

Chapter 6

Hypothesis Testing & ANOVA

Learning Objectives

After reading this chapter you should understand:

- The logic of hypothesis testing.
- The steps involved in hypothesis testing.
- What test statistics are.
- Types of error in hypothesis testing.
- Common types of t-tests, one-way and two-way ANOVA.
- How to interpret SPSS outputs.

Keywords: α -Inflation · Degrees of freedom · F-test · Familywise error rate · Independent and paired samples t-test · Kolmogorov–Smirnov-test · Significance · Levene’s test · Null and alternative hypothesis · One-sample and two-samples t-tests · One-tailed and two-tailed tests · One-way and two-way ANOVA · p-value · Parametric and nonparametric tests · Post hoc tests · Power of a test · Shapiro–Wilk test · Statistical significance · t-test · Type I and type II error · z-test

Does the sponsorship of cultural activities increase the reputation of the sponsoring company? To answer this question, Schwaiger et al. (2010) conducted a large-scale 1-year long experiment which included more than 3,000 consumers. Each month, members of the treatment groups received press reports with a cultural sponsoring theme, while members of the control groups received reports without any reference to culture sponsorship activities. The consumers had to rate to what extent each company supports or sponsors cultural events at the beginning and the end of the experiment. Hypothesis tests indicated that the consumers in the treatment group assessed the companies’ cultural sponsorship activities significantly higher than the year before, while members of the control group did not. Further tests revealed that companies could benefit from such sponsoring activities as these had a significant positive effect on their reputation. Based on these results, companies can strengthen their positioning in the market through the sponsorship of cultural activities.

Introduction

In the previous chapter, we learned about descriptive statistics, such as means and standard deviations, and the insights that can be gained from such measures. Often, we use these measures to compare groups. For example, we might be interested in investigating whether men or women spend more money on the Internet. Assume that the mean amount that a sample of men spends online is 200 USD per year against the mean of 250 USD for the women sample. Two means drawn from different samples are almost always different (in a mathematical sense), but are these differences also statistically significant?

To determine statistical significance, we want to ascertain whether the finding only occurred in the sample analyzed, or whether it also holds for the population (see Chap. 3 for the discussion of samples and population). If the difference between the sample and the population is so large that it is unlikely to have occurred by chance, this is statistical significance. Whether results are statistically significant depends on several factors, including variation in the sample data and the number of observations.

Practitioners, however, usually refer to “significant” in a different, non-statistical context. What they call significant refers to differences that are large enough to influence their decision making process. Perhaps the analysis discloses a significant market segment (i.e., large enough to matter), or the sample reveals such a significant change in consumer behavior that the company needs to change its behavior. Whether results are practically significant (i.e., relevant for decision making), depends on management’s perception of whether the difference is large enough to require specific action.

In this chapter, we will deal with hypothesis testing which allows for the determination of statistical significance. As statistical significance is a precursor to establishing practical significance, hypothesis testing is of fundamental importance for market research.

Understanding Hypothesis Testing

Hypothesis testing is a claim about a statistic characterizing a population (such as a mean or correlation) that can be subjected to statistical testing. Such a claim is called a hypothesis and refers to a situation that might be true or false. A hypothesis may comprise a judgment about the difference between two sample statistics (e.g., the difference in the means of separate corporate reputation assessments made by customers and by non-customers). It can also be a judgment of the difference between a sample statistic and a hypothesized population parameter.

Subject to the type of claim made in the hypothesis, we have to choose an appropriate statistical test, of which there are a great number, with different tests suitable for different research situations. In this chapter, we will focus on

parametric tests used to examine hypotheses that relate to differences in means. Parametric tests assume that the variable of interest follows a specific statistical distribution.

The most popular parametric test for examining means is the *t-test*. The *t-test* can be used to compare one mean with a given value (e.g., do males spend more than 150 USD a year online?). We call this type a *one-sample t-test*. Alternatively, we can test the difference between two samples (e.g., do males spend as much as females?). In this latter case, we talk about a *two-samples t-tests* but then we need to differentiate whether we are analyzing two *independent samples* or two *paired samples*.

Independent samples *t-tests* consider two unrelated groups, such as males vs. females or users vs. non-users. Conversely, paired samples *t-tests* relate to the same set of respondents and thus, occur specifically in pre-test/post-test studies designed to measure a variable before and after a treatment. An example is the before-after design discussed in Chap. 4.

Often, we are interested in examining the differences between means found in more than two groups of respondents. For example, we might be interested in evaluating differences in sales between low, medium, and high income customer segments. Instead of carrying out several paired comparisons through separate *t-tests*, we should use *Analysis of Variance (ANOVA)*. ANOVA is useful for complex research questions, such as when three or more means need to be compared, as ANOVA can analyze multiple differences in one analysis. While there are many types of ANOVA, we focus on the most common types, the one-way and two-way ANOVA.

In the 📖 Web Appendix (→ Chap. 6), we provide an introduction to popular nonparametric tests which do not require the variable of interest to follow a specific distribution and which can also be used for variables measured on a nominal or ordinal scale. There, we specifically discuss the χ^2 -tests (pronounced as *chi-square*).

No type of hypothesis testing can validate hypotheses with absolute certainty. The problem is that we have to rely on sample data to make inferences and, as we have not included the whole population in our analysis, there is some probability that we have reached the wrong conclusion. However, we can (partly) control for this risk by setting an acceptable probability (called significance level) that we will accept a wrong hypothesis.

After this step of hypothesis testing, in which we choose the significance level, information is obtained from the sample and used to calculate the test statistic. Based on the test statistic, we can now decide how likely it is that the claim stated in the hypothesis is correct. Manually, this is done by comparing the test statistic with a critical value that a test statistic must exceed.

SPSS does this work for us by calculating the probability associated with the test statistic, which we only have to compare with the significance level specified earlier. For the sake of clarity, we discuss both manual and computerized ways of making a test decision in the following sections. On the basis of our final decision to either reject or support the hypothesis, we can then draw market research conclusions. Figure 6.1 illustrates the steps involved in hypothesis testing.

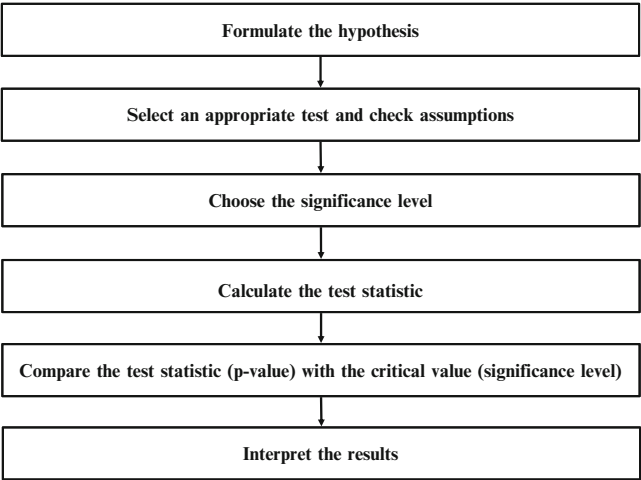


Fig. 6.1 Steps in hypothesis testing

To illustrate the process of hypothesis testing, consider the following example. Suppose that a department store chain wants to evaluate the effectiveness of three different in-store promotion campaigns (1) point of sale display, (2) free tasting stand, and (3) in-store announcements. The management decides to conduct a 1-week field experiment. The management chooses 30 stores randomly and also randomly assigns ten stores to each of the campaign types. This random assignment is important to obtain results that support generalized conclusions, because it equalizes the effect of factors that have not been accounted for in the experimental design (see Chap. 4).

Table 6.1 Sales data

Service type	Sales (units)		
	Point of sale display (stores 1–10)	Free tasting stand (stores 11–20)	In-store announcements (stores 21–30)
Personal	50	55	45
Personal	52	55	50
Personal	43	49	45
Personal	48	57	46
Personal	47	55	42
Self service	45	49	43
Self service	44	48	42
Self service	49	54	45
Self service	51	54	47
Self service	44	44	42

Table 6.1 shows the number of products sold in each store/campaign in 1 week. These sales data are normally distributed, which is a necessary condition for correctly applying parametric tests. We will discuss what this means and how we

can evaluate whether the data are normally distributed later in this chapter. The table also contains information on the service type (personal or self service). We will need this information to illustrate the concept of a two-way ANOVA later in this chapter, so let us ignore this column for now.

In what follows, we want to use these data to carry out different mean comparisons using parametric tests. We illustrate these tests by means of basic formulae. Although we will use SPSS to execute the tests, a basic knowledge of the test formulae will help you understand how the tests work. Ultimately, you will find out that the formulae are not at all as complicated as you might have thought. . . Most of these formulae contain Greek characters with which you might not be familiar. That's why we have included a table with a description of these characters in the [Web Appendix \(→Additional Material\)](#).

Testing Hypotheses About One Mean

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 no effect. Conversely, an alternative hypothesis (represented as H_1) is one in which some difference is expected – it is the opposite of the null hypothesis and supplements it. Examples of potential null and alternative hypotheses on the campaign types are:

- H_0 : There's no difference in the mean sales between stores that installed a point of sale display and those that installed a free tasting stand (statistically, average sales of the point of sale display = average sales of the free tasting stand)
 H_1 : There's a difference in the mean sales between stores that installed a point of sale display and those that installed a free tasting stand (statistically, average sales of the point of sale display \neq average sales of the free tasting stand).
- H_0 : The mean sales in stores that installed a point of sale display are less than (or equal to) 45 units (Note that in this hypothesis, H_0 postulates a specific direction of the effect).
 H_1 : The mean sales in stores that installed a point of sale display are more than 45 units.

When carrying out a statistical test, we always test the null hypothesis. This test can have two outcomes:

The first outcome may be that we reject the null hypothesis, thus finding support for the alternative hypothesis. This outcome is, of course, desirable in most analyses, as we generally want to show that something (such as a promotion campaign) has an effect on a certain outcome (e.g., sales). This is why we generally frame the effect that we want to show in the alternative hypothesis.

The second outcome may be that we do not reject the null hypothesis. However, it would be incorrect to conclude then that the null hypothesis is true, as it is not possible to show 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. Also, it is impossible statistically to show that there is no Yeti or that the Loch Ness Monster Nessie does not exist. Conversely, discovering one Yeti will demonstrate its existence. Thus, we can only reject the null hypothesis.

To make the point again: you can never prove that a hypothesis is true! Each hypothesis test inevitably has a certain degree of uncertainty so that even if we reject a null hypothesis, we can never be fully certain that this was the correct decision. Therefore, knowledgeable market researchers use terms such as “reject the null hypothesis,” or “find support for the alternative hypothesis” when they discuss the findings. Terms like “prove the hypothesis” should never be part of any discussion on statistically tested hypotheses.

Returning to our example, department store managers might be interested in evaluating whether the point of sale display has a positive effect on product sales. Management may consider the campaign a success if sales are higher than the 45 units normally sold (you can likewise choose any other value – the idea is to test the sample against a given standard). But what does “generally higher” mean? Just looking at the data clearly suggests that sales are higher than 45. However, we are not interested in the sample result as much as to assess whether this result is likely to occur in the whole population (i.e., given that the sample is representative, does this result also hold for other stores?).

