The Core Concept of Statistics

3

Contents

What You Will Learn in This Chapter

In Chaps. 1 and 2, we showed that proper conclusions need full information and that many popular measures, such as the Odds Ratio or the Percentage of Correct Responses, provide only partial information. In this chapter, we use the framework of SDT to understand statistical inference, including the role of the *p*-value, a dominant term in statistics. As we show, the *p*-value confounds effect size and sample size and, hence, also provides only partial information.

This chapter is about the essentials of statistics. We explain these essentials with the example of the *t*-test, which is the most popular and basic statistical test. This is the only chapter of this book where we go into details because, we think, the details help a great deal in understanding the fundamental aspects of statistics. Still, only basic math knowledge is required. The hasty or math phobic reader can go directly to Sect. 3.3 *Summary*, where we summarize the main findings and the key steps. Understanding at least this Summary is necessary to understand the rest of the book.

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_3

3.1 Another Way to Estimate the Signal-to-Noise Ratio

In Chap. 2 we defined d' as the distance between population distribution means, divided by the population standard deviation. Typically, we label one of the populations as noise-alone, with mean μ_N, and the other population as signal-and-noise, with mean μ_{SN}. In statistics, the d' of populations is also often referred to as Cohen's δ or effect size. The calculation is the same as in Chap. 2:

$$\delta = d' = \frac{\mu_{SN} - \mu_N}{\sigma} \tag{3.1}$$

Oftentimes we do not have population information, and we want to estimate δ (i.e., d') from empirical data. For example in Chap. 2, we could estimate d' in the patch present vs. absent experiment by computing z(Hit)-z(FA) just from the behavioral data. We did not have any knowledge about the underlying means and variance of the Gaussians. This estimation approach is useful when we cannot directly measure the underlying variables driving the performance of the system but we can measure decision outcomes. In other situations we cannot easily measure decision outcomes but we can estimate the means and variance of the Gaussians directly. For example, we can use a sonar and record the sonar responses in many trials when a rock is present. We plot the results and obtain a graph, from which we can estimate the mean and variance of the Gaussian. We can obtain the mean and variance also for the no-rock condition. Next, the sample means $\overline{x}_{SN}$ and $\overline{x}_N$ and the standard deviation s can be used to compute an estimated effect size called Cohen's d:

$$d = \frac{\overline{x}_{SN} - \overline{x}_N}{s} \tag{3.2}$$

Once again, this standardized effect d is simply an estimate of the d' of the population distributions. If d is large, it will be fairly easy to distinguish a single measurement as coming from the signal-and-noise distribution or from the noise-alone distribution. Regrettably, for many situations that scientists care about the value of d is quite small. For example, within psychology, a value of around $d = 0.8$ is considered to be "large," a value around $d = 0.5$ is considered to be "medium," and a value around $d = 0.2$ is considered to be "small." As Fig. 3.1 shows, even "large" values of d correspond to considerable overlap between distributions. For situations like these we are never going to have a good ability to correctly discriminate *single* measurements. All is not lost, though, as long as you are willing to discriminate the *means* of those measurements. As we will see, the properties of SDT apply for discriminating means similarly to discriminating single measurements.

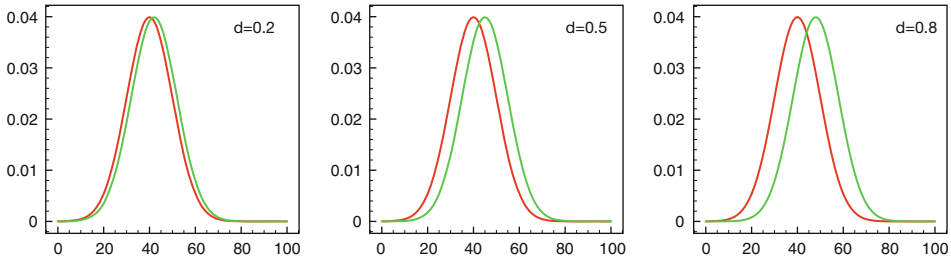

Fig. 3.1 Population distributions with small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) effect sizes

