Hypothesis Testing and ANOVA

Electronic supplementary material
The online version of this chapter (https://doi.org/10.1007/978-3-662-56707-4_6) contains additional material that is available to authorized users. You can also download the "Springer Nature More Media App" from the iOS or Android App Store to stream the videos and scan the image containing the "Play button".

M. Sarstedt, E. Mooi, *A Concise Guide to Market Research*, Springer Texts in Business and Economics, https://doi.org/10.1007/978-3-662-56707-4_6

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 ANOVA.
- How to interpret SPSS output.

Keywords
α-Inflation • *α* error • Adjusted R^2 • Alternative hypothesis • Analysis of variance (ANOVA) • *β* error • Bonferroni correction • Confidence interval • Degrees of freedom • Directional hypothesis • Effect size • Eta squared • Explained variation • Factor level • Factor variable • *F*-test • F-test of sample variance • Familywise error rate • Grand mean • Independent samples • Independent samples *t*-test • Kruskal-Wallis rank test • Left-tailed hypothesis • Levene's test • Mann-Whitney U test • Marginal mean • Noise • Nonparametric test • Non-directional hypothesis • Null hypothesis • Omega squared • One-sample t-test • One-tailed test • One-way ANOVA • p-value • Paired samples • Paired samples t-test • Parametric test • Post hoc tests • Power analysis • Power of a statistical test • Practical significance • Quantile plot • Random noise • R^2 • Right-tailed hypothesis • Sampling error • Shapiro-Wilk test • Significance level • Standard error • Statistical significance • *t*-test • Test statistic • Tukey's honestly significant difference test • Two-samples *t*-test • Two-tailed test • Two-way ANOVA • Type I error • Type II error • Unexplained variation • Welch's correction • Wilcoxon matched-pairs signed-rank test • Wilcoxon signed-rank test • *z*-test.

6.1 Introduction

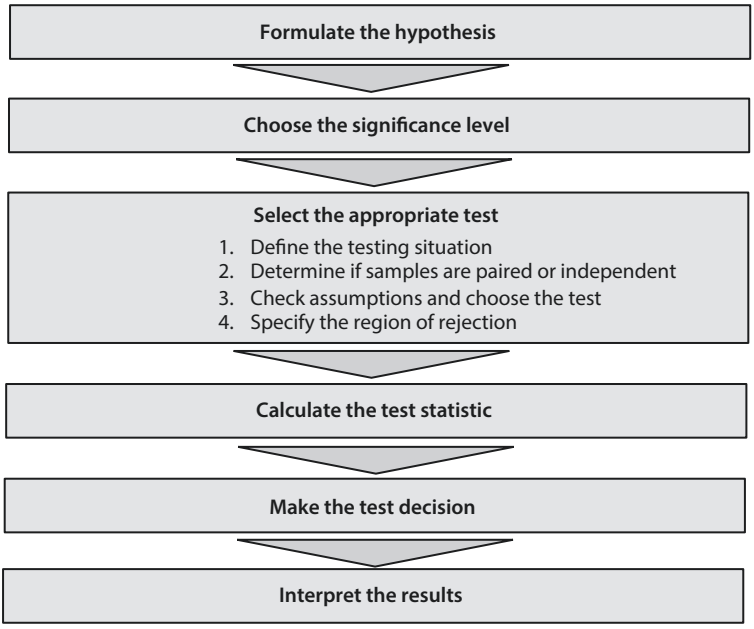
Do men or women spend more money on the Internet? Assume that the mean amount that a sample of men spends online is $200 per year against a women sample's mean of $250. When we compare mean values such as these, we always expect some difference. But, how can we determine if such differences are statistically significant? Establishing statistical significance requires ascertaining whether such differences are attributable to chance or not. If the difference is so large that it is unlikely to have occurred by chance, this indicates **statistical significance**. Whether results are statistically significant depends on several factors, including the difference, variation in the sample data, and the number of observations. 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 condition or parameter (such as a mean or difference) 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 four hours 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 that the stated hypothesis is likely true in the population of interest (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 **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 5 % in market research. 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.

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. Once we have calculated the test statistic, we can decide to either reject or support the hypothesis. This decision enables us to draw market research conclusions in the final step. ◘ Fig. 6.1 illustrates the six steps involved in hypothesis testing.



◘ **Fig. 6.1** 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 include: (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 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). The **marginal mean** represents the means of sales within stores in the last column. Finally, the very last cell shows the **grand mean**, which is the overall average across all service types and campaigns.

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 illustration. You will realize that the formulas are not as complicated as you might have thought! These formulas contain Greek characters and we have included a table describing each Greek character in the ↓ Web Appendix (→ Downloads).

◘ **Table 6.1** Sales data

Service type	Sales (units)			Marginal mean
	Point of sale display (stores 1–10)	Free tasting stand (stores 11–20)	In-store announcements (stores 21–30)	
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

https://www.guide-market-research.com/app/download/13488685427/SPSS+3rd_Greek+Characters.pdf?t=1516714139

6.3 Testing Hypotheses on One Mean

6.3.1 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 is 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 is 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: a first outcome may be that we do not reject the null hypothesis. This suggests there is no effect or no difference and that the null hypothesis can be retained. However, it would be incorrect to conclude from this 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. A second outcome may be that we 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 example 1, 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 of this expectation is:

$$H_0 : \mu \leq 45$$

$$H_1 : \mu > 45$$

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 are equal to the mean sales of the free tasting stand (example 2). This implies a **non-directional hypothesis**, which can be written as:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

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 Choose the Significance Level

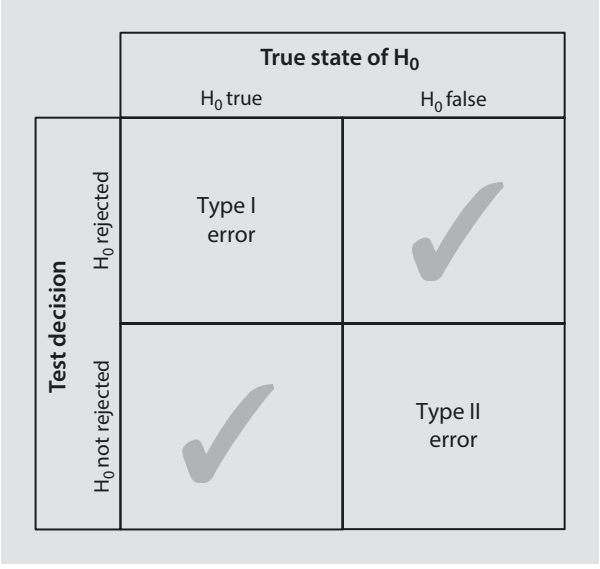
No type of hypothesis testing can evaluate the validity of a hypothesis with absolute certainty. In any study that involves drawing a sample from the population, there is always some probability that we will erroneously retain or reject the null hypothesis due to sampling error. In statistical testing, 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. This situation is also referred to as *false positive*. 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. This situation is also referred to as *false negative*.

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



◘ **Fig. 6.2** Type I and type II errors

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 implication. 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. It is important to note that statistical significance is required to establish practical significance. If we cannot reject that an effect is likely due to chance, such an effect is also not meaningful managerially. Whether results are practically significant depends on the management's perception of the difference or effect and whether this warrants action. For example, a statistical difference of 10 % in sales due to packaging differences, could be practically significant if the packaging change is possible, not too costly, accepted by the retailer etc. In sum, statistical significance does not imply practical significance but practical significance does require statistical 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). Power is 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 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.

Box 6.1 Statistical power of a test

How to calculate what sample size you need? Computing the required sample size (called a **power analysis**) can be complicated, depending on the test or procedure used. Fortunately, SPSS provides an add-on module called "Sample Power," which can be used to carry out such analyses. In addition, the Internet offers a wide selection of downloadable applications and interactive Web programs to conduct power analyses. One particular sophisticated and easy-to-use program is G*Power 3.0 which is available at no charge 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.

6.3.3 Select the 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.

6.3.3.1 Define the Testing Situation

When we test hypotheses, we may find ourselves in one of three situations. First, we may test if we want to compare a group to a hypothetical value (test #1 in Table 6.2). 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 in Table 6.2). Third, we may wish to compare whether the outcome variable differs between three or more levels of a categorical variable with three or more sub-groups (test #4 in Table 6.2). The factor variable is the categorical variable that we use to define the groups (e.g., three types of promotion campaigns).

Table 6.2 Selecting an appropriate test

Test #	Testing situation	Nature of samples	Choice of test			Region of rejection
	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 samples	If either Levene's test: $\sigma_1^2 = \sigma_2^2$ or Shapiro-Wilk test = normal	Paired samples *t*-test		One or two-sided
			Levene's test: $\sigma_1^2 \neq \sigma_2^2$ & Shapiro-Wilk test ≠ normal		Wilcoxon matched-pairs signed-rank test	One or two-sided

Table 6.2 (Continued)

Test #	Testing situation	Nature of samples	Choice of test			Region of rejection
	What do we compare	*Paired vs. independent*	*Assumptions*	*Parametric*	*Non-parametric*	*One or two-sided test*
3	Outcome variable across two groups	Independent samples	Levene's test: $\boldsymbol{\sigma}_1^2 = \boldsymbol{\sigma}_2^2$	Independent samples *t*-test		One or two-sided
			Shapiro-Wilk test = normal & Levene's test: $\sigma_1^2 \neq \sigma_2^2$	Independent samples *t*-test with Welch's correction		One or two-sided
			Shapiro-Wilk test ≠ normal & Levene's test: $\sigma_1^2 \neq \sigma_2^2$		Mann-Whitney U 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*
			Shapiro-Wilk test ≠ 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 ≠ normal & Levene's test: $\sigma_1^2 \neq \sigma_2^2$		Kruskal-Wallis rank test	Two-sided*

* 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.