The appropriate way to formulate the hypotheses is:

$$H_0 : \mu \leq 45$$

(in words, the population mean – indicated by μ – is equal to or smaller than 45)

$$H_1 : \mu > 45$$

(in words, the population mean is larger than 45)

It is important to note that the hypothesis always refers to a population parameter, in this case μ (pronounced as *mu*). It is convention that Greek characters always represent population parameters and not a sample statistic (e.g., the sample mean indicated by \bar{x}). In the end, we want to make inferences regarding the population, not the sample, given the basic assumption that the sample is representative!

If the null hypothesis H_0 is rejected, then the alternative hypothesis H_1 will be accepted and the promotion campaign can be considered a success. On the other hand, if H_0 is not rejected, we conclude that the campaign did not have a positive impact on product sales in the population. Of course, with *conclude*, we still mean that this is what the test indicates and not that this is an absolute certainty.

In this example, we consider a *one-tailed test* (specifically, a right-tailed test) as the two hypotheses are expressed in one direction relative to the standard of 45

units: we presume that the campaign has a positive effect on product sales, which is expressed in hypothesis H_1 . More precisely, we consider a right-tailed test because H_1 presumes that the population mean is actually higher than a given standard. On the other hand, suppose that we were interested in determining whether the product sales differ from the 45 units, either being higher or lower. In this case, a *two-tailed test* would be required, leading to the following hypotheses:

$$H_0 : \mu = 45$$

$$H_1 : \mu \neq 45$$

The difference between the two general types of tests is that a one-tailed test looks for an increase or, alternatively, a decrease in the parameter relative to the standard, whereas a two-tailed test looks for any difference in the parameter from the selected standard. Usually, two-tailed tests are applied because the notion of looking for any difference is more reasonable. Furthermore, two-tailed tests are stricter (and generally considered more appropriate) when it comes to revealing significant effects. Nevertheless, we will stick to our initial set of hypotheses for now.

Select an Appropriate Test and Check Assumptions

To choose an appropriate statistical test, we have to consider the purpose of the hypothesis test as well as the conditions under which the test is applied. This choice depends on the sample size and whether the data are normally distributed.

In this section, we focus on parametric tests to test differences in means. That is, we may want to compare a single sample's mean against some given standard, or evaluate whether the means of two independent or paired samples differ significantly in the population. The *parametric tests* we discuss in this chapter assume that the test variable – which needs to be measured on at least an interval scale – is normally distributed. In our numerical example of the promotion campaign, we already know that the sales are normally distributed. However, in real-world applications, we usually lack this information. To test whether the data are normally distributed, we can use two tests, the *Kolmogorov–Smirnov test (with Lilliefors correction)*, and the *Shapiro–Wilk test*. We describe these in more detail in Box 6.1 and provide an example, executed in SPSS, toward the end of this chapter. If the Kolmogorov–Smirnov or Shapiro–Wilk test suggests that the test variable is not normally distributed, you should use nonparametric tests, which do not make any distributional assumption (🔗 Web Appendix → Chap. 6).

The good news is that parametric tests we discuss in this chapter are generally quite robust against a violation of the normality assumption. That is, if the test statistic obtained is much below the critical value, deviations from normality matter little. If the test statistic is close to the critical value, for example the obtained

p-value (introduced later) is 0.04, and n is smaller than 30, deviations from normality could matter. Still, such deviations are usually no more than a p-value of 0.01 units apart, even if severe deviations from normality are present (Boneau 1960).

Thus, even if the Kolmogorov–Smirnov or Shapiro–Wilk test suggests that the data are not normally distributed, we don’t have to be concerned that the parametric test results are grossly wrong. Specifically, where we compare means, we can simply apply one of the parametric tests with a low risk of error, provided we have sample sizes greater than 30.

Box 6.1 Normality tests

An important (nonparametric) test is the one-sample Kolmogorov–Smirnov test. We can use it to test the null hypothesis that a certain variable follows a specific distribution. In most instances, we use it to examine whether a variable is normally distributed. However, in theory, we can use this test to assess any other type of distribution, such as exponential or uniform.

Somewhat surprisingly, the test’s null hypothesis is that the variable follows a specific distribution (e.g., the normal distribution). This means that only if the test result is insignificant, i.e. the null hypothesis is not rejected, can we assume that the data are drawn from the specific distribution against which it is tested. Technically, when assuming a normal distribution, the Kolmogorov–Smirnov test compares the sample scores with an artificial set of normally distributed scores that has the same mean and standard deviation as the sample data. However, this approach is known to yield biased results which can be corrected for by adopting the Lilliefors correction (Lilliefors 1967). The Lilliefors correction considers the fact that we do not know the true mean and standard deviation of the population. An issue with the Kolmogorov–Smirnov test is that it is very sensitive when used on very large samples and often rejects the null hypothesis if very small deviations are present.

The Shapiro–Wilk test also tests the null hypothesis that the test variable under consideration is normally distributed. Thus, rejecting the Shapiro–Wilk test provides evidence that the variable is not normally distributed. It is best used for sample sizes of less than 50. A drawback of the Shapiro–Wilk test however, is that it works poorly if the variable you are testing has many identical values, in which case you should use the Kolmogorov–Smirnov test with Lilliefors correction.

To conduct the Kolmogorov–Smirnov test with Lilliefors correction and the Shapiro–Wilk test in SPSS, we have to go to ► Analyze ► Descriptive Statistics ► Explore ► Plots and choose the **Normality plots with tests** option (note that the menu option ► Analyze ► Nonparametric Tests ► Legacy Dialogs ► 1-Sample K-S will yield the standard Kolmogorov–Smirnov test whose results oftentimes diverge heavily from its counterpart with Lilliefors correction).

We will discuss these tests using SPSS in this chapter’s case study.

When conducting a one-sample (or independent samples) t-test, we need to have independent observations (see Chap. 3 on independence of observations). Note that this does not apply when using the paired samples t-test and its nonparametric


counterparts (i.e. the *McNemar test* and the *Wilcoxon signed rank test*, see  Web Appendix → Chap. 6), in which we are specifically interested regarding evaluating the mean differences between the same respondents’ assessments at two different points in time!



Figure 6.2 provides a guideline for choosing the appropriate parametric test for comparing means, including the associated SPSS menu options. We will discuss all of these in the remainder of the chapter.

In the present example, we know that the test variable is normally distributed and we are interested in comparing a sample mean with a given standard (e.g., 45). Thus, the one-sample t-test is appropriate.

Choose the Significance Level

As discussed previously, each statistical test is associated with some degree of uncertainty, as we want to make inferences about populations by using sample data. In statistical testing, two types of errors can occur: (1) a true null hypothesis can be incorrectly rejected (*type I or α error*), and (2) a false null hypothesis is not rejected (*type II or β error*). These two types of errors are defined in Table 6.2.

Table 6.2 Type I and type II errors

Decision based on sample data	True state of H_0	
	H_0 true	H_0 false
H_0 rejected	Type I error 	Correct decision
H_0 not rejected	Correct decision	Type II error 

In our example, a type I error would occur if using hypothesis testing we find that the point of sale display increased sales, when in fact it did not affect, or may even have decreased sales. Accordingly, a type II error would occur if we did not reject the null hypothesis, which would suggest that the campaign was not successful, even though, in reality, it did significantly increase sales.

A problem with hypothesis testing is that we don’t know the true state of H_0 . At first sight, there doesn’t seem to be an easy way around this problem; however, we can establish a level of confidence that a true null hypothesis will not be erroneously rejected when carrying out a test that is tolerable given the circumstances that apply.

In other words, before carrying out a test, we should decide on the maximum probability that a type I error can occur that we want to allow. This probability is represented by the Greek character α (pronounced as *alpha*) 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 risk of 5% of mistakenly rejecting a true null hypothesis and that the actual risk, based on the data analysis, turned out to be lower than this.

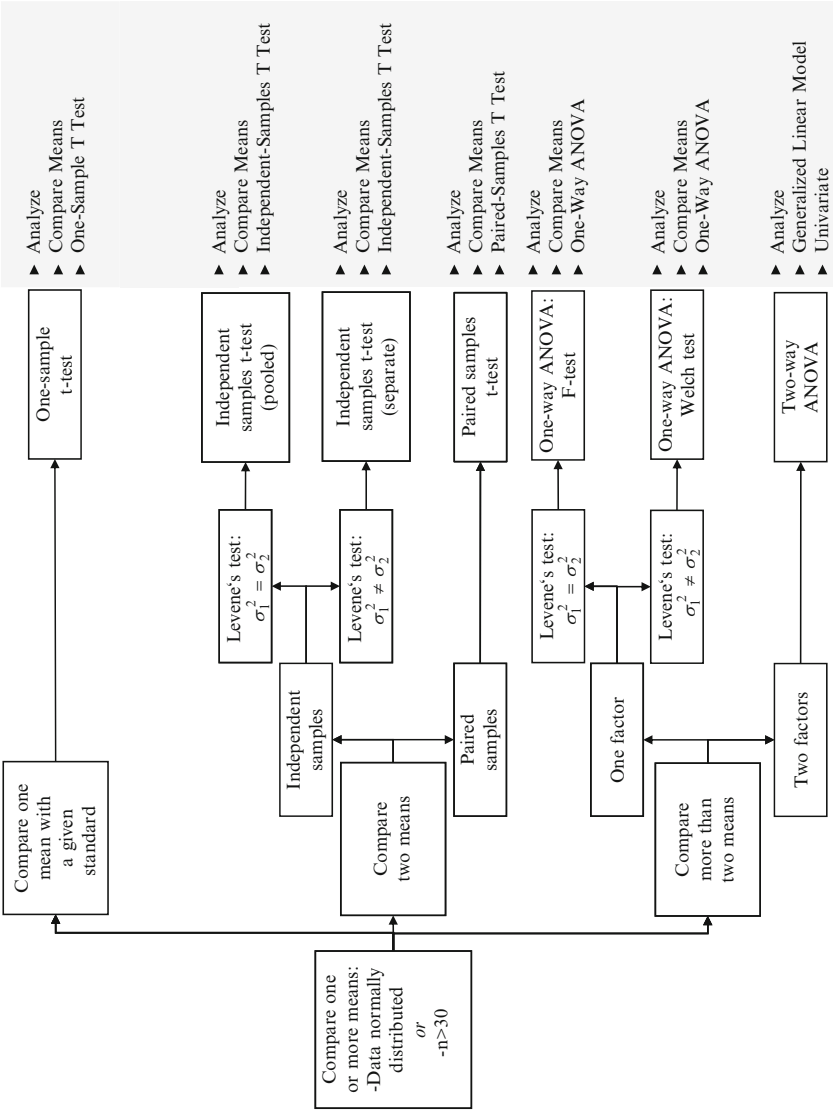


Fig. 6.2 Parametric tests for comparing means

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

Why don't we just set α to 0.0001% to really minimize the probability of a type I error? Obviously, setting α to such a low level would make the erroneous rejection of H_0 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 one type of error, the higher the risk of the other! However, since a type I error is considered more severe than a type II error, we directly control the former by setting α to a desired level.

