided to use a significance level (α) of 0.10, we would have found a significant interaction effect! However, as this is not the case in our example, we conclude that the overall level of price satisfaction by membership status does not depend on gender.

6.8.3.3 Carry Out Post Hoc Tests

To run Tukey's method, go to ► Statistics ► Postestimation. In the window that follows, go to ► Tests, contrasts, and comparisons of parameter estimates ► Pairwise comparisons and click on **Launch**. In the dialog box that opens, select the variable *status#gender* under **Factor terms to compute pairwise comparisons for** and select the **Tukey's method** option from the **Multiple comparisons** drop-down menu. Finally, go to the **Reporting** tab and first tick **Specify additional tables (default is effects table with confidence intervals)** and then tick **Show effects table with confidence intervals and p-values**. Finally, in the same window, tick the box **Sort the margins/differences in each term**, and then click on **OK**. This produces the output as shown in Table 6.15.

In this special case, post hoc tests are carried out across 15 distinct combinations within and between gender and membership status groups. We can check whether the pairwise mean comparisons differ significantly if the *p*-values (under **Tukey P** > **|t|**) are lower than 0.05. The results in Table 6.15 indicate that the overall price/performance satisfaction is significantly lower between: (1) female travelers with a *Silver* and *Blue* membership status, (2) female travelers with a *Gold* and *Blue* membership status, and (3) male and female travelers with a *Silver* membership status. The *p*-values of these pairwise comparisons are respectively **0.006**, **0.002**, and **0.002**, thus smaller than 0.05. All the other effects have a *p*-value higher than 0.05 and are not significant.

6.8.3.4 Measure the Strength of the Effects

In the next step, we focus on the model's effect strength. In Table 6.14 this is shown under **R-squared** and is **0.0255**, indicating that our model, which includes the interaction term, explains 2.5% of the total variance in the overall price satisfaction.

Table 6.15 Output of Tukey's method

```
pwcompare status#gender, mcompare(tukey) effects sort

Pairwise comparisons of marginal linear predictions

Margins      : asbalanced

----------------------------
              |    Number of
              |  Comparisons
--------------+-------------
status#gender |           15
----------------------------

----------------------------------------------------------------------------------------------------
                                  |                            Tukey                   Tukey
                                  |    Contrast  Std. Err.        t    P>|t|     [95% Conf. Interval]
----------------------------------+-----------------------------------------------------------------
                    status#gender |
(Silver#female) vs (Blue#female)  |   -.9876923   .2789273    -3.54   0.006    -1.784013   -.1913714
  (Gold#female) vs (Blue#female)  |      -.7425   .4160734    -1.78   0.476    -1.930365    .4453648
    (Gold#male) vs (Blue#female)  |    -.687874   .1784806    -3.85   0.002    -1.197425   -.1783227
  (Silver#female) vs (Blue#male)  |   -.6771613   .2683811    -2.52   0.118    -1.443373    .0890507
  (Silver#male) vs (Blue#female)  |   -.5829126   .1550695    -3.76   0.002    -1.025627   -.1401984
    (Gold#female) vs (Blue#male)  |    -.431969   .4090783    -1.06   0.899    -1.599863     .735925
      (Gold#male) vs (Blue#male)  |    -.377343   .1615031    -2.34   0.180    -.8384248    .0837387
    (Blue#male) vs (Blue#female)  |    -.310531   .1312038    -2.37   0.169    -.6851101    .0640482
    (Silver#male) vs (Blue#male)  |   -.2723816   .1351832    -2.01   0.334    -.6583217    .1135584
  (Gold#female) vs (Silver#male)  |   -.1595874   .4173454    -0.38   0.999    -1.351083    1.031909
    (Gold#male) vs (Silver#male)  |   -.1049614   .1814259    -0.58   0.992    -.6229216    .4129988
    (Gold#male) vs (Gold#female)  |     .054626    .426598     0.13   1.000    -1.163286    1.272538
(Gold#female) vs (Silver#female)  |    .2451923   .4774212     0.51   0.996    -1.117817    1.608201
  (Gold#male) vs (Silver#female)  |    .2998183   .2943965     1.02   0.912     -.540666    1.140303
(Silver#male) vs (Silver#female)  |    .4047797   .2808212     1.44   0.702    -.3969479    1.206507
----------------------------------------------------------------------------------------------------
```

6.8.3.5 Interpret the Results

The mere increase of 0.7% in the explained variance (from 1.8% in Table 6.11 to 2.5% in Table 6.14) indicates that the interaction term does not add much to the model's fit. This is not surprising, as the interaction term reiterates the main effects in a slightly different form.