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 the samples being compared are paired or independent. 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 is the most serious threat to increases the probability of type I errors (Van Belle 2008). 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 and #4 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.

- **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. Large deviations will therefore relate to p-values smaller than 0.05, suggesting that the sample scores are not normally distributed. An alternative strategy to check visually for normality, which we discuss in Box 6.2.

Box 6.2 Visual check for normality

You can also use plots to visually check for normality. The **quantile plot** (or *Q–Q plot* in SPSS) is a type of probability plot, which compares the quantiles of the sorted sample values with the quantiles of a standard normal distribution. Plotted data that do not follow the straight line reveal departures from normality. The quantile plot is useful to spot non-normality in the tails of the distribution. It is also possible to plot a histogram with a normal curve (discussed in ► Sect. 5.5.1) which is typically most useful when distribution is not quite symmetrical. Note that visual checks are subjective and should always be used in combination with more tests for normality such as the Shapiro-Wilk 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.

■ Equality of variances test

We can 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 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.

■ 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-samples *t*-test** is appropriate. **Independent samples *t*-tests** consider two distinct groups, such as males versus females, or users versus non-users. The **paired samples *t*-tests** is used to test for differences between the same set of twice observed objects (usually respondents). When we are interested in differences between the means of more than two groups of respondents, 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. For 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.

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). Therefore, when 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 (Norman, 2010), provided we have sample sizes greater than 30. The same holds for the ANOVA in cases where the sample sizes per group exceed 30. Nevertheless, if non-normality is an issue, you should use a non-parametric test that is not based on the normality assumption.

We may 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-size sample sizes are different, we recommend using parametric tests, such as the two-samples *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. 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: an independent samples *t*-test with Welch's correction or a one-way ANOVA *F*-test with Welch's correction.

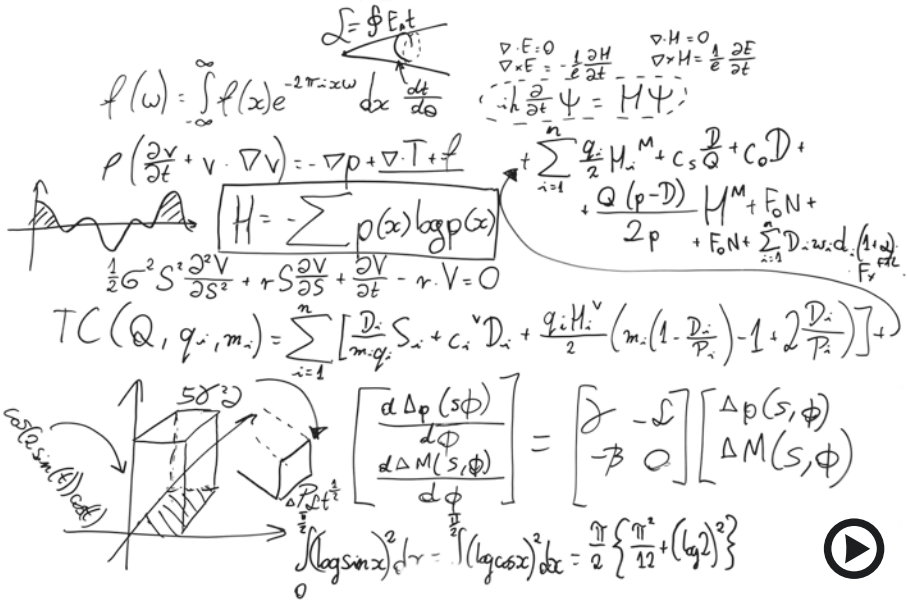
Finally, when both the normality and equality of variance assumptions are violated, non-parametric tests can be chosen directly. In the following, we briefly discuss these non-parametric tests.

■ 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 samples *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.[2] 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** (labelled Kruskal-Wallis H in SPSS) 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. In the ↓ Web Appendix (→ Downloads), we discuss the χ^2-tests.

2 The exact calculation of this test is shown on https://www.ibm.com/support/knowledgecenter/en/SSLVMB_20.0.0/com.ibm.spss.statistics.help/alg_npar_tests_mannwhitney.htm



https://www.guide-market-research.com/app/download/13488667027/SPSS+3rd_Chapter+6_Chi-square+test.pdf?t=1516713011

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 (a **one-tailed test**) or on both sides (a **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., ≠ 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.

Even for directional hypotheses, researchers typically use two-tailed tests (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 (e.g., Ruxton and Neuhaeuser 2010; Van Belle 2008), one-tailed tests should only be used when the opposite direction is theoretically meaningless or impossible (e.g., Field 2013; Kimmel 1957). For example, when testing if sales number

of innovations are greater than zero, it does not make sense to consider negative values as negative sales cannot occur. In such a situation testing for negative outcomes is meaningless because such possibilities are ruled out beforehand. 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.

Overall, the region of rejection for the one-sample *t*-test and the two-samples *t*-test can either be one or two-tailed (however, we recommend the use of two-tailed tests). Although the alternative hypothesis in the ANOVA analysis is non-directional by nature, the underlying *F*-statistic—used to make inferences about group differences—is based on the *F*-distribution that is right-skewed. Specifically, the *F*-statistic is a ratio of two variances and as variances are always positive, the *F*-ratio is never negative. This means that although the underlying alternative hypothesis for the ANOVA analysis is two-sided, all the group differences are assumed to be in the same side of the sampling distribution. We discuss this point in more detail later in this chapter.

6.3.4 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 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:

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

Here $\overline{x}$ is the sample mean, μ is the hypothesized population mean, and $s_{\overline{x}}$ 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 $\overline{x}$ and the hypothesized population mean μ. If the point of sale display was highly successful, we would expect $\overline{x}$ 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:

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

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

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

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 this 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 6.3).

Box 6.3 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), SPSS uses the *t*-test. The *z*-test's statistic closely resembles the *t*-test and is calculated as follows:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

The only minor difference is that we do not write $s_{\bar{x}}$ but $\sigma_{\bar{x}}$ in the denominator. It's the same principle, but the Greek symbol indicates that the measure refers to a (known) population value and not to the sample. If, for example, we assumed that standard deviation in the population is 2.5, we would obtain the following test statistic:

$$\sigma_{\bar{x}} = \frac{2.5}{\sqrt{10}} = 0.791 \text{ and, finally, } z = \frac{2.30}{0.791} = 2.909$$

6

6.3.5 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 6.4.

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, when α is 0.05, the cumulative probability distribution is $1-\alpha/2$ or $1-0.05/2 = 0.975$. We divide the significance

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*.

Box 6.4 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.

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 ± 2.262 on both sides of the distribution. This is indicated by the two rejection regions left and right on the intersection of the vertical line and the curve. These two rejection areas are the upper 2.5 % and bottom 2.5 % while the remaining 95 % non-rejection region is in the middle. Since the test statistic t_{test} (indicated by the line saying t_{test} 2.274) falls in the right rejection 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.

Figure 6.3 is static but using the animation on https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html, you can change, for example, the rejection areas and/or toggle between one- and two-sided rejection regions.

4 To obtain the critical value, you can also use the TINV function provided in Microsoft Excel, whose general form is "TINV(*α, df*)." Here, *α* represents the desired Type I error rate and *df* the degrees of freedom. To carry out this computation, open a new Excel spreadsheet and type in " = TINV(2*0.025,9)." Note that we have to specify "2*0.025" (or, directly 0.05) under *α*, because we are applying a two-tailed instead of a one-tailed test.

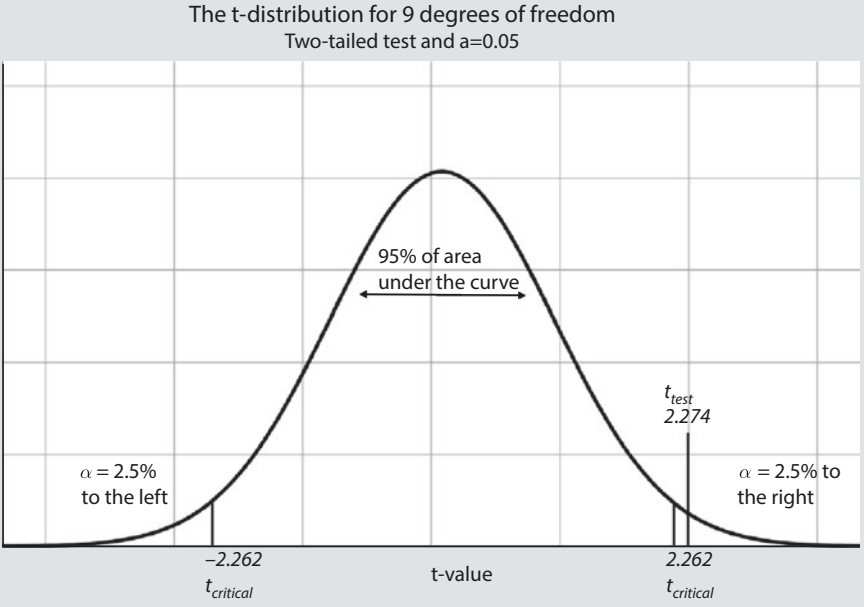


Fig. 6.3 Relationship between test value, critical value, and *p*-value

© Wittayayut/Getty Images/iStock
https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html

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 SPSS. SPSS 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).