Terms

Because very similar concepts were developed in many fields, there are many identical terms, which we list here:

- Hit Rate = Power
- False Positive Rate = False Alarm = Type I error
- Miss Rate = Type II error
- d' = Cohen's δ = effect size = standardized effect size
- Gaussian distribution = Normal distribution = Bell curve
- Sample values, such as tree height, are also called Scores

Some definitions

We collected a sample of n scores x_i

- Sample mean
  $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  where the symbol $\sum$ means "add up all following terms"
- Sample variance
  $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$
- Sample standard deviation
  $s = \sqrt{s^2}$
- Standard error
  $s_{\overline{x}} = s/\sqrt{n}$

Facts about sample means
For $n \to \infty$:

1. The distribution of sample means $\overline{x}$ is Gaussian (Central Limit Theorem; CLT)
2. $\overline{x} \to \mu$
3. $s_{\overline{x}} \to 0$

3.2 Undersampling

Let us start with an example. We are interested in the hypothesis that the mean height of Alpine oak trees at the Northern rim is different than the mean height of oaks from the Southern rim. The straightforward way to address this hypothesis is to measure the height of *all* trees in the North and South rims, compute the means, and compare them. If the means are different, they are different. If they are the same, they are the same. It is that easy.

Fig. 3.2 A small section of a forest in the Swiss Alps

Unfortunately, there are many oaks (Fig. 3.2) and our resources are limited, so we can measure only a certain number, n, of both the North and South trees and measure their heights. The collected trees of a population are called a *sample*, i.e., in this case we collected two samples, one from the North and one from the South trees. The tree heights we measured are also called *scores* and the mean height in a sample is called the *sample mean*. The number n of trees of a sample is called the *sample size*. For each sample, there is most likely a difference between the *mean* height of the *sampled* trees and the *true* mean height of *all* trees of the population. For example, we may have, just by chance, sampled more large than small trees. This difference is called the *sampling error*. Thus, because of *undersampling* (measuring fewer than all trees), we likely do not obtain accurate estimates of the two means. Importantly, we choose the trees for our samples randomly, a procedure called random sampling.

Let us now collect both a sample from the North and a sample from the South trees. If we find a difference of sample means we cannot know whether it was caused by a true difference of the tree population means or whether the population mean heights were the same but the difference was caused by undersampling. Hence, undersampling may lead to wrong conclusions. For example, even though there is no difference between the means for the North and South tree populations, we may decide that there is a difference because there was a difference in the sample means. In this case, we are making a False Alarm, also called a Type I error.

To understand how undersampling influences decisions, we will first study how likely it is that a sample mean deviates from the true mean by a certain amount. As we will see, the sampling error is determined by the standard deviation of the population, σ, and the sample size, n. Second, we will study how undersampling affects how well we can discriminate whether or not there is a difference in the mean height of the *two* tree populations. A simple equation gives the answer. The equation is nothing else but a d for mean values. Hence, we are in a SDT situation. Third, we want to control the Type I error rate. We will see that the famous p-value just sets a criterion for the Type I error.

3.2.1 Sampling Distribution of a Mean

To begin with, let us focus on the North tree population. To arrive at a sample mean, we collect a sample of North trees, measure the height x_i for each tree, sum these heights up, and divide by the sample size n. The sample mean is an estimate of the true mean μ_{North}:

$$\bar{x}_{North} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.3}$$

where x_i is the height of the i-th tree we sampled. Similarly, we can estimate the variance of the tree heights, s^2 (see box). The difference between this sampled mean and the true mean is