Another important concept related to this is the *power of a statistical test* (defined by $1 - \beta$, where β is the probability of a type II error), which represents the probability of rejecting a null hypothesis when it is in fact false (i.e. not making a type II error). Obviously, you would want the power of a test to be as high as possible. However, as indicated before, when α is small, the occurrence of a type I error is much reduced, but then β , the probability of a type II error, is large.¹ This is why α is usually set to the levels described above, which ensures that β is not overly inflated. The good news is that both α and β , the probability of incorrect findings, can be controlled for by increasing the sample size: for a given level of α , increasing the sample size will decrease β . See Box 6.2 for further information on statistical power.

Box 6.2 Statistical power

A common problem that market researchers encounter is calculating the sample size required to yield a certain test power, given a predetermined level of α . Unfortunately, computing the required sample size (*power analyses*) can become very complicated, depending on the test or procedure we use. However, SPSS provides an add-on module called "Sample Power," which can be used to carry

(continued)

¹Note that the power of a statistical test may depend on a number of factors which may be particular to a specific testing situation. However, power nearly always depends on (1) the chosen significance level, and (2) the magnitude of the effect of interest in the population.

out such analyses. In addition, the Internet offers a wide selection of downloadable applications and interactive Web programs for this purpose. Just google “power analysis calculator” and you will find numerous helpful sites. In most sites you can enter the α and sample size and calculate the resulting level of power. By convention, 0.80 is an acceptable level of power.

Nevertheless, power analyses are beyond the scope of basic market research training and you should make use of certain rules of thumb. As a rough guideline, Cohen (1992) provides a convenient presentation of required sample sizes for different types of tests. For example, in order to detect the presence of large differences between two independent sample means for $\alpha = 0.05$ and a power of $\beta = 0.80$, requires $n = 26$ (with $n = 64$ for medium differences and $n = 393$ for small differences) observations in each group.

Calculate the Test Statistic

After having decided on the appropriate test as well as the significance level, we can proceed by calculating the test statistic using the sample data at hand. In our example we make use of a *one-sample t-test*, whose test statistic is simply computed as follows:

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

Here \bar{x} and μ are the sample and assumed population means, respectively, and $s_{\bar{x}}$ is the standard error.

Let’s first take a look at the formula’s numerator, which describes the difference between the sample mean \bar{x} and the hypothesized population mean μ . If the point of sale display was highly successful, we would expect \bar{x} to be much higher than μ , leading to a positive difference between the two in the formula’s numerator.

Comparing the calculated sample mean (47.30) with the hypothesized mean (45), we find that they are mathematically different:

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

At first sight, therefore, it appears as if the campaign was successful; sales during the time of the campaign were higher than those that the store normally encounters.

However, we have not yet accounted for the variation in the dataset, which is taken care of by $s_{\bar{x}}$. The problem is that if we had taken another sample from the population, for example, by using data from a different period, the new sample

mean would most likely have been different from that which we first encountered. To correct this problem, we have to divide the mean difference $\bar{x} - \mu$ 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 this mean? The sample mean is usually used as an estimator of the population mean; that is, we assume that the sample and population means are identical. However, when we draw different samples from the same population, we will most likely obtain different means. The standard error tells us how much variability there is in the mean across different samples from the same population. Therefore, a large value for the standard error indicates that a specific sample mean may not adequately reflect the population mean.

The standard error is computed by dividing the sample standard deviation (s), by the square root of the number of observations (n):

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

Why do we have to divide the difference $\bar{x} - \mu$ by the standard error $s_{\bar{x}}$? Proceeding from the calculation of the test statistic above, we see that when the standard error is very low (i.e., 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 values favor the rejection of the null hypothesis, which implies that the alternative hypothesis is supported. The lower the standard error $s_{\bar{x}}$, the greater the probability that the population represented by the sample truly differs from the selected standard in terms of the average number of units sold.

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. Consequently, for a certain standard deviation, a higher n goes hand in hand with lower $s_{\bar{x}}$ which is why we generally want to have high sample sizes (within certain limits, see Box 6.3).

Box 6.3 Can a sample size be too large?

In a certain sense, even posing this question is heresy! Market researchers are usually trained to think that large sample sizes are good because of their positive relationship with the goodness of inferences.

However, it doesn't seem logical that the claim for high sample sizes should never be discussed. As long as our primary interest lies in the precise estimation of an effect (which will play a greater role when we discuss regression analysis), then the larger the sample, the better. If we want to come as close as possible to a true parameter, increasing the sample size will generally provide more precise results.

(continued)

However, there are two issues. First, all else being equal, if you increase the sample size excessively, even slight or marginal effects become statistically significant. However, statistically significant does not mean managerially relevant! In the end, we might run into problems when wanting to determine which effects really matter and how priorities should be assigned.

Second, academics especially are inclined to disregard the relationship between the (often considerable) costs associated with obtaining additional observations and the additional information provided by these observations. In fact, the incremental value provided by each additional observation decreases with increasing sample size. Consequently, there is little reason to work with an extremely high number of observations as, within reason, these do not provide much more information than fewer observations.

In addition, not every study's purpose is the precise estimation of an effect. Instead, many studies are exploratory in nature, and we try to map out the main relationships in some area. These studies serve to guide us in directions that we might pursue in further, more refined studies. Specifically, we often want to find those statistical associations that are relatively strong and that hold promise of a considerable relationship that is worth researching in greater detail. We do not want to waste time and effort on negligible effects. For these exploratory purposes, we generally don't need very large sample sizes.

In summary, with a higher mean difference $\bar{x} - \mu$ and lower levels of uncertainty in the data (i.e. lower values in $s_{\bar{x}}$), there is a greater likelihood that we can assume that the promotion campaign had a positive effect on sales.

We discuss this test statistic in greater detail because, in essence, all tests follow the same scheme. First, we compare a sample statistic, such as the sample mean, with some standard or the mean value of a different sample. Second, we divide this difference by another measure (i.e., the standard deviation or standard error), which captures the degree of uncertainty in the sample data.

In summary, the test statistic is nothing but the ratio of the variation, which is due to a real effect (expressed in the numerator), and the variation caused by different factors that are not accounted for in our analysis (expressed in the denominator). In this context, we also use the terms systematic and unsystematic variation.

If you understood this basic principle, you will have no problems understanding most other test statistics that are used.


Let's go back to the example and compute the standard error as follows

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


Thus, the result of the test statistic is

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

Note that in this example, we used sample data to calculate the standard error $s_{\bar{x}}$. Even though this occurs rather rarely (actually, almost never) in market research studies, textbooks routinely discuss situations in which we know the population's standard deviation beforehand. From a statistical standpoint, we would have to apply a different test, called the *z-test*, in this situation where the standard deviation in the population is known. The *z-test* follows a different statistical distribution (in this case, a normal instead of a *t*-distribution), which we have to consider when determining the critical value associated with a test statistic (we do this the following step of hypothesis testing). Likewise, the *z-test* is typically used in situations when the sample size exceeds 30, because the *t*-distribution and normal distribution are almost identical in higher sample sizes.

As the *t*-test is slightly more accurate, SPSS only considers the *t*-test in its menu options. To avoid causing any confusion, we do not present the formulae associated with the *z-test* here, but we have included these in the  Web Appendix (→ Chap. 6).

Compare the Test Statistic (*p*-value) with the Critical Value (Significance Level)


In the next step, we have to determine the critical value, which the test statistic must exceed in order for the null hypothesis to be rejected. When calculating test results manually, this is done using statistical tables. In our case, we have to make use of the *t*-distribution table (see Table A1 in the  Web Appendix → Additional Material).

An important characteristic of the *t*-distribution is that – unlike the normal distribution – it also explicitly considers the amount of information (called *degrees of freedom*, usually abbreviated as “df”) available to estimate the test statistic's parameters. In general terms, an estimate's degrees of freedom (such as a variable's variance) are equal to the amount of independent information used (i.e., the number of observations), minus the number of parameters estimated as intermediate steps in the estimation of the parameter itself.

The concept of degrees of freedom is very abstract, so let's look at a simple example: Suppose you have a single sample with n observations to estimate a variable's variance. Then, the degrees of freedom are equal to the number of observations (n) minus the number of parameters estimated as intermediate steps (in this case, one as we need to compute the mean, \bar{x}) and are therefore equal to $n-1$. Field (2009) provides a vivid explanation which we adapted and present in Box 6.4.

Box 6.4 Degrees of freedom

Suppose you have a soccer team and 11 slots in the playing field. When the first player arrives, you have the choice of 11 positions in which you can place a player. By allocating the player to a position (e.g., right defense) one position is occupied. When the next player arrives, you can chose from ten positions. With every additional player that arrives, you will have fewer choices where to position each player. For the very last player, you have no freedom to choose where to put that player – there is only one spot left. Thus, there are ten degrees of freedom. For ten players you have some degree of choice, but for one player you don't. The degrees of freedom are the number of players minus one.

The degrees of freedom for the t-statistic to test a hypothesis on one mean are also $n - 1$; that is, $10 - 1 = 9$ in our example. With 9 degrees of freedom and using a significance level of, for example, 5% (i.e. $\alpha = 0.05$), the critical value of the t-statistic is 1.833 (just look at the “significance level = 0.05” column and at line “df = 9” in Table A1 in the  Web Appendix (→ Additional Material)).

This means that for the probability of a type I error (i.e. falsely rejecting H_0) to be less than or equal to 0.05, the value of the test statistic must be 1.833 or greater. In our case (right-tailed test), the test statistic (2.274) clearly exceeds the critical value (1.833), which suggests that we should reject the null hypothesis.²

Table 6.3 shows 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 of a significance level α . We always consider absolute test values as these may well be negative (depending on the test's formulation), while the tabulated critical values are always positive.

Important: Unlike with right and left-tailed tests, we have to divide α by two when obtaining the critical value of a two-tailed test. This means that when using a significance level of 5%, we have to look under the 0.025 column in the t-distribution table. The reason is that a two-tailed test is somewhat “stricter” than a one-tailed test, because we do not only test whether a certain statistic differs in one direction

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	$\mu \leq \text{value}$	$\mu > \text{value}$	$ t_{test} > t_{critical}(\alpha)$
Left-tailed test	$\mu \geq \text{value}$	$\mu < \text{value}$	$ t_{test} > t_{critical}(\alpha)$
Two-tailed test	$\mu = \text{value}$	$\mu \neq \text{value}$	$ t_{test} > t_{critical}(\frac{\alpha}{2})$

²To 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*0.05,9).” Note that we have to specify “2*0.05” (or, directly 0.1) under α as we are applying a one-tailed instead of a two-tailed test.

from a population value, but rather test whether it is either lower *or* higher than a population value. Thus, we have to assume a tighter confidence level of $\alpha/2$ when looking up the critical value.

You might remember the above things from an introductory statistics course and 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 at least as extreme as the one that is actually observed. This probability is also referred to as the *p-value* (rather confusingly, it is usually denoted with **Sig.** – also in the SPSS outputs).

With regard to our example, the p-value (or probability value) is the answer to the following question: If the population mean is really lower than or equal to 45 (i.e., in reality, therefore, H_0 holds), what is the probability that random sampling would lead to a difference between the sample mean \bar{x} and the population mean μ as large (or larger) as the difference observed? In other words, the p-value is the probability of erroneously rejecting a true null hypothesis.

This description is very similar to the significance level α , which describes the tolerable level of risk 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.³ Thus, we cannot simply transfer the interpretation that a long series of α -level tests will reject no more than $100 \cdot \alpha$ of true null hypotheses. Note that the p-value is not the probability of the null hypothesis being true! Rather, we should interpret it as evidence against the null hypothesis. The α -level is an arbitrary and subjective value that the researcher assigns for the level of risk of making a Type I error; the p-value is calculated from the available data.