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) and a suitable threshold (Benjamin et al. 2017; Lakens et al. 2017).

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 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

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 rule doesn't always apply such as for exact tests of probabilities.

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

Tip

There is another way of hypothesis testing without applying a significance test. Instead of following the steps just described, we can construct a 100(1–*α*)% **confidence interval** for μ. The confidence interval is a range of values calculated, which includes the desired true parameter (here, the population mean μ) with a certain probability that we need to define in advance. The confidence interval has a number of critical ingredients, the estimate itself, the standard deviation of the sample, the sample size, the distribution belonging to these estimates, and the confidence level (typically 95 % but 90 % or 99 % are also common). The boundaries of the mean's confidence interval are defined as:

- Lower boundary: $x - t_{critical}\left(\frac{\alpha}{2}\right) \cdot s_{\bar{x}}$
- Upper boundary: $\bar{x} + t_{critical}\left(\frac{\alpha}{2}\right) \cdot s_{\bar{x}}$

In other words, the confidence interval is defined as:
$\left(x - t_{critical}\left(\frac{\alpha}{2}\right) \cdot s_{\bar{x}}; \bar{x} + t_{critical}\left(\frac{\alpha}{2}\right) \cdot s_{\bar{x}}\right)$. In our case, the confidence interval looks like this:

$$(47.3 - 2.262 \cdot 1.012; 47.3 + 2.262 \cdot 1.012) = (45.01; 49.59)$$

From this, we would conclude that the true population mean, with 95 % certainty, lies in the range of 45.01 and 49.59. As we can see, the 95 % confidence interval does not include 45. Hence, we reject the null hypothesis and conclude that the population mean is not 45. Indeed, from the lower and upper boundaries, we can see that the population mean is likely larger than 45.

6.3.6 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.

8 We don't have to conduct manual calculations and tables when working with SPSS. However, we can calculate the *p*-value 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., https://graphpad.com/quickcalcs/pvalue1.cfm) that calculate *p*-values and test statistic values.

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 a 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 two-sided null and alternative hypotheses are now:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

The test statistic of the two independent samples *t*-test—which is now 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:

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

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

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

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:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

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:

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

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

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

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

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

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

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 need to look at the column labeled $\alpha = 0.025$ and the line labeled $df = 18$). The absolute value of −2.819 is greater than 2.101 and, thus, falls within the bottom 2.5 % of the distribution (◘ Table 6.3). 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 we have a paired-samples case as the sample is the same before and after. ◘ Table 6.4 shows the sales of the 10 stores before and after the point of display and in each case showing the sales in units. You can again assume that the data are normally distributed.

Table 6.4 Sales data (extended)

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

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 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:

$$H_0 : \mu_d = 0$$
$$H_1 : \mu_d \neq 0$$

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:

$$d_1 = 50 - 46 = 4$$
$$d_2 = 53 - 51 = 2$$
$$\ldots$$
$$d_9 = 51 - 49 = 2$$
$$d_{10} = 44 - 43 = 1$$

Based on these results, we calculate the mean difference:

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

as well as the standard error of this difference:

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

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

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 $\alpha = 0.025$. 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 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:

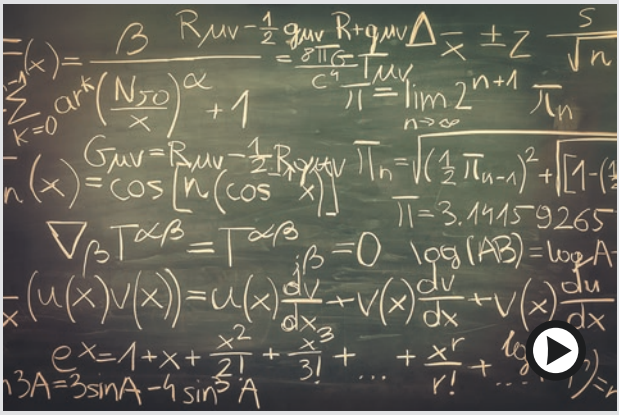
1. The point of sale display versus the free tasting stand,
2. the point of sale display versus the in-store announcements, and
3. the free tasting stand versus the in-store announcements.

While three comparisons are doable, 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]

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 the *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.



© domin_domin/Getty Images/iStock
https://www.guide-market-research.com/app/download/13488670527/SPSS+3rd_Chapter+6_Two-way+ANOVA.pdf?t=1516713104

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

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

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. ANOVA is even more flexible, because you can also integrate interval or ratio-scaled independent variables and even multiple dependent variables as well as more than one factor variable (**Two-way ANOVA**).

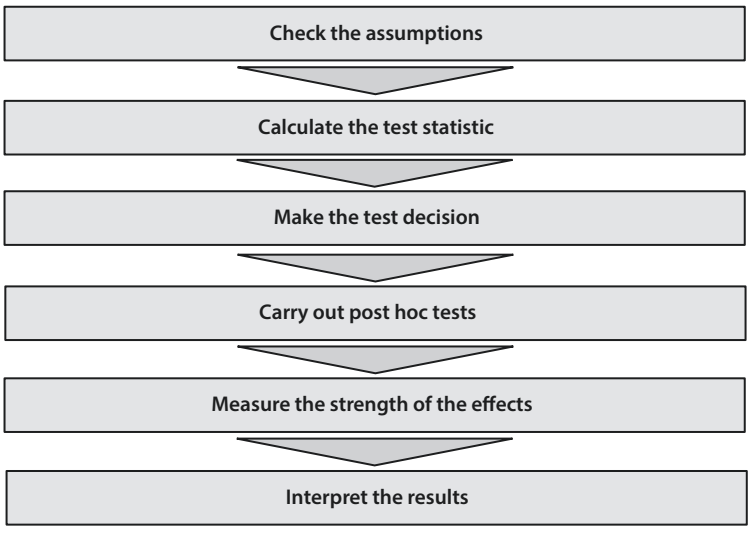
We now know that the one-way ANOVA is used to examine the mean differences between more than two groups. In more formal terms, the objective of the one-way ANOVA is to test the null hypothesis that the population means of the groups (defined by the factor and its levels) are equal. If we compare three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

$$H_0: \mu_1 = \mu_2 = \mu_3$$

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:

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

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.

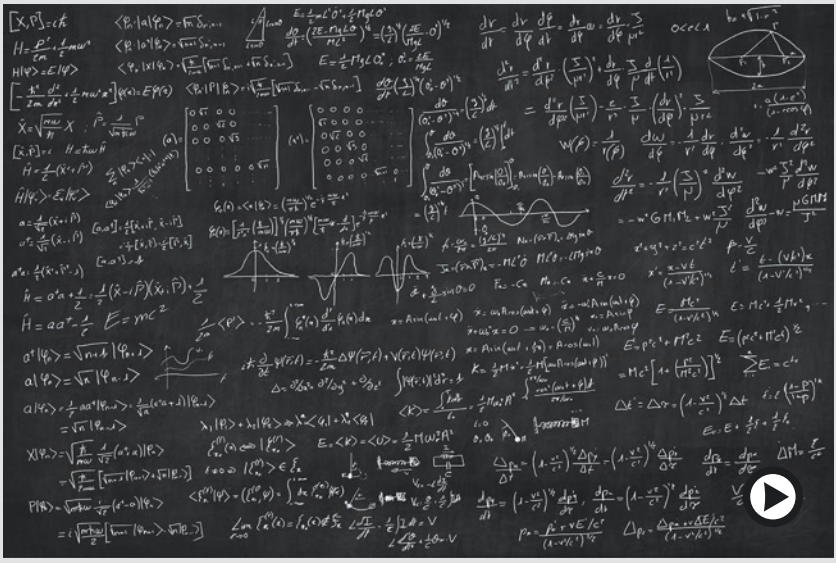


◘ **Fig. 6.4** Steps in conducting an ANOVA

6.5.1 Check the Assumptions

ANOVA is a parametric test that relies on the same distributional assumptions as discussed in ► Sect. 6.3. 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 per group, 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.[10] 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 (→ Downloads).



© virtualphoto/Getty Images/iStock
https://www.guide-market-research.com/app/download/13488668127/SPSS+3rd_Chapter+6_Levene%27s+test.pdf?t=1516713071

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

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.

6.5.2 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).[11]

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.

- **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 stand, and in-store announcements) relative 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 stem from different groups can explain. 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.