$$\bar{x}_{North} - \mu_{North} \tag{3.4}$$

How large is this difference *on average*? To answer this question, assume—for the sake of the argument—we are going many times to the woods and randomly choose a sample of a fixed sample size, n. The various samples most likely contain different trees because we chose the trees randomly. How many sample means are close to the true mean and how many are far from it? Let us assume that we know the true population mean and standard deviation. Let us first collect a sample of only two trees and compute the mean. In the example of Fig. 3.3, the true mean is 20 m and the mean of the sample is 19 m. Thus, the difference is 1 m. Let us collect another sample of two trees. This error is likely different from the previous one because we have likely chosen two trees with different heights. Let us continue measuring two trees and see how the sample means are distributed. The Central Limit Theorem tells us that the distribution of the sample means is similar to a Gaussian function. The Gaussian is centered around the true mean, thus many sample means reflect well the true mean. However, the Gaussian is quite broad, i.e., the standard deviation is large, and hence quite some sample means deviate substantially from the true mean.

Let us now collect 9 instead of 2 samples and repeat the procedure as before. Again, we obtain a Gaussian distribution, which is, however, narrower than for the sample of two trees, i.e., the standard deviation is smaller, and, thus, it is much more unlikely that the mean of a randomly chosen sample deviates strongly from the true mean. In general, for each sample size n, there is such a sampling distribution. The larger the sample size n, the smaller is the standard deviation of the sampling distribution. This is not surprising because the error is zero if we measure all trees and small if we fail to measure only a few trees. Hence, the standard deviation $\sigma_{\bar{x}}$ of the sampling distributions is a measure of how good we expect our estimate of the mean to be. $\sigma_{\bar{x}}$ is called the *standard error of the mean*, and it can be shown that $\sigma_{\bar{x}}$ is equal to the standard deviation of the true population distribution σ divided by the square root of the sample size n:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3.5}$$

If we do not know σ, then we can estimate the standard error by using the sample estimate of the standard deviation:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{3.6}$$

The equation shows again why with larger sample sizes the sampling error becomes smaller: as $\sqrt{n}$ goes larger, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ goes to zero. $\sigma_{\bar{x}}$ depends on both n and σ. Suppose