Comparison of these two values allows the researcher to reject or not reject the null hypothesis. Specifically, if the p-value is smaller than or equal to the significance level, we reject the null hypothesis. Thus, when looking at test results in SPSS, we have to make use of the following decision rule – this should become second nature!

- p-value (**Sig.** in SPSS) $\leq \alpha \rightarrow$ reject H_0
- p-value (**Sig.** in SPSS) $> \alpha \rightarrow$ do not reject H_0

No matter how small the p-value is in relation to α , these decision rules only guarantee that you will erroneously reject the null hypothesis $100 \cdot \alpha$ times out of 100 repeated tests. Otherwise stated, it is only of interest whether $p\text{-value} \leq \alpha$ but not the specific value of p itself. Thus, a statement such as “this result is highly significant” is inappropriate, as the test results are binary in nature (i.e., they reject or don’t reject, nothing more). Compare it with any binary condition from real life, such as pregnancies – it is not possible to be just a little bit pregnant!

³Unfortunately, there is quite some confusion about the difference between α and p-value. See Hubbard and Bayarri (2003) for a discussion.

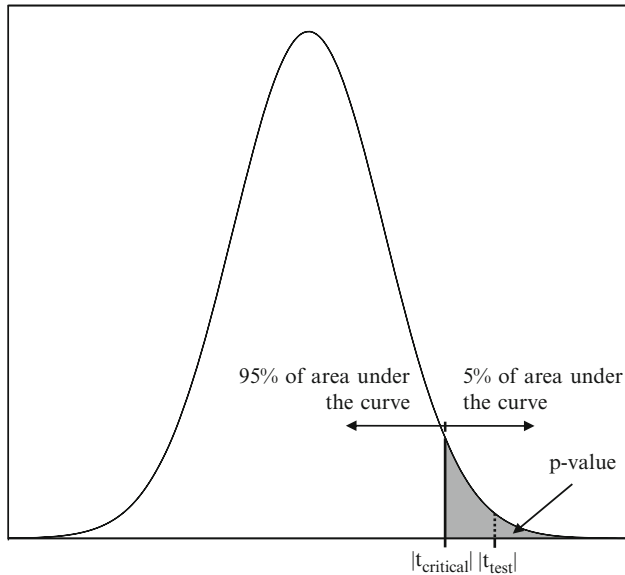


Fig. 6.3 Relationship between test value, critical value, and p-value

In our example, the actual p-value is about 0.024, which is clearly below the significance level of 0.05. Thus, we can reject the null hypothesis and find support for the alternative hypothesis.⁴

Figure 6.3 summarizes these concepts graphically. In this figure, you can see that the critical value t_{critical} for an α -level of 5% divides the area under the curve into two parts. One part comprises 95% of the area, whereas the remaining part (also called the acceptance region, meaning that we accept the alternative hypothesis) represents the significance level α , that is, the remaining 5%. The test value $|t_{\text{test}}|$ defines the actual probability of erroneously rejecting a true null hypothesis, which is indicated by the area right of the dotted line. This area is nothing but the p-value, which is smaller than the 5%. Thus, we can reject the null hypothesis.

⁴We don't have to conduct manual calculations and tables when working with SPSS. However, we can easily compute the p-value ourselves 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 ($\text{tails} = 1$) or two-tailed test ($\text{tails} = 2$). For our example, just open a new spreadsheet and type in "=TDIST(2.274,9,1)". Likewise, there are several webpages with Java-based modules (e.g., <http://www.graphpad.com/quickcalcs/index.cfm>) that calculate p-values and test statistic values.

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 significantly increased sales during the week it was installed.

Comparing Two Means: Two-samples t-test

In the previous example, we examined a hypothesis relating to one sample and one mean. However, market researchers are often interested in comparing two samples regarding their means. As indicated in Fig. 6.2, two samples can either be *independent* or *paired*, depending on whether we compare two distinct groups (e.g., males vs. females) or the same group at different points in time (e.g., customers before and after being exposed to a treatment or campaign). Let's begin with two independent samples.

Two Independent Samples

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

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

Each of the foregoing hypotheses aims 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 existing example of a promotion campaign in a department store. Specifically, we want to test whether the population mean of the sales of the point of sale display (μ_1) differs from that of the free tasting stand (μ_2). Thus, the resulting null and alternative hypotheses are now:

$$H_0 : \mu_1 = \mu_2 \text{ (or put differently : } H_0 : \mu_1 - \mu_2 = 0 \text{)}$$

$$H_1 : \mu_1 \neq \mu_2 \text{ (or } H_1 : \mu_1 - \mu_2 \neq 0 \text{)}$$

In its general form, the test statistic of the two independent samples t-test – which is now distributed with $n_1 + n_2 - 2$ degrees of freedom – seems very similar to the one-sample t-test:

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

Here, \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 likewise use another value in cases where we hypothesize a specific difference in population means. Lastly, $s_{\bar{x}_1 - \bar{x}_2}$ describes the estimated standard error, which comes in two forms:

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

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

2. Alternatively, if we assume that the population variances differ (i.e. $\sigma_1^2 \neq \sigma_2^2$), things become a little bit easier as we can use a *separate* variance estimate:


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

How do we determine whether the two populations have the same variance? This is done by means of an intermediate step that consists of another statistical test. Not surprisingly, this test, known as the F-test of sample variance (also called Levene's test), considers 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 is that they differ.

As the computation of this test statistic is rather complicated, we refrain from discussing it in detail. If you want to learn more about Levene's test and its application to the promotion campaign example, read up on it in the  Web Appendix (→ Chap. 6).

In this example, the Levene's test provides support for the assumption that the variances in the population are equal, so that we have to make use of the pooled variance estimate. First, we estimate the variances of samples 1 and 2

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{10} (x_{1i} - \bar{x}_1)^2 = \frac{1}{10 - 1} [(50 - 47.30)^2 + \dots + (44 - 47.30)^2] \\ = 10.233,$$

$$s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{10} (x_{2i} - \bar{x}_2)^2 = \frac{1}{10 - 1} [(55 - 52)^2 + \dots + (44 - 52)^2] = 17.556,$$


and use these to obtain the estimated standard error:

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

Inserting this into the test statistic results in:

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

As you can see, calculating these measures manually is not very difficult. Still, it is much easier to let SPSS do the calculations.

The test statistic follows a t-distribution with $n_1 + n_2 - 2$ 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 (\rightarrow Additional Material), we can see that the critical value for a significance level of 5% is 2.101 (note that we are conducting a two-tailed test and, thus, have to look in the $\alpha = 0.025$ column). As the absolute value of -2.819 is greater than 2.101, we can reject the null hypothesis at a significance level of 5% and conclude that the means of the sales of the point of sale display (μ_1) and those of the free tasting stand (μ_2) differ in the population.

If we evaluated the results of the left-tailed test (i.e., $H_0 : \mu_1 \geq \mu_2$ and $H_1 : \mu_1 < \mu_2$) we find that sales of the point of sale display are significantly lower than those of the free tasting stand. Thus, the managerial recommendation would be to make use of free tasting stands when carrying out promotion campaigns, as this increases sales significantly over those of the point of sale display. Of course, in making the final decision, we would need to weigh the costs of the display and free tasting stand against the expected increase in sales.⁵

⁵There may be a situation in which we know the population standard deviation beforehand, for example, from a previous study. From a strict statistical viewpoint, it would be appropriate to use a z-test in this case, but both tests yield results that only differ marginally.

Two Paired Samples

In the previous example we compared the mean sales of two independent samples. Now, imagine that management wants to evaluate the effectiveness of the point of sale display in more detail. We have sales data for the week before point of sale display was installed, as well as the following week when this was not the case. Table 6.4 shows the sale figures of the ten stores under consideration for the two experimental conditions (the point of sale display and no point of sale display). Again, you can assume that the data are normally distributed.

Table 6.4 Sales data (extended)

Store	Sales (units)	
	No point of sale display	Point of sale display
1	46	50
2	51	52
3	40	43
4	48	48
5	46	47
6	45	45
7	42	44
8	51	49
9	49	51
10	43	44

At first sight, it appears that the point of sale display yielded higher sales numbers. The mean of the sales in the week during which the point of sale display was installed (47.30) is slightly higher than in the week when it was not (46.10). However, the question is whether this difference is statistically significant.

Obviously, we cannot assume that we are comparing two independent samples, as each set of two samples originates from the same store, but at different points in time under different conditions. This means we have to examine the differences by means of a *paired samples t-test*. In this example, we want to test whether the sales of the point of sale display condition are significantly higher than when no display was installed. We can express this by means of the following hypotheses, where μ_d describes the population difference in sales:

$$H_0 : \mu_d \leq 0$$

$$H_1 : \mu_d > 0$$

Obviously, we assume that the population difference is greater than zero since we suspect that the point of sale display ought to have significantly increased sales. This is expressed in the alternative hypothesis H_1 , while the null hypothesis assumes that the point of sale display made no difference or even resulted in fewer sales

To carry out this test, we have to define a new variable d_i , which captures the differences in sales between the two treatment conditions (point of sale display installed and not installed) for each of the stores. Thus:

$$d_1 = 50 - 46 = 4$$

$$d_2 = 52 - 51 = 1$$

...

$$d_9 = 51 - 49 = 2$$

$$d_{10} = 44 - 43 = 1$$

Based on these results, we calculate the mean difference


$$\bar{d} = \frac{1}{n} \sum_{i=1}^{10} d_i = \frac{1}{10} (4 + 1 + \dots + 2 + 1) = 1.2$$

as well as the standard error of this difference

$$\begin{aligned} s_{\bar{d}} &= \frac{\sqrt{\frac{1}{n-1} \sum_{i=1}^{10} (d_i - \bar{d})^2}}{\sqrt{n}} \\ &= \frac{\sqrt{\frac{1}{9} [(4 - 1.2)^2 + (1 - 1.2)^2 + \dots + (2 - 1.2)^2 + (1 - 1.2)^2]}}{\sqrt{10}} = 0.533 \end{aligned}$$

As you might suspect, the test statistic is very similar to the ones discussed before. Specifically, 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

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}} = \frac{1.2 - 0}{0.533} = 2.250,$$

which follows a t-distribution with $n - 1$ degrees of freedom, where n is the number of pairs that we compare. Assuming a significance level of 5%, we obtain the critical value by looking at Table A1 in the  Web Appendix (→Additional Material). In our example, with 9 degrees of freedom and using a significance level of 5% (i.e. $\alpha = 0.05$), the critical value of the t-statistic is 1.833. Since the test value is larger than the critical value, we can reject the null hypothesis and presume that the point of sale display really did increase sales.