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

■ The Within-Group Variation (*SSW*)

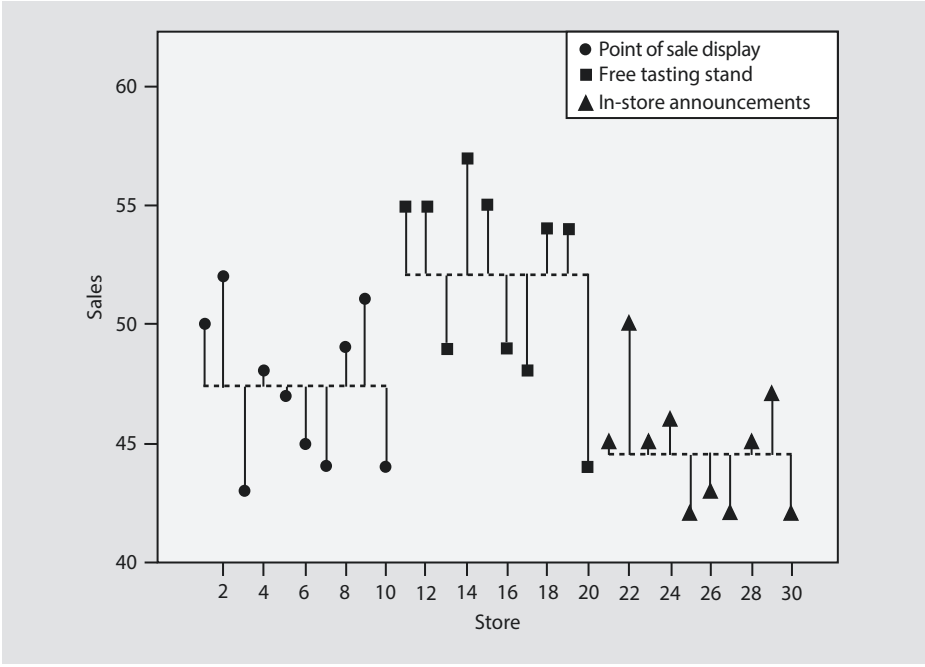
As the name already suggests, SS_W describes the variation in the dependent variable within each of the groups. In our example, SS_W simply represents the variation in sales in each of the three treatment conditions. The smaller the variation within the groups, the greater the probability that the grouping 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. While 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**.

■ 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 (▲).



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

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 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. We can calculate the three types of variation as follows:

1. The total variation, computed by comparing each store's sales with the overall mean $\bar{x}$, which is equal to 48 in our example:[12]

$$SS_T = \sum_{i=1}^{n} (x_i - \bar{x})^2 = (50-48)^2 + (52-48)^2 + \ldots + (47-48)^2 + (42-48)^2 = 584$$

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

$$SS_B = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2$$

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:

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

12 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:

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

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:

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

6.5.3 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 occur, it makes no sense to split the significance level in half (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.

6.5.4 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.

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—SPSS provides no less than 18! Generally, these tests detect pairs of groups whose mean values do not differ significantly (*homogeneous subsets*). However, all these tests are based on different assumptions and designed for different purposes, whose details are clearly beyond the scope of this book. Check out the SPSS help function for an overview and references.

The most widely used post hoc test in market research is **Tukey's honestly significant difference test** (usually simply called *Tukey's HSD*). Tukey's HSD is a very versatile test which controls for the type I error and is conservative in nature. A less conservative alternative is

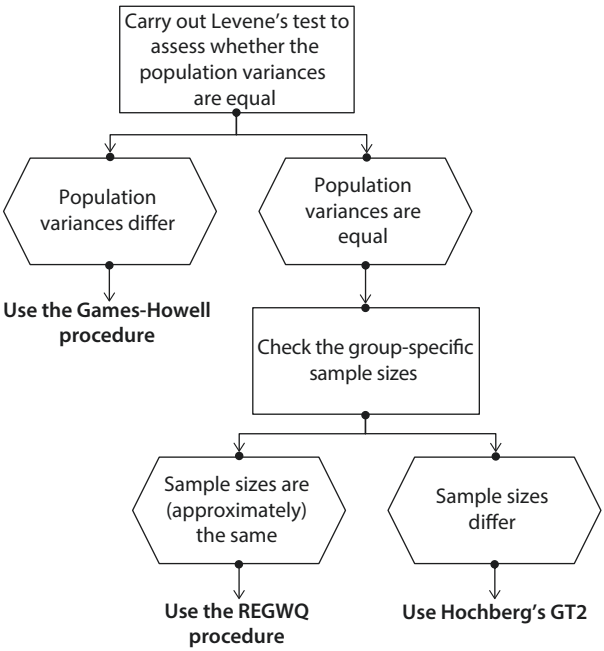


Fig. 6.6 Guidelines for choosing the appropriate post hoc test

the *Ryan/Einot-Gabriel/Welsch Q* (*REGWQ)* procedure, which also controls for the type I error rate but has a higher statistical power. These post hoc tests share two important properties:

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

Fortunately, research has provided alternative post hoc tests for situations in which these properties are not met. When sample sizes differ clearly, it is advisable to use *Hochberg's GT2*, which has good power and can control the type I error. However, when population variances differ, this test becomes unreliable. Thus, in cases where our analysis suggests that population variances differ, it is best to use the *Games-Howell procedure* because it generally seems to offer the best performance. Fig. 6.6 provides a guideline for choosing the appropriate post hoc test.

While post hoc tests provide a suitable way of carrying out pairwise comparisons among the groups while maintaining the familywise error rate, they do not allow making any statements regarding the strength of a factor's effects on the dependent variable. This is something we have to evaluate in a separate analysis step, which is discussed next.

6.5.5 Measure the Strength of the Effects

We can compute the η^2 (the **eta squared**) coefficient manually 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 (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 ($\mathbf{R^2}$), known from regression analysis (► Chap. 7). 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 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 $n \leq 50$ and η^2 for $n > 50$.

Tip

It is difficult to provide firm rules of thumb regarding when η^2 or ω^2 is appropriate, 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.

6.5.6 Interpret the Results

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.

Table 6.5 provides an overview of steps involved when carrying out the following tests in SPSS: One-sample *t*-test, independent samples *t*-test, paired samples *t*-test, and the one-way ANOVA.

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

Theory (number in brackets referring to Table 6.2)	Action
(1) One-sample t-test	
Formulate the hypothesis	
Formulate the study's hypothesis:	*For example:* $H_0: \mu = \#$ $H_1: \mu \neq \#$

Table 6.5 (Continued)

Theory (number in brackets referring to Table 6.2)	Action
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?	*Check for normality* Go to ► Analyze ► Descriptive Statistics ► Explore. Then add the test variable(s) to the **Dependent list** box. Click on **Plots**, uncheck **Stem-and-leaf** and check **Normality plots with tests** and click on **Continue** and then **OK**. Interpret the **Tests of Normality** table and check in the **Tests of Normality** table under **Sig.** of **Shapiro-Wilk** if $p > 0.05$, which suggests normality.
Specify the type of t-test	
Is the test one or two-sided?	Determine the region of rejection, one-sided (left or right) or two-sided.
Calculate the test statistic	
Specify the test variable and the fixed value	Either normally distributed or equal group variances (one sample *t*-test): Go to ► Analyze ► Compare Means ► One-Sample T test. Put the test variable(s) under **Test Variable(s):**. Then enter under **Test Value** the mean value you want to compare the test variable(s) against. Note: it is possible to compare multiple test variables but no Bonferroni correction is applied and it may not be appropriate to compare multiple test variables against one test value. Consider selecting a different confidence interval different from the default 95 % by clicking on **Options** and to enter the desired confidence interval (e.g. 99 or 90) under **Confidence Interval percentage:**. Not normally distributed and unequal group variances (Wilcoxon signed-rank test): First you will need to compute a new variable with the fixed value you want to test for. Do this by going to ► Transform ► Compute variable and as the target variable type (for example) *testval* and enter the value you want to test for under **Numeric Expression**. Click on **OK**. Then go to ► Analyze ► Nonparametric Tests ► Legacy Dialogs ► 2 Related Samples. Then add the variable you want to test for to Test Pairs: and also add the variable you just constructed (e.g. *testval*). Make sure only **Wilcoxon** is ticked and click on **OK**.

Table 6.5 (Continued)

Theory (number in brackets referring to Table 6.2)	Action
Interpret the results	
Look at the test results	One sample *t*-test: In the output tables, check in the **One-Sample Test** table if **Sig. (2-tailed)** is less than .05. For two-sided tests, you need to multiply the value under **Sig. (2-tailed)** by two. Wilcoxon signed-rank test, check under **Test Statistics** under the row **Asymp. Sig. (2-tailed)** if the value is less than .05.
What is your conclusion?	Reject the null hypothesis that the population mean of the test variable is equal to the mean value you want to compare the test variable(s) against if the *p*-value is lower than 0.05.
(2) 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.
Are the test variables normally distributed in each of the groups?	*Check for normality* Go to ► Analyze ► Descriptive Statistics ► Explore. Then add the test variable(s) to the **Dependent list:** and under **Factor List:** add the variable that indicates the two groups. Then click on Plots, uncheck **Stem-and-leaf** and check **Normality plots with tests** and click on **Continue** and then **OK**. Interpret the **Tests of Normality** table and check in the **Tests of Normality** table under **Sig.** of **Shapiro-Wilk** if $p > 0.05$, which suggests normality.
Specify the type of *t*-test	If the data appear to be normally distributed across all groups or have equal group variances, we can proceed with the paired samples *t*-test. If the data are non-normally distributed and have unequal variances, we should apply the Wilcoxon matched-pairs signed-rank test.