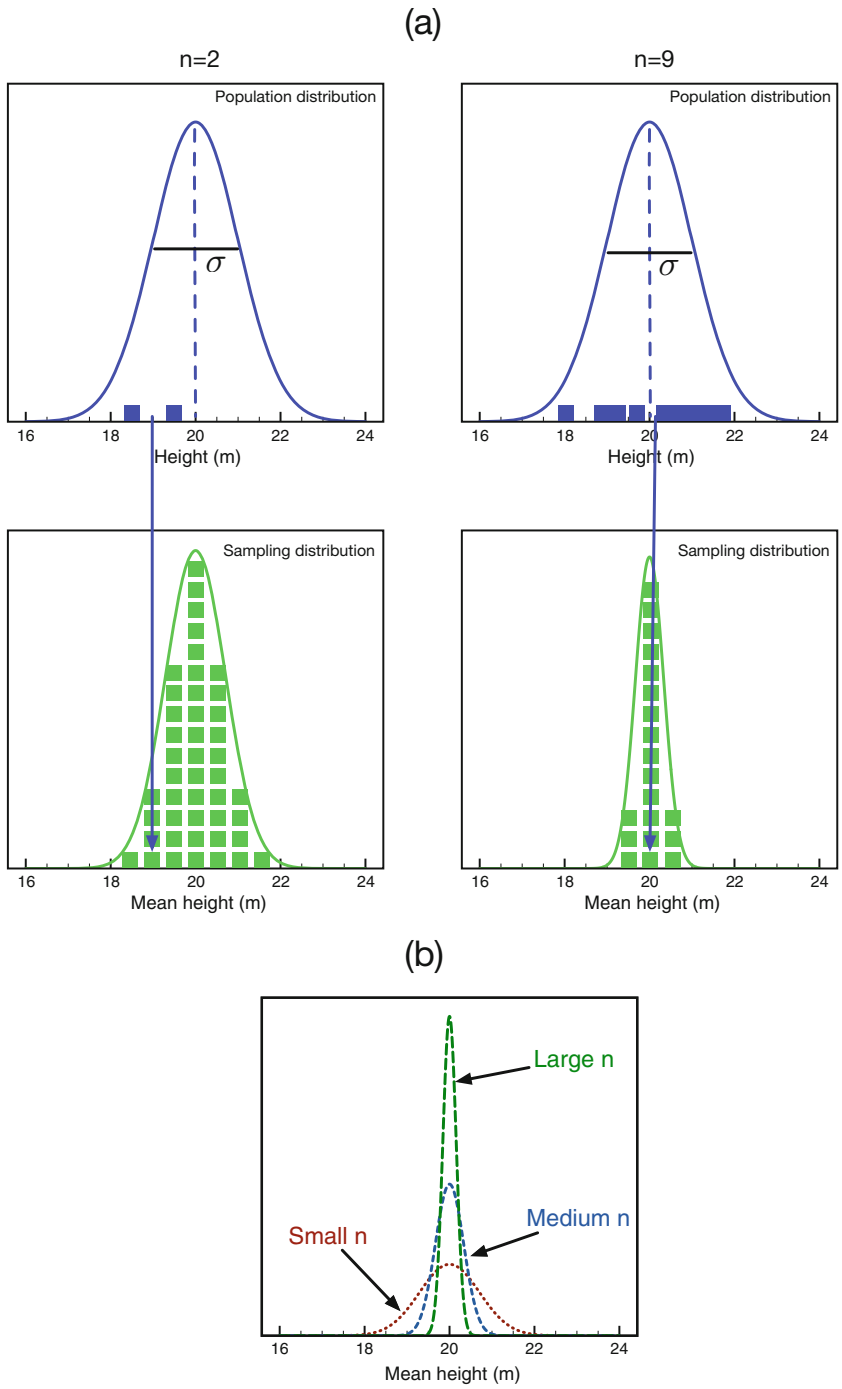

Fig. 3.3 Let us focus on the North trees only. Since we cannot measure the height of all trees, we collect samples with a mean value that, because of undersampling, is likely different from the

σ is zero, then all trees in our sample have the same height, which means all trees have the mean height μ_{North}, and hence we need to only measure the height of one tree. On the other hand, if σ is large, we need to sample many trees to obtain a good estimate of the population mean.

Summary Because of undersampling, sample means likely differ from the true mean. The standard error, $s_{\overline{x}}$, is a measure for the expected sampling error.

3.2.2 Comparing Means

Let us next see how undersampling affects a *comparison* of the means of the North and South trees. Obviously, there is a family of sampling distributions for the South trees too. In the following, we assume that the sample sizes and the population variances are the same for both tree populations. As mentioned, if our two samples contain all trees from both populations, we can simply compare the means and note any difference. For smaller samples, both sample means may strongly differ from the true means. First, we subtract the two sample means: $\overline{x}_{North} - \overline{x}_{South}$. For each pair of samples of North and South trees, we can compare the difference of sample means with the difference of the true means $\mu_{North} - \mu_{South}$. Hence, we have only one sampling distribution and the same situation as in the last subsection.

As in the last subsection, the sampling distribution is a Gaussian with a mean equal to the difference of the population means, $\mu_{North} - \mu_{South}$. Moreover, something called the "variance sum law" tells us that the standard deviation of this sampling distribution is