Comparing More Than Two Means: Analysis of Variance ANOVA

Researchers are often interested in examining mean differences between more than two groups. Some illustrative research questions might be:

- Do light, medium and heavy users differ with regard to their monthly disposable income?
- Do customers across four different types of demographic segments differ with regard to their attitude towards a certain brand?
- Is there a significant difference in hours spent on Facebook between US, UK and German 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 groups under consideration. In this example, doing so would require the following comparisons (1) the point of sale display vs. the free tasting stand, (2) the point of sale display vs. the in-store announcements, and (3) the free tasting stand vs. the in-store announcements. While three comparisons seem to be easily manageable, you can appreciate the difficulty that will arise when a greater number of groups are compared. For example, with ten groups, we would have to carry out 45 group comparisons.⁶

Although such high numbers of comparisons become increasingly time consuming, there is a more severe problem associated with this approach, called α -inflation. This refers to the fact that the more tests you conduct at a certain significance level, the more likely you are to claim a significant result when this is not so (i.e., a 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 found by

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


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 provide results that are wrong and which are right.

Instead of carrying out many pairwise tests and creating α -inflation, market researchers use ANOVA, which allows a comparison of averages between three

⁶The number of pairwise comparisons is calculated as follows: $k(k - 1)/2$, with k the number of groups to compare.

or more groups. In ANOVA, the variable that differentiates the groups is referred to as the *factor* (don't confuse this with the factors from factor analysis which we will discuss in Chap. 8!). The values of a factor (i.e., as found for the different groups under consideration) are also referred to as *factor levels*.

In the example above on promotion campaigns, we considered only one factor with three levels, indicating the type of campaign. This is the simplest form of an ANOVA and is called *one-way ANOVA*. However, ANOVA allows us to consider more than one factor. For example, we might be interested in adding another grouping variable (e.g., the type of service offered), thus increasing the number of treatment conditions in our experiment. In this case, we would use a *two-way ANOVA* to analyze the factors' effect on the units sold. ANOVA is in fact even more flexible in that you can also integrate metric independent variables and even several additional dependent variables.

For the sake of clarity, we will focus on the more fundamental form of ANOVA, the one-way ANOVA, with only a brief discussion of the two-way ANOVA.⁷ For a more detailed discussion of the latter, you can turn to the  Web Appendix (→ Chap. 6).

Understanding One-Way ANOVA

As indicated above, ANOVA is used to examine mean differences between more than two groups.⁸ In more formal terms, the objective of one-way ANOVA is to test the null hypothesis that the population means of the groups under consideration (defined by the factor and its levels) are equal. If we compare three groups, as in our 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 mean sales). The alternative hypothesis is

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

which implies that at least two population means differ significantly. Of course, before we even think of running an ANOVA in SPSS, we have to come up with a problem formulation, which requires us to identify the dependent variable and the

⁷Field (2009) provides a detailed introduction to further ANOVA types such as multiple ANOVA (MANOVA) or an analysis of covariance (ANCOVA).

⁸Note that you can also apply ANOVA when comparing two groups. However, in this case, you should rather revert to the two independent samples t-test.

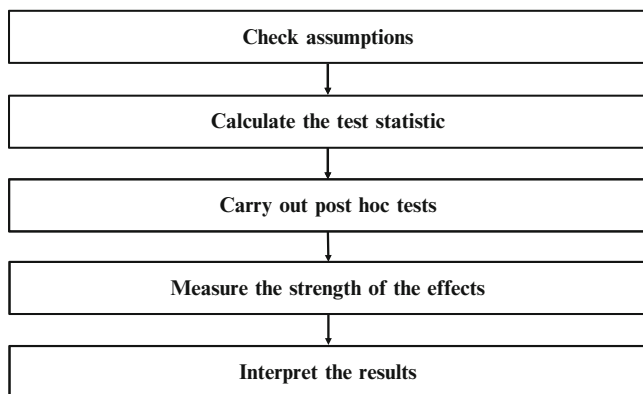


Fig. 6.4 Steps in ANOVA

factor, as well as its levels. Once this task is done, we can dig deeper into the ANOVA by following the steps described in Fig. 6.4. We will discuss each step in more detail in the following sections.

Check Assumptions

The assumptions under which ANOVA is reliable are similar as to the parametric tests discussed above. That is, the dependent variable is measured on at least an interval scale and is normally distributed in the population (see the normality tests in Box 6.1). In terms of violations of these assumptions, ANOVA is rather robust, at least in cases where the group sizes are equal. Consequently, we may also use ANOVA in situations when the dependent variable is ordinaly scaled and not normally distributed, but then we should ensure that the group-specific sample sizes are equal.⁹

Thus, when carrying out experiments, it is useful to collect equal-sized samples of data across the groups (as in our example, in which we have observations in ten stores per promotion campaign type). When carrying out ANOVA the population variances in each group should also be similar. Even though ANOVA is rather robust in this respect, violations of the assumption of homogeneous variances can significantly bias the results, especially when groups are of very unequal sample size.¹⁰ Consequently, we should always test for homogeneity of variances, which is

⁹Nonparametric alternatives to ANOVA are, for example, the χ^2 -test of independence (for nominal variables) and the Kruskal–Wallis test (for ordinal variables). See, for example, Field (2009).

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

commonly done by using Levene's test. We already briefly touched upon this test and you can learn more about it in [Web Appendix \(→ Chap. 6\)](#).

Box 6.5 Tests to use when variances are unequal and group-specific sample sizes different

When carrying at ANOVA, violations of the assumption of homogeneity of variances can have serious consequences, especially when group sizes are unequal. Specifically, the within-group variation is increased (inflated) when there are large groups in the data that exhibit high variances. There is however a solution to this problem when it occurs. Fortunately, SPSS provides us with two modified techniques that we can apply in these situations: Brown and Forsythe (1974) and Welch (1951) propose modified test statistics, which make adjustments if the variances are not homogeneous. While both techniques control the type I error well, past research has shown that the Welch test exhibits greater statistical power. Consequently, when population variances are different and groups are of very unequal sample sizes, the application of the Welch test is generally recommended.

If Levene's test indicates that population variances are different, it is advisable to use modified F-tests such as Brown and Forsythe's test or the Welch test, which we discuss in Box 6.5 (the same holds for post hoc tests which we will discuss later in this chapter). The most severe violation of assumptions is when observations are not independent. Research has shown that when observations across groups are correlated, the probability of a type I error increases greatly. For example, when comparing ten observations across three groups, even where observations correlate only weakly at 0.05, the probability of a type I error increases to 74% instead of the 5% that is often selected. Thus, we have to make sure that the same respondents do not accidentally fill out multiple surveys or that we use questionnaires collected from the same respondents at two different points in time.

Calculate the Test Statistic

The basic idea underlying the ANOVA is that it examines the dependent variable's variation across the samples and, based on this variation, determines whether there is reason to believe that the population means of the groups (or factor levels) differ significantly.

With regard to our example, each store's sales will likely deviate from the overall sales mean, as there will always be some variation. The question is whether the difference between each store's sales and the overall sales mean is likely to be caused by a specific promotion campaign or be due to a natural variation between the stores. In order to disentangle the effect of the treatment (i.e., the promotion campaign type) and the natural deviation, ANOVA splits up 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 maths, let's see what SS_B and SS_W are all about.

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 sales mean values across the three treatment conditions (i.e., point of sale display, free tasting stand, and in-store announcements) in relation to the overall mean. However, 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, sales are always the same. Obviously, in such a case, we cannot claim that the different types of promotion campaigns had any influence on sales. On the other hand, if mean sales differ substantially across the three treatment conditions, we can assume that the campaigns influenced the sales to different degrees.

This is what is expressed by means of SS_B ; it tells us how much variation can be explained by the fact that the differences in observations truly stem from different groups. Since SS_B can be considered "explained variation" (i.e., variation explained by the grouping of data and thus reflecting different effects), we would want SS_B 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 an income variable?). Consequently, we can only interpret the explained variation expressed by SS_B in relation to the variation that is not explained by the grouping of data. This is where SS_W comes into play.

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 all the observed variation can be explained by the grouping of data. 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 this variation seems to be caused by some extraneous factor that was not accounted for in the experiment and not the grouping of data. For this reason, SS_W is also referred to as "unexplained variation." This unexplained variation can occur if we fail to account for important factors in our experimental design. For example, in some of the stores, the product might have been sold through self-service while in others personal service was available. This is a factor that we have not yet considered in our analysis, but which will be used when we look at two-way ANOVA later in the chapter. Nevertheless, some unexplained variation will always be present,

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

regardless of how good our experimental design is and how many factors we consider. That is why unexplained variation is frequently called (random) noise.

Therefore, 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. More precisely, ideally we want SS_B to be as large as possible, whereas SS_W should be as small as possible. We'll take a closer look at this in the next section.

This relationship is described in Fig. 6.5, which shows a scatter plot, visualizing sales across stores of our three different campaign types: the point of sale display, the free tasting stand, and the in-store announcements. We indicate the group mean of each level by dashed lines. If the group means were all the same, the three dashed lines would be aligned and we would have to conclude that the campaigns have the same effect on sales. In such a situation, we could not expect the point of sale group to differ from the free tasting stand group or the in-store announcements group. Furthermore, we could not expect the free tasting stand group to differ from the in-store announcements group. On the other hand, if the dashed lines differ, we would probably conclude that the campaigns' effects differ significantly.

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 the dashed lines, implying that the grouping explains the variation in sales perfectly, something that will, however, hardly ever occur.

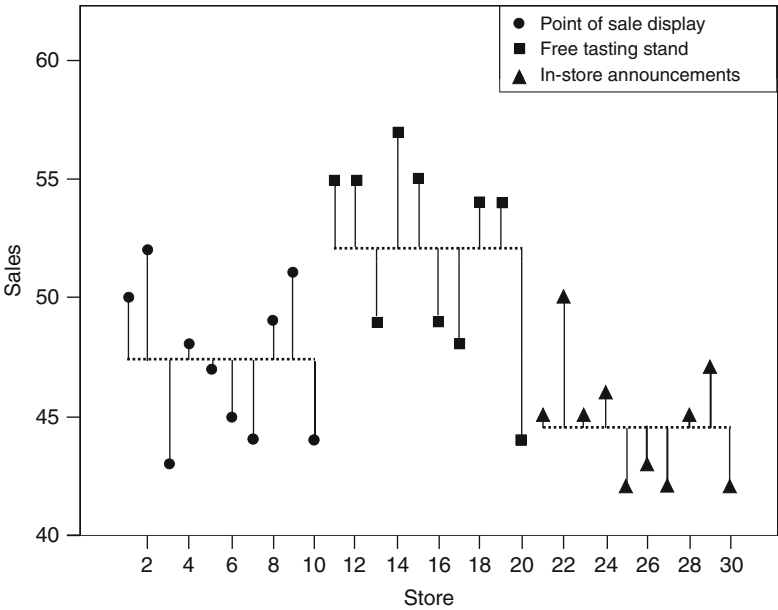


Fig. 6.5 Scatter plot of stores vs. sales

It is easy to visualize from this diagram that if the vertical bars were all, say, twice as long, then it would be difficult or impossible to draw any meaningful conclusions about the effects of the different campaigns. Too great a variation within the groups then swamps any variation across the groups.