Table 6.5 (Continued)

Theory (number in brackets referring to Table 6.2)	Action
Is the test one or two-sided?	Determine the region of rejection.
Calculate the test statistic	
Select the paired test variables	Normally distributed or equal group variances (Paired *t*-test): Go to ► Analyze ► Compare Means ► Paired-Samples T test. Put the test variables into the **Paired Variable(s)** box. Note: it is possible to compare multiple test variables but no Bonferroni correction is applied. Consider selecting a different confidence interval different from the default 95 % by clicking on **Options** and to enter the desired confidence interval (e.g. 99 or 90) under **Confidence Interval percentage**. – Not normally distributed and unequal variances (Wilcoxon matched-pairs signed-rank test): Go to ► Analyze ► Nonparametric tests ► Related Samples and click on **Customize analysis**. Under **Fields** click on **Use custom field assignments** first. Then add the variable you want to test to the **Test Fields:** box. Then click on **Settings** and tick **Customize tests:** and then **Wilcoxon matched-pair signed-rank (2 samples)** and click on **Run**.
Interpret the results	
Look at the test results	– Paired *t*-test: test in the **Paired Samples Test** table if **Sig. (2-tailed)** is below .05. – Wilcoxon matched-pairs signed-rank test: in the Hypothesis Test Summary, check if the *p*-value under **Sig.** is below .05.
What is your conclusion?	Reject the null hypothesis that the population mean of the test variables are equal if the *p*-value under **Sig. (2-tailed)** is below 0.05.
(3) Independent 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.

Table 6.5 (Continued)

Theory (number in brackets referring to Table 6.2)	Action
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 in each of the groups and are the group variances the same?	*Check for normality* Go to ► Analyze ► Descriptive Statistics ► Explore. Then add the test variable(s) to the **Dependent list:**. Click on Plots, uncheck **Stem-and-leaf** and check **Normality plots with tests** and click on **Continue** and then **OK**. Interpret the **Tests of Normality** table and check in the **Tests of Normality** table under **Sig.** of **Shapiro-Wilk** if $p > .05$, which suggests normality. SPSS conducts the group variances equality test automatically as part of the Paired samples t-test.
Specify the type of t-test	
Is the test one or two-sided?	Determine the region of rejection.
Calculate the test statistic	
Select the test variable and the grouping variable	Normally or non-normally distributed but equal group variances (independent samples *t*-test): Go to ► Analyze ► Compare Means ► Independent Samples T test. Put the test variable into the **Test Variable(s)** box and add the variable indicating the grouping into the **Grouping Variable** box. Click on **Define Groups** and under **Use specified values** enter which values indicate the two groups you want to compare. Note: it is possible to compare multiple test variables but no Bonferroni correction is applied. Consider selecting a confidence interval different from the default 95 % by clicking on **Options** and to enter the desired confidence interval (e.g. 99 or 90) under **Confidence Interval percentage**. Not normally distributed and unequal variances (Mann-Whitney U test): Go to ► Analyze ► Nonparametric tests ► Independent Samples. Under **Objective** tick **Customize analysis**. Under **Fields**, click on **Use customer field assignments** and move the test variable into the **Test Fields** box. Under **Settings** click on **Customize tests** and click on **Mann-Whitney U (2-samples). Followed by Run**.

Table 6.5 (Continued)

Theory (number in brackets referring to Table 6.2)	Action
Interpret the results	
Look at the test results	If the data appear to be normally or non-normally distributed with equal group variances then from the **Independent Samples Test** table we should read the row saying **Equal variance assumed**. If they come from normally distributed data but unequal variances we should use the row saying **Equal variances not assumed.** For both situations, check if **Sig. (2-tailed)** is less than 0.05. For two-sided tests, you need to multiply the value under **Sig. (2-tailed)** by two. For the Mann-Whitney U test, we should check in the **Hypothesis Test Summary** table if **Sig. (2-tailed)** is less than 0.05. For two-sided tests, you need to multiply the value under **Sig. (2-tailed)** by two.
What is your conclusion?	Reject the null hypothesis that the population mean of the test variables are equal if the *p*-value under **Sig. (2-tailed)** is lower than 0.05.
(4) 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.
Are the observations independent?	Next, consult ► Chap. 3 to determine whether the observations are independent.

Table 6.5 (Continued)

Theory (number in brackets referring to Table 6.2)	Action
Is the test variable normally distributed and are the group variances the same?	*Check for normality* Go to ► Analyze ► Descriptive Statistics ► Explore. Then move the test variable into the **Dependent list** box and under **Factor List:** add the variable that indicates the two groups. Click on **Plots**, uncheck **Stem-and-leaf** and check **Normality plots with tests** and click on **Continue** and then **OK**. Interpret the **Tests of Normality** table for all groups and check in the **Tests of Normality** table under **Sig.** of **Shapiro-Wilk** if $p > 0.05$, which suggests normality. *Check for Equality of Variances Assumption* This test is included in the ANOVA analysis (in the ANOVA dialog box go to **Options** and tick **Homogeneity of variance test** and click on **Continue**.
Select the type of the test	If the assumption of normality and equality of the variance are met, proceed with the one-way ANOVA analysis. If not, conduct the Kruskal-Wallis rank test.
Calculate the test statistic	
Specify the dependent variable and the factor (grouping variable)	Normally or non-normally distributed but equal group variances (One-way ANOVA: *F*-test): Go to ► Analyze ► Compare Means ► One-Way ANOVA. Put the test variable into the **Dependent List** box and add the variable indicating the grouping into the **Factor** box. Click on **Options** and also tick **Homogeneity of variance test**. Normally or non-normally distributed but equal group variances (One-way ANOVA: *F*-test with Welch's correction): Go to ► Analyze ► Compare Means ► One-Way ANOVA. Put the test variable into the **Dependent List** box and add the variable indicating the grouping into the **Factor** box. Click on **Options** and tick **Homogeneity of variance test** and **Welch**. Not normally distributed and unequal variances (Kruskal-Wallis rank test): Go to ► Analyze ► Nonparametric Tests ► Legacy Dialogs ► K Independent Samples. Put the test variable into the **Test Variable List** box. Add the factor into the **Grouping Variable** box after which you should click on **Define Range** to set the **Minimum** and **Maximum** value of the factor and click on **Continue**. Make sure also **Kruskal-Wallis H** is ticked (only). Click on **OK**.

Table 6.5 (Continued)

Theory (number in brackets referring to Table 6.2)	Action
Interpret the results	
Look at the test results	Compare the *p*-value in the **ANOVA** table with the significance level under **Sig.** The *p*-value should be lower than 0.05 to reject the null hypothesis. For the One-way ANOVA: *F*-test with Welch's correction, check the table **Robust Tests of Equality of Means** under **Sig.** The *p*-value should be lower than 0.05 to reject the null hypothesis. For the Kruskal-Wallis rank test, check the table **Test Statistics** under **Asymp. Sig.** is lower than 0.05 to reject the null hypothesis.
Carry out pairwise comparisons	You can carry out post hoc tests as part of the ANOVA analysis. In the **Univariate** dialog box, click on **Post Hoc** and move the factor(s) to the **Post Hoc Tests for:** box. If you found evidence of equal variances, tick **Tukey**. If you found unequal variances, tick **Games-Howell**.
Look at the strength of the effects	In the Univariate dialog box, go to ► **Options** and tick **Estimates of effect size**. 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 in the **Paired Comparisons** table. If the *p*-values under **Sig.** 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 **Tests of Between Subjects Effects** table, reject the null hypothesis that at least two population means are equal if the *p*-value is lower than 0.05 for the corrected model.

6.6 Example

Let's now turn to the Oddjob Airways case study and apply what we 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?

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*).

6.6.1 Research Question 1

6.6.1.1 Formulate the 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 that the overall price/performance satisfaction means of male and female travelers differs.

6.6.1.2 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.

6.6.1.3 Select the 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 is to compare 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 they are independent. Next, we need to check if the dependent variable *overall_sat* is normally distributed. To do this go to ► Analyze ► Descriptive Statistics ► Explore. Then put the test variable *overall_sat* into the **Dependent list** box. Under **Factor List:** enter the *gender* variable as you want to assess normality within each of the two groups. For the ANOVA example that follows, we also need the *status* variable so please add this as well as it avoids having to redo the analysis later. All of this is shown in ◘ Fig. 6.7. Next, click on **Plots**, which will open a dialog box similar to ◘ Fig. 6.8, uncheck **Stem-and-leaf**, check **Normality plots with tests** and click on **Continue** and then **OK**.