Fig. 3.3 (continued) true mean. (**a**) Left. The Gaussian at the top row shows the true population distribution. Let us first measure the height of only two randomly collected trees (blue boxes). The heights are 18.5 m and 19.5 m. Hence, the mean of the two samples is 19 m, shown as the green box in the graph below, highlighted by the arrow. Hence, the sampling error of the mean is 1 m since the true mean is 20 m. Let us now go to the woods and collect more samples of two trees. Each green box shows one of the corresponding sample means. After having collected many samples, we obtain a Gaussian function. The *y*-axis shows the probability how likely it is that a certain mean value occurs. Right. Now let us sample nine trees. The true population distribution is the same as on the Left and shown again on the top row along with one sample of nine trees. The corresponding sample mean is shown by the arrow. Again, if we sample more trees, we arrive at more mean values (green boxes). The Gaussian is more narrow, i.e., the standard deviation of the sample means is much smaller. Hence, whereas for samples of two trees, an error of 2 m is not unlikely, such an error is unlikely for samples of nine trees. (**b**) For any population distribution there is a family of sampling distributions, one for each sample size n. All sampling distributions are Gaussians. For increasingly large sample sizes n, we obtain sampling distributions with decreasingly smaller standard deviations. Hence, when n increases it becomes much more unlikely that the sample mean strongly differs from the true mean. This is not surprising because if we measure the height of all trees, there is no sampling error. The standard deviation is zero. If we fail to measure just a few trees, the error is very small

related to the standard deviation of the populations and the sample size:

$$\sigma_{\overline{x}_{North}-\overline{x}_{South}} = \sigma\sqrt{\frac{2}{n}} \tag{3.7}$$

This term is the *standard error* for the difference of sample means.[1] If we do not know the population means and standard deviation, we consider the estimate:

$$s_{\overline{x}_{North}-\overline{x}_{South}} = s\sqrt{\frac{2}{n}} \tag{3.8}$$

Let us recall our main question. We have collected one sample from the North trees and one sample from the South trees, respectively, with a sample size of n. Most likely there is a difference between the two sample means, i.e., $\overline{x}_{North} - \overline{x}_{South} \neq 0$. Does this difference come from undersampling even though there is no difference between the population means, or does this difference reflect a true difference in mean height? This is a classic SDT situation—just with means instead of single measurements. How well can we discriminate between the two alternatives? We can answer the question by computing the d' or Cohen's δ between the two alternatives. For the first alternative $\mu_{North} - \mu_{South} = 0$, meaning that there is no difference between the mean heights of the North and South trees, i.e., the noise alone distribution. The corresponding sampling distribution is centered around 0 since there is no difference. For the second alternative, there is a real difference and the sampling distribution is centered at $\mu_{North} - \mu_{South}$. Because we do not know the true values, we use estimates.[2]

So, we have now estimated two sampling distributions: one for when there is a real difference (signal-and-noise) with mean $\overline{x}_{North} - \overline{x}_{South}$ and one when there is no difference (noise-alone) with mean 0. Hence, we have exactly the same situation as in the yellow submarine example and estimate d' or Cohen's δ of the sampling distributions, which is usually called t, by:

$$t = \frac{(\overline{x}_{North} - \overline{x}_{South}) - (0)}{s_{\overline{x}_{North}-\overline{x}_{South}}} \tag{3.9}$$

The t-value is nothing else as a d' applied to sampling distributions. Just as for all SDT situations, if t is large, it is fairly easy to distinguish whether a difference of means comes

[1] If the samples from the North and South trees are of different sizes, then the formula is $\sigma_{\overline{x}_{North}-\overline{x}_{South}} = \sigma\sqrt{\frac{1}{n_{North}} + \frac{1}{n_{South}}}$.

[2] Typically, the value s is computed by pooling the variances from the two samples. We describe one way of doing this pooling in Sect. 3.4.

from the signal with noise distribution or from the noise-alone distribution. If t is small, then it will be difficult to determine whether there is a true difference.[3]

Let us substitute the estimate of the standard error into the t-equation:

$$t = \frac{(\overline{x}_{North} - \overline{x}_{South}) - (0)}{s_{\overline{x}_{North} - \overline{x}_{South}}} = \frac{(\overline{x}_{North} - \overline{x}_{South})}{s\sqrt{\frac{2}{n}}} = \frac{(\overline{x}_{North} - \overline{x}_{South})}{s}\sqrt{\frac{n}{2}} = d\sqrt{\frac{n}{2}} \tag{3.10}$$

By splitting up the standard error, which is the measure for the sampling error, into s and n, we see that the t-value is d (the estimated δ of the population distributions) multiplied by the square root of half the sample size.