6.8.3.6 Plot the Results

Finally, we move to plotting the results, which is optional. Like research question 2, we go to ▶ Statistics ▶ Postestimation. In the dialog box that follows, go to ▶ Marginal analysis ▶ Marginal means and interaction analysis ▶ At sample means and click on **Launch**. Next, enter the first variable *status* under **Covariate** and tick the box **Interaction analysis with another covariate** where you enter the second variable *gender*. Then tick the box **Draw profile plots of results** and click **OK**. Stata will produce the output in Table 6.16 and the plot displayed in Fig. 6.11.

Under **status#gender** in Table 6.16, we see the *average marginal effects* of overall price satisfaction in respect of all status groups as they vary by gender. Here we find that the mean overall price satisfaction of female *Blue* status members (indicated by **Blue#female**) is **4.68**, while the mean of male *Blue* status members equals **4.369469**, and so on. Figure 6.11 depicts exactly these margins and their corresponding confidence intervals. Overall, we conclude that the relationship between the overall price satisfaction and travelers' membership status does not vary by gender.

Table 6.16 Predicted average marginal effects of a combination between flight frequency and gender

```
margins status#gender, plot

  Adjusted predictions                          Number of obs     =      1,065

  Expression   : Linear prediction, predict()
  at           : 1.status        =    .6356808 (mean)
                 2.status        =    .2300469 (mean)
                 3.status        =    .1342723 (mean)
                 1.gender        =    .2629108 (mean)
                 2.gender        =    .7370892 (mean)
------------------------------------------------------------------------------
                  |            Delta-method
                  |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
    status#gender |
     Blue#female  |       4.68   .1072066    43.65   0.000     4.469639    4.890361
       Blue#male  |   4.369469   .0756386    57.77   0.000      4.22105    4.517888
   Silver#female  |   3.692308   .2575019    14.34   0.000     3.187036     4.19758
     Silver#male  |   4.097087   .1120415    36.57   0.000     3.877239    4.316936
     Gold#female  |     3.9375   .4020247     9.79   0.000     3.148645    4.726355
       Gold#male  |   3.992126   .1426957    27.98   0.000     3.712128    4.272124
   ---------------------------------------------------------------------------
```

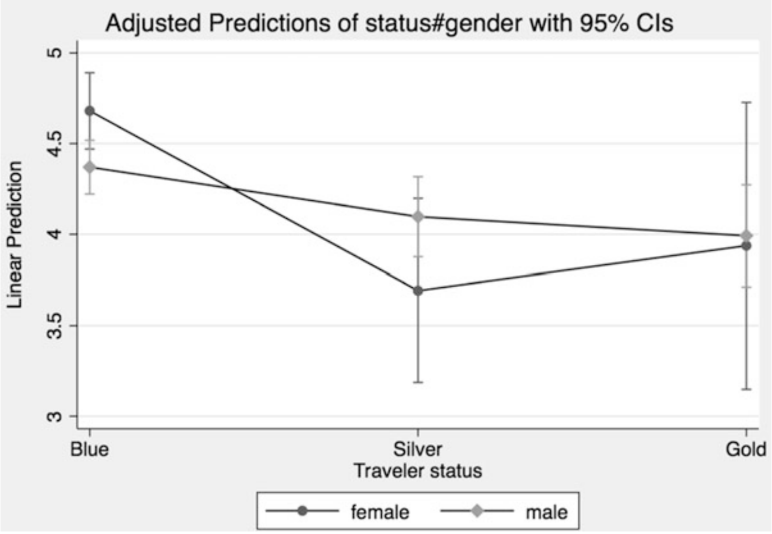


Fig. 6.11 Mean predicted level of overall satisfaction by membership status and gender

6.9 Customer Analysis at Crédit Samouel (Case Study)

CS
Crédit Samouel

In 2017, Crédit Samouel, a globally operating bank underwent a massive re-branding campaign. In the course of this campaign, the bank's product range was also restructured and its service and customer orientation improved. In addition, a comprehensive marketing campaign was launched, aimed at increasing the bank's customer base by one million new customers by 2025.

In an effort to control the campaign's success and to align the marketing actions, the management decided to conduct an analysis of newly acquired customers. Specifically, the management is interested in evaluating the segment customers aged 30 and below. To do so, the marketing department surveyed the following characteristics of 251 randomly drawn new customers (variable names in parentheses):

- Gender (*gender*: male/female).
- Bank deposit in Euro (*deposit*: ranging from 0 to 1,000,000).
- Does the customer currently attend school/university? (*training*: yes/no).
- Customer's age specified in three categories (*age_cat*: 16–20, 21–25, and 26–30).