Based on the discussion above, 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:

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

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:

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

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

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

In the previous steps, we discussed the comparison of the between-group and within-group variation. The higher the between-group variation is in relation to the within-group variation, the more likely it is that the grouping of data is responsible for the different levels in the stores' sales and not the natural variation in all 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, as these are based on summed values and, thus, are influenced by the number of scores summed. These results for SS_B and SS_W have to be normalized, which we do by dividing the values 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 squares 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 test statistic as follows:

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

This test statistic follows an F-distribution. 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 ($n-k$). Turning back to our example, we calculate the F-value as:

$$F = \frac{MS_B}{MS_W} = \frac{SS_{B/k-1}}{SS_{W/n-k}} = \frac{273.80/3-1}{310.20/30-3} = 11.916$$

For the promotion campaign example, the degrees of freedom are 2 and 27; therefore, looking at Table A2 in the  Web Appendix (→ Additional Material), we obtain a critical value of 3.354 for $\alpha = 0.05$. Note that we don't have to divide α by two when looking up the critical value! The reason is that we always test for equality of population means in ANOVA, rather than one being larger than the others. Thus, the distinction between one-tailed and two-tailed tests does not apply in this case.

Because the calculated F-value is greater than the critical value, the null hypothesis is rejected. Consequently, we can conclude that at least two of the population sales means for 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, note that rejecting the null hypothesis does not mean that all population means differ – it only means that at least two of the population means differ significantly! Market researchers often make this mistake, assuming that all means differ significantly when interpreting ANOVA results. Since we cannot, of course, conclude that all means differ from one another,

this can present a problem. Consider the more complex example in which the factor under analysis does not only have three different levels, but ten. In an extreme case, nine of the population means could be the same while one is significantly different from the rest and could even be an outlier. It is clear that great care has to be taken not to misinterpret the result of our F-test.

How do we determine which of the mean values differ significantly from the others without stepping into the α -inflation trap discussed above? One way to deal with this problem is to use *post hoc tests* which we discuss in the next section.¹²

Carry Out Post Hoc Tests

The basic idea underlying post hoc tests is to perform tests on each pair of groups, but to correct the level of significance for each test so that the overall type I error rate across all comparisons (i.e. the familywise error rate) remains constant at a certain level such as $\alpha = 0.05$. The easiest way of maintaining the familywise error rate is to carry out each comparison at a statistical significance level of α divided by the number of comparisons made. This method is also known as the *Bonferroni correction*. In our example, we would use $0.05/3 = 0.017$ as our criterion for significance. Thus, in order to reject the null hypothesis that two population means are equal, the p-value would have to be smaller or equal to 0.017 (instead of 0.05!).

As you might suspect, the Bonferroni adjustment is a very strict way of maintaining the familywise error rate. While this might at first sight not be problematic, there is a trade-off between controlling the familywise error rate and increasing the type II error, which would reduce the test's statistical power. By being very conservative in the type I error rate, such as when using the Bonferroni correction, a type II error may creep in and cause us to miss out on revealing some significant effect that actually exists in the population.

The good news is that there are alternatives to the Bonferroni correction. 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 the *Ryan/Einot-Gabriel/Welsch Q procedure* (REGWQ), 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, and (2) they assume that the population variances are equal. Fortunately, research has provided alternative post hoc tests for

¹²Note that the application of post hoc tests only makes sense when the overall F-test finds a significant effect.

situations in which these properties are not met. When sample sizes differ, 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.

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.

Measure the Strength of the Effects

To determine the strength of the effect (also *effect size*) that the factor exerts on the dependent variable, we can compute the η^2 (pronounced as *eta squared*) coefficient. It is the ratio of the between-group variation (SS_B) to the total variation (SS_T) and, as such, expresses the variance accounted for of the sample data. η^2 which is often simply referred to as effect size, 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 pronounced influence on the dependent variable. In our example η^2 is:

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

The outcome indicates that 46.9% of the total variation in sales is explained by the promotion campaigns. Note that η^2 is often criticized as being inflated, for example, due to small sample sizes, which might in fact apply to our analysis.

To compensate for small sample sizes, we can compute ω^2 (pronounced *omega squared*) which adjusts for this bias:

$$\omega^2 = \frac{SS_B - (k - 1) \cdot MS_W}{SS_T + MS_W} = \frac{273.80 - (3 - 1) \cdot 11.489}{584 + 11.489} = 0.421$$

In other words, 42.1% of the total variation in sales is accounted for by the promotion campaigns.

Generally, you should use ω^2 for small sample sizes (say 50 or less) and η^2 for larger sample sizes. 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 η^2 resembles the Pearson's correlation coefficient (Chap. 5) of linear relationships, we follow the suggestions provided in Chap. 5. Thus, we can consider values below 0.30 weak, values from 0.31 to 0.49 moderate and values of 0.50 and higher as strong.

Unfortunately, the SPSS one-way ANOVA procedure does not compute η^2 and ω^2 . Thus, we have to do this manually, using the formulae above.

Interpret the Results

Just as in any other type of analysis, the final step is to interpret the results. Based on our results, we can conclude that the promotion campaigns have a significant effect on sales. An analysis of the strength of the effects revealed that this association is moderate. Carrying out post hoc tests manually is difficult and, instead, we have to rely on SPSS to do the job. We will carry out several post hoc tests later in this chapter when dealing with an example application.

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

A logical extension of one-way ANOVA is to add a second factor to the analysis. For example, we could assume that, in addition to the different promotion campaigns, management also varied the type of service provided by providing either self-service or personal service (see column “Service type” in Table 6.1). In principle, a two-way ANOVA works the same way as a one-way ANOVA, except that the inclusion of a second factor necessitates the consideration of additional types of variation. Specifically, we now have to account for two types of between-group variations (1) the between-group variation in factor 1 (i.e., promotion campaigns), and (2) the between-group variation in factor 2 (i.e., service type).

In its simplest usage, the two-way ANOVA assumes that these factors are mutually unrelated. However, in real-world market research applications this is rarely going to be the case, thereby requiring us to use the more complex case of related factors.

When we take two related factors into account, we not only have to consider each factor’s direct effect (also called *main effect*) on the dependent variable, but also the factors’ *interaction effect*. Conceptually, an interaction effect is the additional effect due to combining two (or more) factors. Most importantly, this extra effect cannot be

Box 6.6 A different type of interaction



<http://tinyurl.com/interact-coke>

observed when considering each of the factors separately and thus reflects a concept known as *synergy*. There are many examples in everyday life where the whole is more than simply the sum of the parts as we know from cocktail drinks, music or paintings (for a very vivid example of interaction, see the link provided in Box 6.6).


In our example, the free tasting stand might be the best promotion campaign when studied separately, but it could well be that when combined with personal service, the point of sale display is much more effective. A significant interaction effect indicates that the combination of the two factors is particularly effective or, on the other hand, ineffective, depending on the direction of the interaction effect. Conversely, an insignificant interaction effect suggests that we should choose the best level of the two factors and then use them in combination. The calculation of these effects as well as discussion of further technical aspects go beyond the scope of this book but are discussed in the  Web Appendix (→ Chap. 6).

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 t-tests and one-way ANOVA in SPSS

Theory	Action
<i>One-sample t-test</i>	
Compare mean value with a given standard	► Analyze ► Compare Means ► One-Sample T Test
Assumptions:	
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 if the observations are independent.
Is the test variable normally distributed or is $n > 30$?	If necessary, carry out normality tests: ► Analyze ► Descriptive Statistics ► Explore ► Plots. Check Normality plots with tests. For $n < 50$, interpret the Shapiro–Wilk test. If test variable exhibits many identical values or for higher sample sizes, use the Kolmogorov–Smirnov test (with Lilliefors correction).
Specification:	
Select the test variable	Enter the variable in the Test Variable(s) box
Specify the standard of comparison	Specify the test value in the Test Value box.
Results interpretation:	
Look at test results	Look at the t-value and its significance level.
<i>Independent samples t-test</i>	
Compare the differences in the means of the independent samples	► Analyze ► Compare Means ► Independent-Samples T Test
Assumptions:	
Is the test variable measured on at least an interval scale?	Check Chap. 3 to determine the measurement level of your variables.
Are the observations independent?	Consult Chap. 3 to determine if the observations are independent.

(continued)

Table 6.5 (continued)

Theory	Action
Is the test variable normally distributed or is $n > 30$?	If necessary, carry out normality tests: ► Analyze ► Descriptive Statistics ► Explore ► Plots. Check Normality plots with tests. For $n < 50$, interpret the Shapiro–Wilk test. If test variable exhibits many identical values or for higher sample sizes, use the Kolmogorov–Smirnov test (with Lilliefors correction).
Specification: Select the test variable and the grouping variable Results interpretation: Look at test results	Enter these in the Test Variable(s) and Grouping Variable boxes. If the Levene’s test suggests equal population variances, interpret the t-value and its significance level based on the pooled variance estimate (upper row in SPSS output); if Levene’s test suggests unequal population variances, interpret the t-value and its significance level based on the separate variance estimate (lower row in SPSS output).
<i>Paired samples t-test</i> Compare the differences in the means of the paired samples Assumptions: Are the test variables measured on an interval or ratio scale? Are the observations dependent? Are the test variable normally distributed or is $n > 30$?	► Analyze ► Compare Means ► Paired-Samples T Test Check Chap. 3 to determine the measurement level of your variables. Consult Chap. 3 to determine if the observations are dependent. If necessary, carry out normality tests: ► Analyze ► Descriptive Statistics ► Explore ► Plots. Check Normality plots with tests. For $n < 50$, interpret the Shapiro–Wilk test. If test variable exhibits many identical values or for higher sample sizes, use the Kolmogorov–Smirnov test (with Lilliefors correction).
Specification: Select the paired test variables Results interpretation: Look at test results	Enter these in the Paired Variables box. Look at the t-value and its significance level.
<i>One-way ANOVA</i> Compare the means of three or more groups Specification: Select the dependent variable and the factor (grouping variable) Assumptions: Is the dependent variable measured on an interval or ratio scale? Are the observations independent? Is the dependent variable normally distributed?	► Analyze ► Compare Means ► One-Way ANOVA Enter these in the Dependent List and Factor box. Check Chap. 3 to determine the measurement level of your variables. Consult Chap. 3 to determine if the observations are independent. Carry out normality tests: ► Analyze ► Descriptive Statistics ► Explore ► Plots. Check Normality plots with tests. For $n < 50$, interpret the Shapiro–Wilk test.

(continued)

Table 6.5 (continued)

Theory	Action
	If test variable exhibits many identical values or for higher sample sizes, use the Kolmogorov–Smirnov test (with Lilliefors correction).
Are the population variances equal?	Carry out Levene’s test: ► Analyze ► Compare Means ► One-Way ANOVA ► Options ► Homogeneity of variance test
Results interpretation: Look at test results	Check the F-value and its significance level; if Levene’s test suggests different population variances and the group sizes differ significantly, interpret using Welch test. ► Analyze ► Compare Means ► One-Way ANOVA ► Options ► Welch
Look at the strength of the effects Carry out pairwise comparisons	Use SPSS output to compute η^2 and ω^2 manually Carry out post hoc tests: ► Analyze ► Compare Means ► One-Way ANOVA ► Post Hoc Equal population variances assumed: - Use R-E-G-WQ if group sizes are equal - Use Hochberg’s GT2 if group sizes differ. If unequal population variances assumed, use the Games–Howell procedure.

Example

Founded in 2001, vente-privee.com organizes exclusive sales events for international designer brands covering a broad range of product categories, ranging from fashion to home wares, sports products to electronics. Sales events last from 2 to 4



<http://en.vente-privee.com>

days offering 50–70% discounts exclusively to its 9.5 million members. vente-privee.com therefore offers a new distribution channel, which complements of traditional channels. Here, companies are given the opportunity to sell their stock in a quick, discreet, and qualitative way while still protecting and enhancing their brand image. In 2008, vente-privee.com sold 38 million products through 2,500 sales, achieving a turnover of 610 million GBP.