◘ Table 6.6 displays the **Tests of Normality** table for both *gender* and *status*. SPSS also reports both the Kolmogorov-Smirnov test (which we do not need) and the Shapiro-Wilk test with its corresponding *p*-values under (**Sig.**). The results show that the *p*-values of the Shapiro-Wilk test (**.000**) are smaller than 0.05, indicating that the normality assumption is violated for both *gender* groups and all tree *status* groups. SPSS also produces five quantile plots in ◘ Fig. 6.9 as discussed in Box 6.2. In every plot, the dots appear to follow the straight line reasonably well and do not clearly suggest where the normality is off.

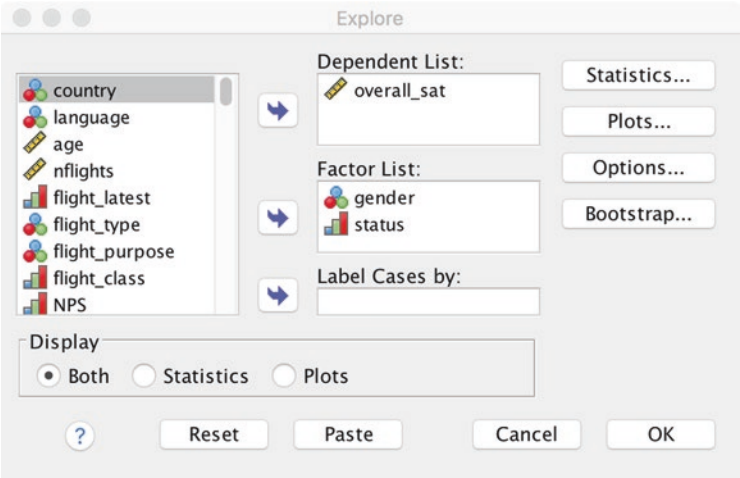
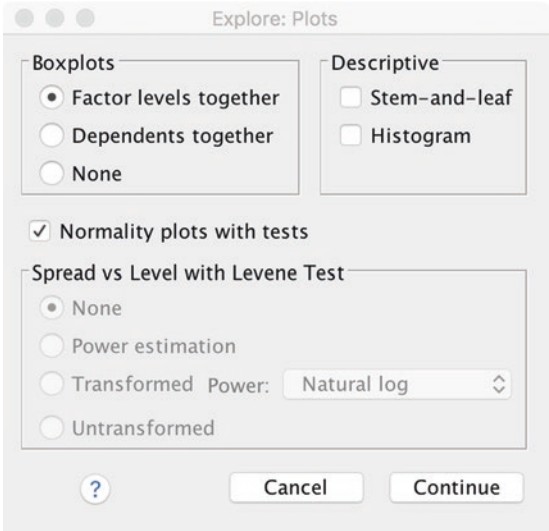


Fig. 6.7 Testing for normality

Fig. 6.8 Normality plots with tests



6.6.1.4 Calculate the Test Statistic and Make the Test Decision

As we find no support for normality, we may have to use the independent samples *t*-test or the Mann-Whitney U test. The decision on which to choose depends on whether the variances are equal. To do this, go to ► **Analyze** ► **Compare Means** ► **Independent Samples T test**, which will open a dialog box similar to ◘ Fig. 6.10.

Put the test variable *overall_sat* into the **Test Variable(s)** box and add *gender* into the **Grouping Variable** box. Next, click on **Define Groups** and enter *1* next to **Group 1** and *2* next to **Group 2**. Proceed by clicking on **Continue** and then **OK**.

Tip

To indicate different groups, you can also specify a **Cut point**, which is particularly useful when you want to compare two groups based on an ordinal or continuous variable. For example, if you want to compare younger vs. older members you could put all members below 30 years of age into one category and all who are 30 or above into the other category. When you indicate a cut point, observations with values less than the cut point form one group, while observations with values greater than or equal to the cut point form the other group.

Table 6.6 Tests of normality

Tests of Normality

	Gender	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Overall, I am satisfied with the price performance ratio of Oddjob Airways.	female	.201	280	.000	.929	280	.000
	male	.167	785	.000	.941	785	.000

[a] Lilliefors Significance Correction

Tests of Normality

	Traveler Status	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Overall, I am satisfied with the price performance ratio of Oddjob Airways.	Blue	.181	677	.000	.930	677	.000
	Silver	.161	245	.000	.947	245	.000
	Gold	.183	143	.000	.941	143	.000

[a] Lilliefors Significance Correction

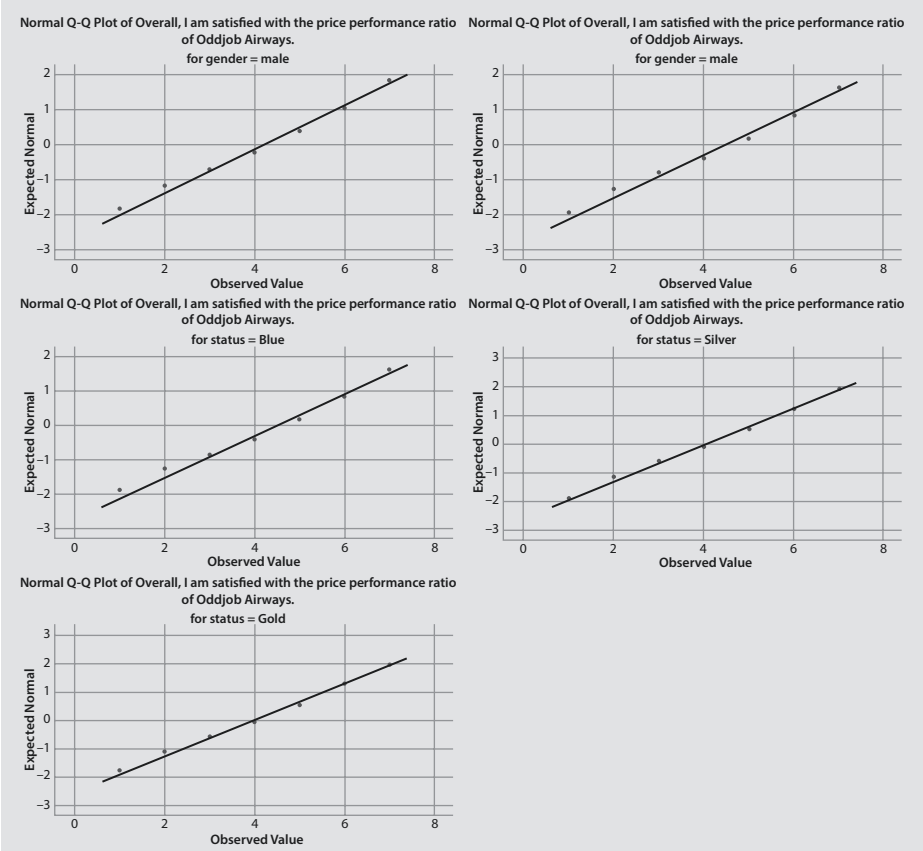


Fig. 6.9 Normal Q–Q plot output

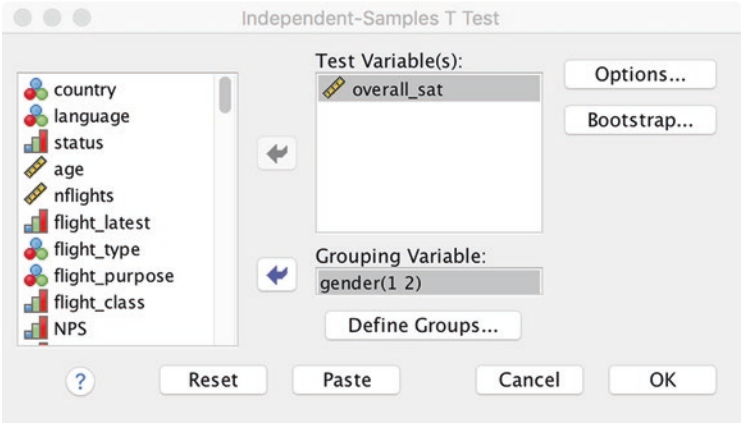


Fig. 6.10 Independent samples *t*-test

Table 6.7 Group statistics

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
overall_sat	female	280	4.5000	1.64611	.09837
	male	785	4.2369	1.61306	.05757

SPSS will produce Tables 6.7 and 6.8. Table 6.7 shows the sample size (**N**) which is **280** females and **785** males. The mean (**Mean**) scores are also show as **4.500** and **4.2369** respectively and these mean differences, at first sight, appear different between the two groups. However, as we learned before, we have to take the variation in the data into account to test whether this difference is also present in the population. The standard deviation (**Std. Deviation**) gives some indication but the formal test is shown in Table 6.8. On the left of the output, we can see the test results of Levene's test for the equality of population variances. The low *F*-value (**F**) of **0.418** suggests that we cannot reject the null hypothesis that the population variances are equal. This is also mirrored in the large *p*-value of **0.518**, which lies far above 0.05. Because we obtained evidence that the variances are equal, we also find that the independent samples *t*-test is appropriate. We can thus proceed to interpret the output further.

If both the normality and equality of the variance assumptions were not met, the Mann-Whitney U test should have been performed. This is the non-parametric counterpart of the independent samples *t*-test. Please see Table 6.5 for details on how to carry out this test.