We can interpret the t-value in two ways. First, we are interested in whether there is a real difference between the mean values. For this reason, we subtract the estimates of the two means to see how much they differ. However, a large difference is meaningless when the noise, i.e., the standard deviation is high. For this reason, as in Chap. 2, we divide by an estimate of the standard deviation, which in this case is the estimated standard deviation of the sampling distribution of the difference of means. The estimated standard deviation of the sampling distribution is the standard error:

$$s_{\overline{x}_{North} - \overline{x}_{South}} = \frac{s}{\sqrt{\frac{n}{2}}} \tag{3.11}$$

Hence, the standard error of the sampling distribution of the means combines both sources of uncertainty, namely the population standard deviation and the uncertainty from undersampling. Second, we see that the t-value is the product of the estimated d of the population distribution and a function of the sample size n. The t-value combines effect and sample size.

Summary We wanted to know whether or not two means are identical but have difficulties to decide because we have only inaccurate estimates caused by undersampling. This is a classic discrimination task, just with means instead of single measurements. The t-value, which is easy to calculate from the samples we collected, is nothing else than the estimated d' for this situation. Most importantly, the t-value is a function of the estimated effect size d and the sample size n, namely, a multiplication of the estimated effect size d and the square root of the sample size $n/2$.

[3] Following the convention of SDT, we will always interpret t as being a positive number, unless we specifically say otherwise. Should the computed value be negative, one can always just switch the means in the numerator.

3.2.3 The Type I and II Error

Undersampling may create an error, which may lead to wrong conclusions. For example, we may decide that there is a true mean difference in the height of the North and South trees—even though there is none—because there was a difference in the sample means (False Alarm). Likewise, we may decide that there is no true difference—even though there is one—because the sample mean difference was small (Miss). Following the statistical conventions, we call a False Alarm a Type I error and a Miss a Type II error. How do we cope with these errors? As we have seen in Chap. 2, False alarms and Misses depend on where we set our criterion. The same is true here and the four possible outcomes of a decision are summarized in Fig. 3.4. Commonly, people focus on a special hypothesis called the *null hypothesis*: there is no difference between the population means. In terms of SDT, the null hypothesis claims that, even though an observed difference of sample means occurs, the difference comes from undersampling, i.e., from the noise-alone distribution. The alternative hypothesis, H_a or H_1, is that there is a difference between the two population means. In terms of SDT, the alternative hypothesis states: the observed difference of sample means comes from the signal-and-noise distribution. In Fig. 3.4 we refer to this as "H_0 is False."

As in the yellow submarine example, a large t tells us that the discrimination between population means should be easy, while a small t-value indicates that discrimination should be hard and we may easily arrive at wrong conclusions. It is now straightforward to decide about the null hypothesis. We compute t and then apply a criterion. If the computed t-value is greater than the criterion, we take that as evidence that the estimated difference