In 2008, *vente-privee.com* successfully launched its new online appearance for the UK. After 2 years of operations, the company decided to consult a market research firm to explore the customers' buying behavior and their perception of the website. In a small-scale ad hoc study, the firm gathered data from a sample of 30 randomly selected customers on the following variables (variable names in parentheses):

- Identification number (*id*)
- Respondent's gender (*gender*)
- Perceived image of *vente-privee.com* (*image*)
- Perceived ease-of-use of *vente-privee.com* (*ease*)
- Sales per respondent per year in GBP (*sales*)

Using these data, the market research firm sought to identify potential areas of improvement for the website design and ideas on how to further explore relevant target groups in future advertising campaigns. Specifically, the firm wanted to answer the following research questions:

- Is there a significant difference in sales between male and female customers?
- Does the website's perceived ease-of-use influence the customers' buying behavior?

The dataset used is *vente_privee.sav* (📁 Web Appendix → Chap. 6).¹³

Before conducting any testing procedure, we have to examine whether the variable of interest (i.e. sales in GBP) is normally distributed, as this determines whether we should consider parametric or nonparametric tests. This is done using normality test, which we can access by clicking ► Analyze ► Descriptive Statistics ► Explore. This will open a dialog box similar to Fig. 6.6.

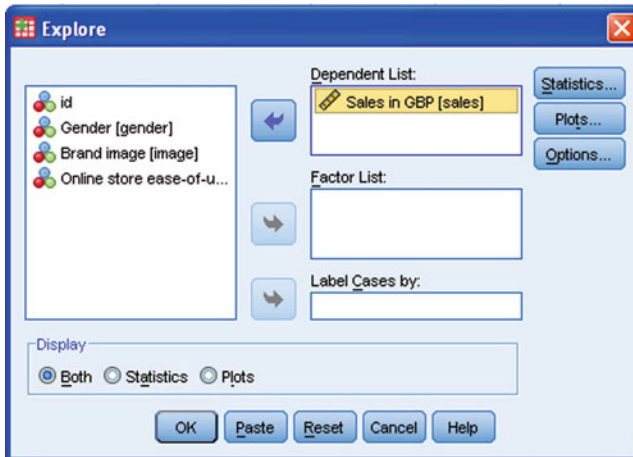


Fig. 6.6 One-sample Kolmogorov–Smirnov test dialog box

¹³Note that the data are artificial.

Table 6.6 Normality tests results

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Sales in GBP	.077	30	.200 [*]	.985	30	.931

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

Simply enter the variable *sales* in the **Dependent List** box. Under **Plots**, we need to check **Normality plots with tests**. Next, click on **Continue** and then **OK** to run the analysis.

The output in Table 6.6 shows the results for both, the Kolmogorov–Smirnov as well as the Shapiro–Wilk test. As we only have 30 observations in this example, we should interpret the Shapiro–Wilk test as it exhibits more statistical power in situations with small sample sizes. Remember that the p-value is the probability of obtaining a test statistic at least as extreme as the one actually observed, assuming that the null hypothesis is true. This means that if the p-value is smaller than or equal to our level of tolerance of the risk of rejecting a true null hypothesis (i.e., α), we should reject the null hypothesis.

In this analysis, we follow the general convention by using a 5% significance level. Because the p-value (0.931) is much larger than 0.05, the null hypothesis that the data are normally distributed should not be rejected at a 5% level. Consequently, we may assume the sales data are normally distributed in the population.

The next question that we want to address is whether there is a difference in sales based on gender. Using the previous test’s results, we know that the sales variable is normally distributed and since it is also measured on an interval scale, we have to use a parametric test. In this case, we examine two distinct groups of customers (i.e., males and females). Thus, looking at our guideline in Fig. 6.2, we have to use the independent samples t-test. To run this test, go to ► Analyze ► Compare Means ► Independent-Samples T Test to display a dialog box similar to Fig. 6.7.

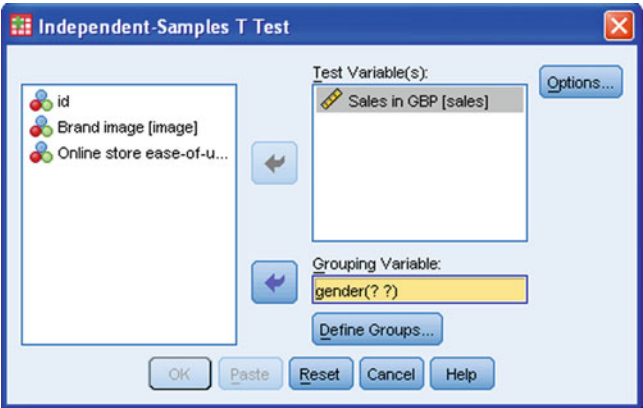


Fig. 6.7 Independent-samples t-test dialog box

Move *sales* to the **Test Variable(s)** box and move *gender* to the **Grouping Variable** box.

Click **Define Groups** and a menu similar to that shown in Fig. 6.8 will appear in which you have to specify the grouping variable’s values that identify the two groups you wish to compare.

You can also specify a **Cut point**, which is particularly useful when you want to compare two groups based on an ordinal or continuous grouping variable. For example, if you want to compare younger vs. older members you could arbitrarily 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. However, we will stick to the first option, using 1 for group 1 (males) and 2 for group 2 (females). Click on **Continue** and then on **OK** in the main menu. This will yield the outputs shown in Tables 6.7 and 6.8.

Looking at the descriptive statistics in Table 6.7, we can see that female customers have a higher mean in sales (310 GBP) than male customers (270 GBP). At first sight, it appears that the sales are really different between these two groups, but, 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 output of this test appears 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 very low F-value of 0.001 already suggests that we cannot reject the null hypothesis that the population variances are equal. This is also mirrored in the large p-value of 0.974 (**Sig.** in Table 6.8), which lies far above our α -level of 0.05. Looking at the central and right part of the output, we can see that SPSS carries out two tests, one



Fig. 6.8 Definition of groups

Table 6.7 Descriptive statistics (group statistics)

Group Statistics					
Gender		N	Mean	Std. Deviation	Std. Error Mean
Sales in GBP	male	15	270.00	50.723	13.097
	female	15	310.00	52.299	13.503

Table 6.8 Independent samples t-test result

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
Sales in GBP	Equal variances assumed	.001	.974	-2.126	28	.042	-40.000	18.811	-78.533	-1.467
	Equal variances not assumed			-2.126	27.974	.042	-40.000	18.811	-78.535	-1.465

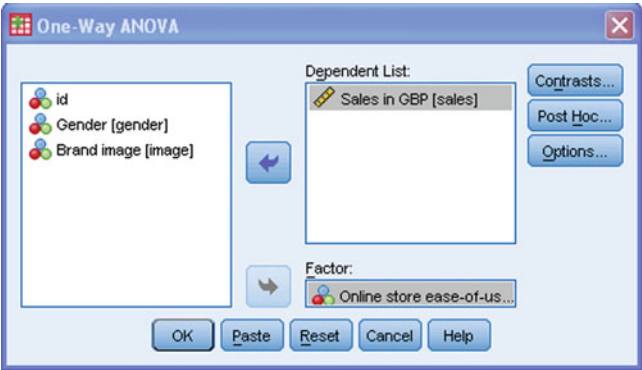


Fig. 6.9 One-way ANOVA dialog box

based on the pooled variance estimate (upper row) and the other based on separate variance estimates (lower row). Since we assume that the population variances are equal, we have to consider the pooled variance estimate. The resulting t-test statistic of -2.126 is negative due to the negative mean difference of -40 (sample mean of group 1 – sample mean of group 2). This yields a p-value of 0.042 , which is barely below the threshold value of 0.05 . Therefore, we can still reject the independent samples t-test’s null hypothesis, namely that there is no difference in the population mean sales between male and female consumers. We can therefore conclude that the sales for men and women differ significantly.

In the next analysis, we want to examine whether the customers’ perceived ease-of-use of *vente-privee.com*’s online appearance has a significant effect on sales. Reconsidering the guideline in Fig. 6.2, we can easily see that a one-way ANOVA is the method of choice for answering this research question. Clicking on ► Analyze ► Compare Means ► One-Way ANOVA will open a dialog box similar to Fig. 6.9.

We start off by moving the *sales* variable to the **Dependent List** box and the *ease-of-use* variable (*ease*) to the **Factor** box (remember that in ANOVA, factor refers to the grouping variable). Under **Options**, we can request several statistics.

As discussed above, we have to determine whether the assumption of homogenous variances is met. For this purpose – just as in the independent samples t-test – we have to consider Levene's test, which we can request by choosing the **Homogeneity of variance test**. Because we do not yet know whether the population variances are equal or not, we should choose the statistic **Welch** (for the Welch test). Also, make sure you tick the boxes next to **Descriptive** as well as **Means plot** (Fig. 6.10). Click **Continue**.

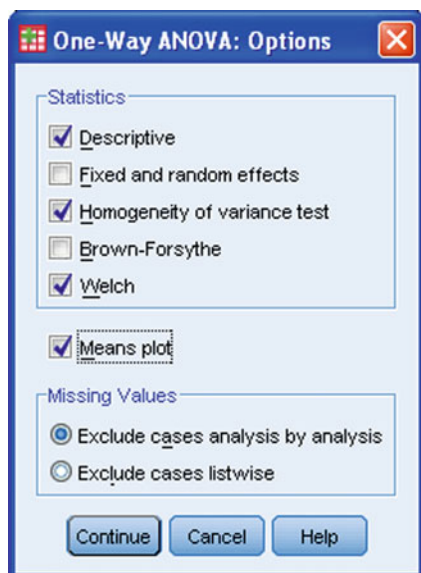


Fig. 6.10 Options for one-way ANOVA

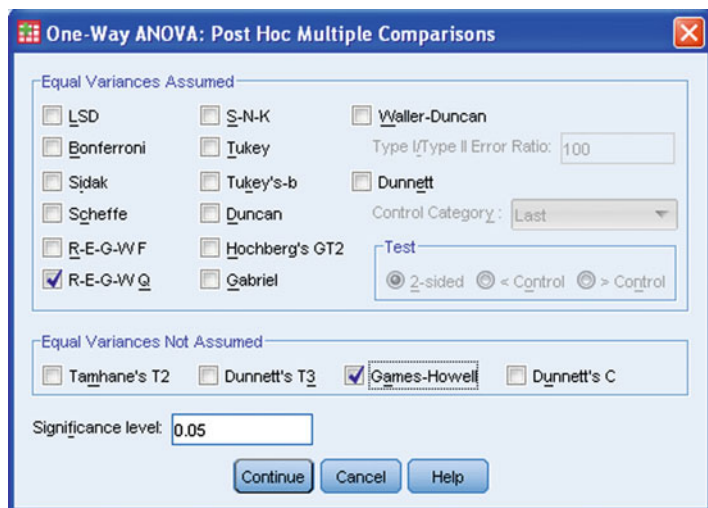


Fig. 6.11 Post hoc tests

Under **Post Hoc** (Fig. 6.11) we can specify a series of post hoc tests for multiple group comparisons. Since we do not yet know the result of Levene’s test, let’s choose the Ryan/Einot-Gabriel/Welsch Q procedure (**R-E-G-W Q**) if we can assume equal variances, as well as **Games-Howell** if we have to reject Levene’s test’s null hypothesis that population variances are equal. Since group sizes are equal, there is no need to select Hochberg’s GT2. Next, click on **Continue** to get back to the main menu.