6.6.1.5 Interpret the Results

Looking at the central and right part of the output in Table 6.8, we can see that SPSS carries out two tests, one based on the pooled variance estimate (upper row) and the other based on separate variance estimates using Welch's correction (lower row). Since we assume that the population variances are equal, we should interpret the upper row. When comparing the *p*-value under **Sig. (2-tailed)** with the significance level, we learn that the *p*-value (**0.020**) is smaller than the significance level (0.05). Therefore, we can reject the independent samples *t*-test's null hypothesis that there is no difference in satisfaction between female and male customers and conclude that that the overall price/performance satisfaction differs significantly between female and male travelers.

6.6.2 Research Question 2

In the second research question, we examine whether customers' membership status (i.e., *status*) relates to their overall price/performance satisfaction (i.e., *overall_sat*) with *Oddjob Airways*. The status *variable* defines three groups: *Blue, Silver*, and *Gold*. Again, we start by formulating a null hypothesis that is again non-directional in nature, expecting that

Table 6.8 Independent samples *t*-test

Independent Samples Test

		Levene's Test for Equality of Variances	t-test for Equality of Means							
									95 % Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
overall_sat	Equal variances assumed	.418	.518	2.330	1063	.020	.26306	.11289	.04154	.48457
	Equal variances not assumed			2.308	482.698	.021	.26306	.11398	.03909	.48702

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.

6.6.2.1 Check the Assumptions

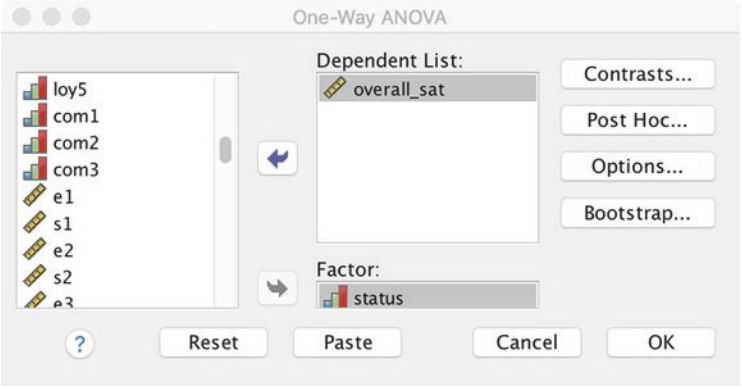
In checking the assumptions, we already know that the sample is a random subset of the population and we also know that other respondents' responses do not influence those of the respondents (i.e., they are independent). Next, we check the normality but as we had already tested for normality of *overall_sat* for the three loyalty groups (*Blue, Silver*, and *Gold,* see ◘ Table 6.6) we know that this variable is not normally distributed. If we thus follow ◘ Table 6.2., we should either conduct a one-way ANOVA *F*-test when the variances are equal or a Kruskal-Wallis rank test when the variances are not equal.

To test for the equality of variances, we need to carry out an ANOVA analysis first. To do this, go to ► Analyze ► Compare Means ► One-Way ANOVA, which opens a dialog box similar to ◘ Fig. 6.11. Move the test variable *overall_sat* into the **Dependent List** box and add the variable *status* that indicates the three groups (*Blue, Silver*, and *Gold*) into the **Factor** box.

Next, under **Post Hoc** (◘ Fig. 6.12) we should specify a series of post hoc tests for multiple group comparisons that are all shown in ◘ Fig. 6.12. Since we do not yet know the result of Levene's test, we choose Ryan/Einot-Gabriel/Welsch Q (**R-E-G-W Q**) and **Games-Howell**. Since group sizes are fairly unequal, we also select **Hochberg's GT2**. Next, click on **Continue** to get back to the main menu.

Go back to the main menu and click on **Options**. In the dialog box that opens, (◘ Fig. 6.13), tick **Descriptive** and **Homogeneity of variance test**. Note that if you were to have support for normality (which we don't), it is useful to tick **Welch** as well. Go back to the main menu by clicking on **Continue** and then **OK**.

The descriptive statistics in ◘ Table 6.9 show that the groups clearly differ in size. For example, the *Blue* group comprises 677 customers, whereas the *Gold* group comprises 143 customers. The results also indicate that the means of *overall_sat* differ between the



◘ **Fig. 6.11** One-way ANOVA dialog box

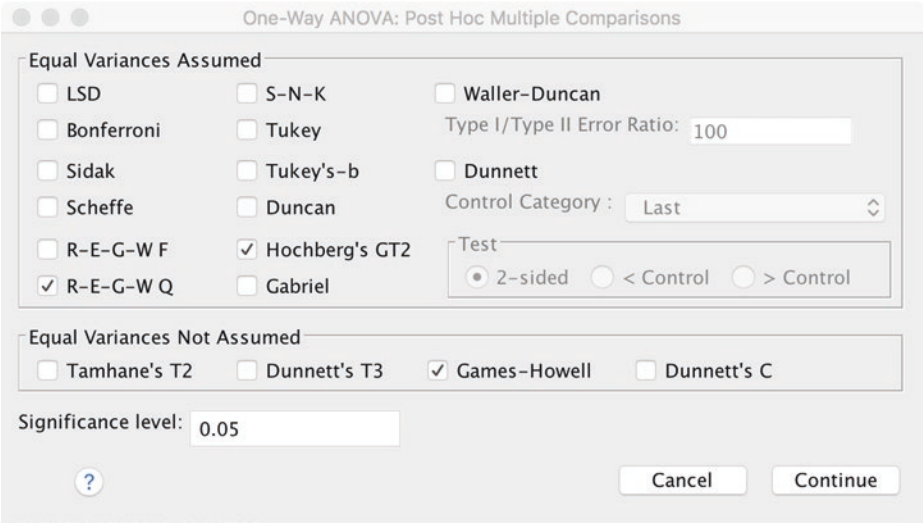
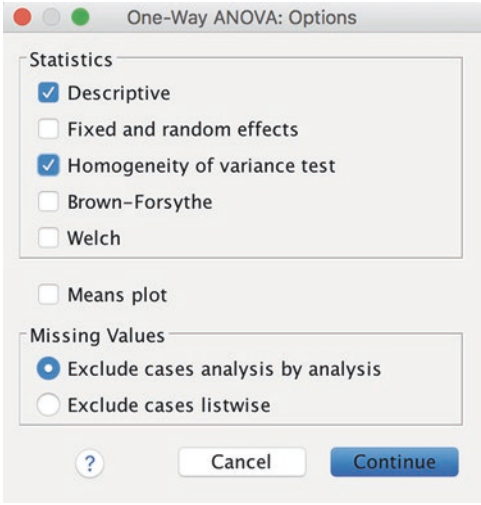


Fig. 6.12 Post hoc multiple comparisons dialog box

Fig. 6.13 One-way ANOVA options



three groups. For example, in the *Blue* group, the mean is 4.47 compared to 3.99 in the *Gold* group. Before evaluating the ANOVA's results to assess whether these differences are significant however, we have to take a closer look at the results of Levene's test as shown in Table 6.10. The Levene's test statistic clearly suggests that the population variances are equal, as the test's p-value (**0.404**) is well above 0.05. Thus, following from Table 6.2, we should use the F-test to decide whether at least one group mean differs from the others.

Table 6.9 Descriptive statistics

Descriptives

Overall, I am satisfied with the price performance ratio of Oddjob Airways.

	N	Mean	Std. Deviation	Std. Error	95 % Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Blue	677	4.47	1.641	.063	4.35	4.60	1	7
Silver	245	4.03	1.560	.100	3.84	4.23	1	7
Gold	143	3.99	1.556	.130	3.73	4.24	1	7
Total	1065	4.31	1.625	.050	4.21	4.40	1	7

Table 6.10 Test of homogeneity of variances

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Overall, I am satisfied with the price performance ratio of Oddjob Airways.	Based on Mean	.907	2	1062	.404
	Based on Median	.068	2	1062	.934
	Based on Median and with adjusted df	.068	2	1017.925	.934
	Based on trimmed mean	.771	2	1062	.463

Table 6.11 ANOVA table

ANOVA

Overall, I am satisfied with the price performance ratio of Oddjob Airways.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	51.755	2	25.878	9.963	.000
Within Groups	2758.455	1062	2.597		
Total	2810.210	1064			

6.6.2.2 Calculate the Test Statistic

SPSS will calculate the test statistic and provide a table as shown in Table 6.11. The top part of the table reports several measures of model significance and fit which we will interpret next to make the test decision.

Table 6.12 Post hoc test results (Hochberg's GT2 and Games Howell)

Multiple Comparisons

Dependent Variable: Overall, I am satisfied with the price performance ratio of Oddjob Airways.

	(I) Traveler Status	(J) Traveler Status	Mean Difference (I-J)	Std. Error	Sig.	95 % Confidence Interval	
						Lower Bound	Upper Bound
Hochberg	Blue	Silver	.440*	.120	.001	.15	.73
		Gold	.487*	.148	.003	.13	.84
	Silver	Blue	–.440*	.120	.001	–.73	–.15
		Gold	.047	.170	.990	–.36	.45
	Gold	Blue	–.487*	.148	.003	–.84	–.13
		Silver	–.047	.170	.990	–.45	.36
Games-Howell	Blue	Silver	.440*	.118	.001	.16	.72
		Gold	.487*	.145	.003	.15	.83
	Silver	Blue	–.440*	.118	.001	–.72	–.16
		Gold	.047	.164	.956	–.34	.43
	Gold	Blue	–.487*	.145	.003	–.83	–.15
		Silver	–.047	.164	.956	–.43	.34

* The mean difference is significant at the 0.05 level.