Use the data provided in *bank.dta* (⤓ Web Appendix → Downloads) to answer the following research questions:

1. Which test do we have to apply to find out whether there is a significant difference in bank deposits between male and female customers? Do we meet the assumptions required to conduct this test? Also use an appropriate normality test and interpret the result. Does the result give rise to any cause for concern? Carry out an appropriate test to answer the initial research question.
2. Is there a significant difference in bank deposits between customers who are still studying and those who are not?
3. Which type of test or procedure would you use to evaluate whether bank deposits differ significantly between the three age categories? Carry out this procedure and interpret the results.
4. Reconsider the previous question and, using post hoc tests, evaluate whether there are significant differences between the three age groups.
5. Is there a significant interaction effect between the variables *training* and *age_cat* in terms of the customers' deposit?

6. Estimate and plot the average marginal effects of bank deposits over the different combinations of training groups and the customers' different age categories.
7. Based on your analysis results, please provide recommendations for the management team on how to align their future marketing actions.

6.10 Review Questions

1. Describe the steps involved in hypothesis testing in your own words.
2. Explain the concept of the p-value and explain how it relates to the significance level α.
3. What level of α would you choose for the following types of market research studies? Give reasons for your answers.
   (a) An initial study on preferences for mobile phone colors.
   (b) The production quality of Rolex watches.
   (c) A repeat study on differences in preference for either Coca Cola or Pepsi.
4. Write two hypotheses for each of the example studies in question 3, including the null hypothesis and alternative hypothesis.
5. Describe the difference between independent and paired samples t-tests in your own words and provide two examples of each type.
6. What is the difference between an independent samples t-test and an ANOVA?
7. What are post hoc test and why is their application useful in ANOVA?

6.11 Further Readings

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measure of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician, 57*(3), 171–178.

The authors discuss the distinction between p-value and α and argue that there is general confusion about these measures' nature among researchers and practitioners. A very interesting read!

Kanji, G. K. (2006). *100 statistical tests* (3^{rd} ed.). London: Sage.

If you are interested in learning more about different tests, we recommend this best-selling book in which the author introduces various tests with information on how to calculate and interpret their results using simple datasets.

Mooi, E., & Ghosh, M. (2010). Contract specificity and its performance implications. *Journal of Marketing, 74*(2), 105–120.

This is an interesting article that demonstrates how directional hypotheses are formulated based on theory-driven arguments about contract specificity and performance implications.

Stata.com (http://www.stata.com/manuals14/rmargins.pdf).

Stata offers a very thorough explanation of marginal effects and the corresponding Stata syntaxes for the estimation of marginal means, predictive margins, and marginal effects.

References

Agresti, A., & Finlay, B. (2014). *Statistical methods for the social sciences* (4th ed.). London: Pearson.

Bartus, T. (2005). Estimation of marginal effects using marge off. *The Stata Journal, 5*(3), 309–329.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin, 57*(1), 49–64.

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for equality of variances. *Journal of the American Statistical Association, 69*(346), 364–367.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Everitt, B. S., & Skrondal, A. (2010). *The Cambridge dictionary of statistics* (4th ed.). Cambridge: Cambridge University Press.

Field, A. (2013). *Discovering statistics using SPSS* (4th ed.). London: Sage.

Greene, W. H. (1997). *Econometric analysis* (3rd ed.). Upper Saddle River: Prentice Hall.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measure of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician, 57*(3), 171–178.

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin, 54*(4), 351–353.

Lehmann, E. L. (1993). The Fischer, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88*(424), 1242–1249.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278–292). Palo Alto: Stanford University Press.

Liao, T. F. (2002). *Statistical group comparison*. New York: Wiley-InterScience.

Lichters, M., Brunnlieb. C., Nave, G., Sarstedt, M., & Vogt, B. (2016). The influence of serotonin defficiency on choice deferral and the compromise effect. *Journal of Marketing Research*, 53(2), 183–198.

Lilliefors, H. W. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62*(318), 399–402.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics, 18*(1), 50–60.

Mitchell, M. N. (2015). *Stata for the behavioral sciences*. College Station: Stata Press.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature, 506*(7487), 150–152.

Ruxton, G. D., & Neuhaeuser, M. (2010). When should we use one-tailed hypothesis testing? *Methods in Ecology and Evolution, 1*(2), 114–117.

Schuyler, W. H. (2011). *Readings statistics and research* (6th ed.). London: Pearson.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3/4), 591–611.

Van Belle, G. (2008). *Statistical rules of thumb* (2nd ed.). Hoboken: Wiley.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38*(3/4), 330–336.