In the main menu, click on **OK** to run the analysis. Table 6.9 shows the descriptive results. Not unexpectedly, the higher the customers rate the perceived ease-of-use of the online store, the greater the sales. This is also illustrated in the means plot in Fig. 6.12.

Table 6.9 Descriptive statistics

Descriptives								
Sales in GBP								
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
easy	10	318.90	49.660	15.704	283.38	354.42	256	413
neutral	10	291.80	41.464	13.112	262.14	321.46	230	348
difficult	10	259.30	58.532	18.509	217.43	301.17	178	330
Total	30	290.00	54.555	9.960	269.63	310.37	178	413

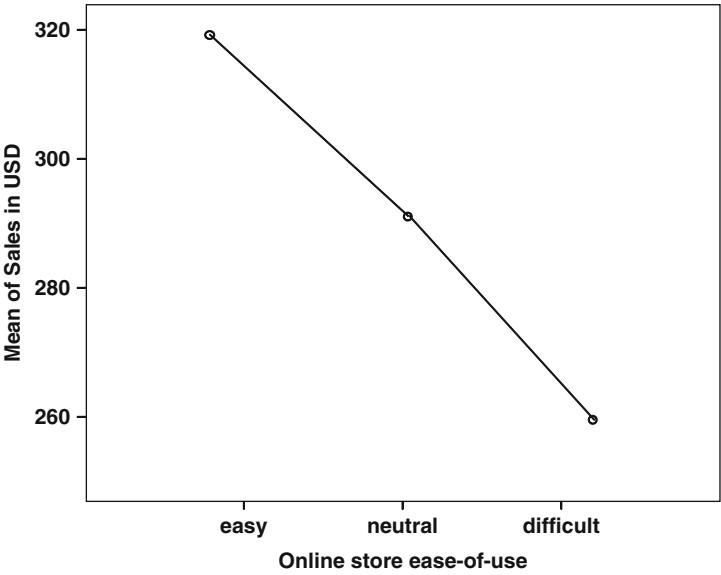


Fig. 6.12 Means plot

Table 6.10 Levene’s test output (test of homogeneity of variances)

Test of Homogeneity of Variances			
Sales in GBP			
Levene Statistic	df1	df2	Sig.
.924	2	27	.409

Table 6.11 ANOVA results

ANOVA					
Sales in GBP					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	17809.400	2	8904.700	3.510	.044
Within Groups	68502.600	27	2537.133		
Total	86312.000	29			

Table 6.12 Ryan/Einot–Gabriel/Welsch Q-procedure’s result

Sales in GBP				
Online store ease-of-use		N	Subset for alpha = 0.05	
			1	2
Ryan-Einot-Gabriel- Welsch Range	difficult	10	259.30	
	neutral	10	291.80	291.80
	easy	10		318.90
	Sig.		.161	.239

Before we can evaluate the ANOVA’s results, we have to take a closer look at the results of Levene’s test as shown in Table 6.10. Levene’s test clearly suggests that the population variances are equal, as the test’s p-value is well above 0.05. Thus, to decide whether at least one group mean differs from the others, we should consider using the regular F-test, rather than the Welch test.

Table 6.11 shows that we can reject the null hypothesis that the sales of the three groups of customers are equal (**Sig.** = **0.044** which is smaller than 0.05). More precisely, this result suggests that at least two group means differ significantly.¹⁴

To evaluate whether all groups are mutually different or only two, we take a look at the post hoc test results. As Levene’s test showed that the population variances are equal, we are primarily interested in the Ryan/Einot-Gabriel/Welsch Q-procedure’s

¹⁴Contrary to this, the Welch test suggests that there are no differences between the three groups (p-value 0.081 > 0.05). This divergent result underlines the importance of carefully considering the result of the Levene’s test.

Table 6.13 Games–Howell test result

Dependent Variable: Sales in GBP		Multiple Comparisons					
(I) Online store ease-of-use		(J) Online store ease-of-use	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Games-Howell	easy	neutral	27.100	20.458	.401	-25.26	79.46
		difficult	59.600	24.274	.061	-2.49	121.69
	neutral	easy	-27.100	20.458	.401	-79.46	25.26
		difficult	32.500	22.683	.348	-25.95	90.95
	difficult	easy	-59.600	24.274	.061	-121.69	2.49
		neutral	-32.500	22.683	.348	-90.95	25.95

result (Table 6.12). This table is a summary of the differences in the means. It organizes the means of the three groups into “homogeneous subsets” – subsets of means that do not differ at $p \leq 0.05$ are grouped together, and subsets that do differ are placed in separate columns. According to the Ryan/Einot-Gabriel/Welsch Q-procedure, groups that do not appear in the same column differ significantly at $p \leq 0.05$. Notice how the difficult and the easy groups show up in separate columns. This indicates that those groups differ significantly. The neutral group shows up in each column, indicating that it is not significantly different from either of the other two groups. Means for groups in homogeneous subsets are displayed.

Even though the population variances are assumed to be equal, let us take a look at the results of the Games-Howell procedure (Table 6.13) for the sake of comprehensiveness. There are several comparisons listed in the table. In the first row, you can see the comparison between the easy group and neutral group. The difference between the means of these two groups is 27,100 units. Following this row across, we see that this difference is statistically insignificant (p -value = 0.401). Similarly, in the row below this, we can see that the difference between the easy and difficult group (59.600) is also insignificant, as the p -value (0.061) lies above 0.05. Lastly, the comparison of the neutral and difficult group (two rows below) also renders a insignificant result. This difference in results compared to those of the Ryan/Einot-Gabriel/Welsch Q-procedure underlines the importance of closely considering Levene’s test before interpreting the analysis results.

Finally, we want to examine the strength of the effect by computing η^2 and ω^2 manually (remember that SPSS does not provide us with these measures). Using the information from Table 6.11, we can easily compute the measures as follows:

$$\eta^2 = \frac{SS_B}{SS_T} = \frac{17,809.40}{86,312.00} = 0.206$$

$$\omega^2 = \frac{SS_B - (k - 1) \cdot MS_W}{SS_T + MS_W} = \frac{17,809.40 - (3 - 1) \cdot 2,537.133}{86,312.00 + 2,537.133} = 0.143$$

As one can see, the strength of the effect is rather weak. Taken jointly, these results suggest that the buying behavior of customers is little influenced by their perceived ease-of-use of vente-privee.com’s online store. Of course, improving the

online store's navigation or layout might boost sales across all three groups jointly, but it will not necessarily introduce differences in sales between the groups.

So far, we have considered only one factor in the ANOVA but we could easily extend this example by simultaneously considering a second factor, say one which captures the customers' perceived image of *vente-privee.com* as a company. This would then require the application of a two-way ANOVA, which we discuss in the Web Appendix (🔗 Web Appendix → Chap. 6).

Case Study

In the spring of 2010, the German unit of Citibank was sold to *Crédit Mutuel*, a French bank. After the purchase, the former Citibank was renamed *Targobank* and a massive re-branding campaign was launched. The new *Targobank*'s product range was also restructured and service and customer orientation were stressed. In addition, a comprehensive marketing campaign was launched, aimed at increasing the bank's customer base by one million new customers in 2015.

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

- Gender (*gender*)
- Bank deposit in Euro (*deposit*)
- Does the customer still attend school/university? (*training*)
- Customer's age specified in three categories (*age_cat*)

Use the data provided in *bank.sav* (🔗 Web Appendix → Chap. 6) to answer the following research questions.¹⁵

1. Which test do we have to apply to find out whether there is a significant difference in bank deposits between male and female customers? Do we meet the assumptions necessary to conduct this test? Also use an appropriate normality test and interpret the result. Does the result give rise to any cause for concern? Carry out an appropriate test to answer the initial research question.
2. Is there a significant difference in bank deposits between customers who are still studying and those that are not?
3. Which type of test or procedure would you use to evaluate whether bank deposits differ significantly between the three age categories? Carry out this procedure and interpret the results.

¹⁵Note that the data are artificial.

4. Reconsider the previous question and, using post hoc tests, evaluate whether there are significant differences between the three age groups.
5. Is there a significant interaction effect between the variables *training* and *age_cat* on *deposit*?
6. On the basis of your analysis results, please provide recommendations on how to align future marketing actions for Targobank's management.

Questions

1. Describe the steps involved in hypothesis testing in your own words.
2. Explain the concept of the p-value and explain how it relates to the significance level α .
3. What level of α would you choose for the following types of market research studies? Give reasons for your answers.
 - (a) An initial study on the preferences for mobile phone colors
 - (b) The production quality of Rolex watches
 - (c) A repeat study on differences in preference for either Coca Cola or Pepsi
4. Write two hypotheses for each of the example studies in question 3, including the null and alternative hypotheses.
5. Describe the difference between independent and paired samples t-tests in your own words and provide two examples of each type.
6. Use the data from the *vente-privee.com* example to run a two-way ANOVA, including the factors (1) ease-of-use and (2) brand image, with sales as the dependent variable. To do so, go to Analyze ► General Linear Model ► Univariate and enter *sales* in the Dependent Variables box and *image* and *online_store* in the Fixed Factor(s) box. Interpret the results.

Further Readings

Field A (2009) *Discovering statistics using SPSS*, 3rd edn. Sage, London

Good reference for advanced types of ANOVA.

Hubbard R, Bayarri MJ (2003) Confusion over measure of evidence (p's) versus errors (α 's) in classical statistical testing. *Am Stat* 57(3):171–178

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

Kanji GK (2006) *100 statistical tests*, 3rd edn. Sage, London

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

Sawyer AG, Peter JP (1983) The significance of statistical significance tests in marketing research. *J Mark Res* 20(2):122–133

Interesting article in which the authors discuss the interpretation and value of classical statistical significance tests and offer recommendations regarding their usage.

References

- Boneau CA (1960) The effects of violations of assumptions underlying the t test. *Psychol Bull* 57(1):49–64
- Brown MB, Forsythe AB (1974) Robust tests for the equality of variances. *J Am Stat Assoc* 69(346):364–367
- Cohen J (1992) A power primer. *Psychol Bull* 112(1):155–159
- Field A (2009) *Discovering statistics using SPSS*, 3rd edn. Sage, London
- Hubbard R, Bayarri MJ (2003)[AU4] Confusion over measure of evidence (p 's) versus errors (α 's) in classical statistical testing. *Am Stat* 57(3):171–178
- Lilliefors HW (1967) On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *J Am Stat Assoc* 62(318):399–402
- Schwaiger M, Sarstedt M, Taylor CR (2010) Art for the sake of the corporation: Audi, BMW Group, DaimlerChrysler, Montblanc, Siemens, and Volkswagen Help Explore the Effect of Sponsorship on Corporate Reputations. *Journal of Advertising Research* 50(1):77–90
- Welch BL (1951) On the comparison of several mean values: an alternative approach. *Biometrika* 38(3/4):330–336