6.6.2.3 Make the Test Decision

Let's now focus on the *F*-test result with respect to the overall model in ◘ Table 6.11. The model has an *F*-value of **9.963**, which yields a *p*-value of 0.00 (less than 0.05), suggesting that at least two of the three groups differ significantly with regard to the overall price/performance satisfaction. SPSS also indicates the between-group sum of squares (i.e., SS_B), which is labelled **Between Groups** in ◘ Table 6.11 and the within-group sum of squares (i.e., SS_W), which is labelled **Within Groups**. The total sum of squares (i.e., SS_T) is the sum of this. We will need these figures to calculate the effect size later.

6.6.2.4 Carry Out Post Hoc Tests

To evaluate whether all groups are mutually different or only two, we take a look at the post hoc test results. The Levene's test showed that the population variances are equal. The descriptive statistics in ◘ Table 6.9 indicate that the group-specific sample sizes are clearly different. Hence, following the procedure in ◘ Fig. 6.6, we should focus on the interpretation of Hochberg's GT2.

The results in ◘ Table 6.12 list a series of comparisons based on Hochberg's GT2 (upper part) and the Games-Howell procedure (lower part). In the first row, you can see the comparison between the *Blue* group and *Silver* group. The difference between the means of these two groups is **0.440** units. Following this row across, we see that this difference is statistically significant (*p*-value = **0.001**). On the contrary, further below, we can see that the difference between the Silver and Gold groups (**0.047**) is not significant (*p*-value = **0.990**).

Even though the population variances are assumed to be equal and group-specific sample sizes differ, let us take a look at the results of the Ryan/Einot-Gabriel/Welsch Q procedure for the sake of completeness. ◘ Table 6.13 organizes the means of the three groups into *homogeneous subsets*. Subsets that do not differ at a significance level of 0.05 are grouped together, and subsets that differ are placed in separate columns. Notice how the *Blue* group shows up in a separate column than the *Gold* and *Silver* groups. This indicates that the *Blue* group is significantly different from the other two in terms of the overall price/performance satisfaction. The *Gold* and *Silver* groups show up in the same column, indicating that they are not significantly different from each other. The lower part of ◘ Table 6.13 uses the same display to illustrate the results of Hochberg'S GT2.

6.6.2.5 Measure the Strength of the Effects

Finally, we want to examine the strength of the effect by computing η^2 and ω^2. However, the One-Way ANOVA analysis in SPSS does not provide these measures.[13] Nevertheless we can easily compute η^2 and ω^2 manually using the information from ◘ Table 6.10. For this we need to look up SS_B, SS_T, SS_W, n and k from the output. Note also that $MS_W = SS_W/(n\text{-}k)$ and that k represents the number of groups to compare while n represents the sample size.

13 Note that when initiating the analysis by going to ► Analyze ► General Linear Model ► Univariate, we can request these statistics under **Options** (**Estimates of effect size**).

Table 6.13 Post hoc test results (Ryan/Einot-Gabriel/Welsch Q procedure)

Overall, I am satisfied with the price performance ratio of Oddjob Airways.

	Traveler Status	N	Subset for alpha = 0.05	
			1	2
Ryan-Einot-Gabriel-Welsch Range	Gold	143	3.99	
	Silver	245	4.03	
	Blue	677		4.47
	Sig.		.807	1.000
Hochberg[a, b]	Gold	143	3.99	
	Silver	245	4.03	
	Blue	677		4.47
	Sig.		.985	1.000

Means for groups in homogeneous subsets are displayed.
[a] Uses Harmonic Mean Sample Size = 239.011.
[b] The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

$$\eta^2 = \frac{SS_B}{SS_T} = \frac{51.755}{2810.210} = 0.0184$$

$$\omega^2 = \frac{SS_B - (k-1)\cdot MS_W}{SS_T + MS_W} = \frac{51.755 - (3-1)\cdot 2.560}{2810.210 + 2.560} = 0.0166$$

The effect size η^2 is 0.0184. This means that differences in the travelers' status explain **1.841 %** of the total variation in the overall price/performance satisfaction. The ω^2 displayed is 0.0166, which is low.

6.6.2.6 Interpret the Results

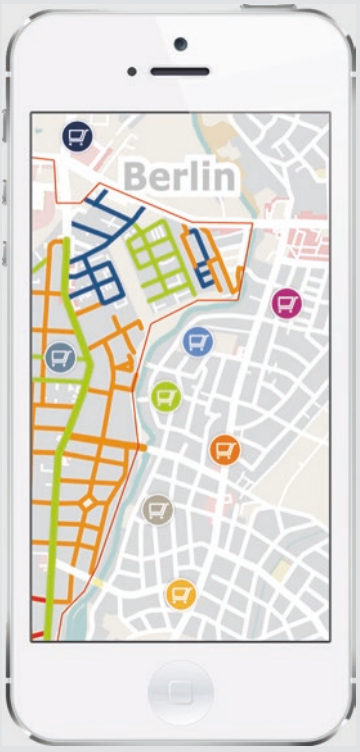
Overall, based on the outputs of the ANOVA in Table 6.11 and the post-hoc output in Tables 6.12, we conclude that:

1. *Gold* and *Blue* members, as well as *Silver* and *Blue* members, differ significantly in their mean overall price/performance satisfaction.
2. There is no difference between *Gold* and *Silver* members in their mean overall price/performance satisfaction.
3. Membership status explains only a minimal share of the customers' price/performance satisfaction. Hence, other factors—presently not included in the model—explain the remaining variation in the outcome variable.

6.7 Customer Spending Analysis with IWD Market Research (Case Study)

Case Study

Founded in 1998, IWD Market Research Institute (https://www.iwd-marketresearch.com/index.php) is a major German full-service provider, specialized in smartphone-supported face-to-face surveys, particularly in the retail sector. One of the company's main fields of expertise are catchment area studies. These studies involve analyzing from which streets or districts potential customers originate, along with their demographics and spending behavior.



In 2017, one of UK's major supermarket chains commissioned a study to better understand their customer base. In the course of this project, IWD surveyed the following items among 7199 customers who frequented the store during the project period (variable names in parentheses):

- Could you please tell me your current postal code? (*postal_code*)
- How often do you shop in this store? (*frequency*)
- How much have you spent today? (*spending*)
- Do you regularly receive our leaflet at home? (*leaflet*)
- How did you get here today? (*journey*)
- May I ask you for your age? (*age*)
- How many people live in your household (including yourself)? (*household*)
- The customer's gender (*gender*)

Use the data provided in *IWD.sav* (⇩ Web Appendix → Downloads) to answer the following research questions:

1. Is there a significant difference in spending between male and female customers?
2. Does the intensity with which customers receive a leaflet variable have a significant effect on their spending? Use an appropriate test to check for significant differences in spending between each of the three categories of the leaflet variable ("Yes, regularly (weekly);" "Yes, irregularly;" "No").
3. Does the customers' spending depend on how they got to the supermarket (*journey*)? Are there significant differences between the journey types?
4. Extend the previous analysis by researching the impact of *leaflet* and *journey* on *spending*. Is there a significant interaction between the *leaflet* and *journey* variables?
5. Based on your analysis results, please provide recommendations for the management team on how to align their future marketing actions.

6.8 Review Questions

1. Describe the steps involved in hypothesis testing in your own words.
2. If you take a look at the following video, you will see many situations that could be tested. Please identify these and discuss what tests are needed.

https://www.youtube.com/watch?v=CJOfmHmwGO0

3. Explain the concept of the *p*-value and explain how it relates to the significance level α.
4. 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.
5. Write two hypotheses for each of the example studies in question 4, including the null hypothesis and alternative hypothesis.
6. Describe the difference between independent and paired samples *t*-tests in your own words and provide two examples of each type.
7. What is the difference between an independent samples *t*-test and an ANOVA?
8. What are post hoc test and why is their application useful in ANOVA?

References

Agresti, A., & Finlay, B. (2014). *Statistical methods for the social sciences* (4th ed.). London: Pearson.
Benjamin, D. J., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10.
Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, *57*(1), 49–64.
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.
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.
Lakens, D., et al. (2018). Justify your alpha. *Nature Human Behaviour*, *2*, 168–171.
Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.) Contributions to probability and statistics (pp. 278–292). Palo Alto, CA: Stanford University Press.
Liao, T. F. (2002). *Statistical group comparison*. New York, NJ: Wiley-InterScience.
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.
Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632.
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, N.J.: John Wiley & Sons.
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.

Further Readings

Kanji, G. K. (2006). *100 statistical tests* (3rd ed.). London: Sage.
Van Belle, G. (2011). *Statistical rules of thumb* (2nd ed.). Hoboken, N.J.: John Wiley & Sons.