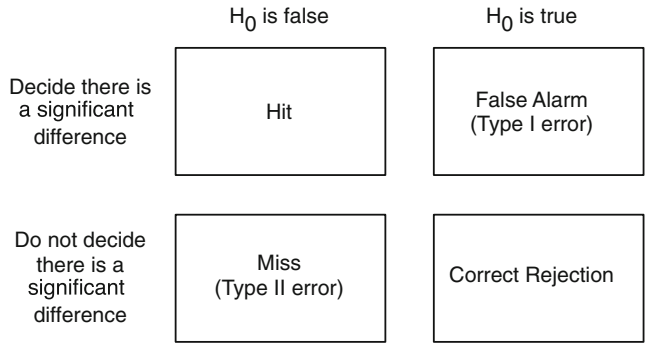

Fig. 3.4 Statistics is about making conclusions about a hypothesis. We have, similar to Chaps. 1 and 2, four outcomes. (1) The null hypothesis is false and we come to the conclusion that there is a difference in the means (Hit), (2) the null hypothesis is false and we have insufficient evidence against the null hypothesis (Miss, or Type II error). (3) The null hypothesis is true and we come to the conclusion that there is a difference in the means (False Alarm, Type I error). (4) The null hypothesis is true and we come to the conclusion we have insufficient evidence against the null hypothesis (Correct Rejection)

of means did not come from the noise-alone distribution: there is a difference between the two means. If the computed t-value is smaller than the criterion, then we do not have confidence that there is a difference between the two means. Maybe there is a difference, maybe not. We do not make any conclusions.

In practice, different fields use different criteria, which reflects their relative comfort levels with making Hits or False Alarms. For example, physics often follows the "5σ rule," which requires $t > 5$ to claim that an experiment has found sufficient evidence that there is a difference between mean values. Compared to other fields, this is a very high criterion; and it partly reflects the fact that physicists often have the ability (and resources) to greatly reduce σ and s by improving their measurement techniques. In particle physics, the Large Hadron Collider produces trillions of samples. Fields such as medicine, psychology, neuroscience, and biology, generally use a criterion that (to a first approximation) follows a "2σ rule." This less stringent criterion partly reflects the circumstances of scientific investigations in these fields. Some topics of interest are inherently noisy, and the population differences are small. Simultaneously, the per-unit cost for medical or biological samples is often much higher than for many situations in physics; and in some situations (e.g., investigations of people with rare diseases) a large sample size is simply impossible to acquire.

SDT also tells us that any chosen criterion trades off Hits and False Alarms, and the 5σ and 2σ rules are no exception. Everything else equal, the 5σ rule will have fewer Hits than the 2σ rule. Likewise, everything else equal, the 5σ rule will have fewer False Alarms than the 2σ rule.

Rather than setting a criterion in terms of standard deviation σ, in many fields (including medicine, psychology, neuroscience, and biology), scientists want to keep the Type I error smaller than a certain value, e.g., 0.05. It should be clear why one wants to limit this kind of error: it would cause people to believe there is an effect when there really is not. For example, one might conclude that a treatment helps patients with a disease, but the treatment is actually ineffective, and thus an alternative drug is not used. Such errors can lead to deaths. From a philosophical perspective, scientists are skeptical and their default position is that there is no difference: a treatment does not work, an intervention does not improve education, or men and women have similar attributes. Scientists will deviate from this default skepticism only if there is sufficient evidence that the default position is wrong.

Summary A Type I error occurs when there is no difference in the means, i.e. the null hypothesis is true, but we decide there is one. The Type I error is a False Alarm in terms of SDT. To decide about the null hypothesis, we compute t, which reflects the discriminability in terms of SDT, and then apply a criterion.

3.2.4 Type I Error: The p-Value is Related to a Criterion

Here, we show how the criterion determines the Type I error rate. Let us consider what the sampling distribution of the difference of sample means looks like when the Null hypothesis H_0 is true, i.e., $\mu_{North} - \mu_{South} = 0$. The distribution is centered on zero with a standard error that we estimated from our data. Suppose we set the criterion to $t = 2.0$, which is often called the critical value (cv) and written $t_{cv} = 2.0$. If our data produces a t-value larger than $t_{cv} = 2.0$, we will decide that there is a difference between the sample means-even though there is none, i.e, a Type I error. The probability of such a t-value is the area under the curve beyond $t_{cv} = 2.0$ (see Fig. 3.5a). This kind of test is called a "one-tailed t test". Calculating this area (assuming large sample sizes) gives 0.0228. Thus, if you use the criterion $t_{cv} = 2.0$, you will make a Type I error with a probability of only

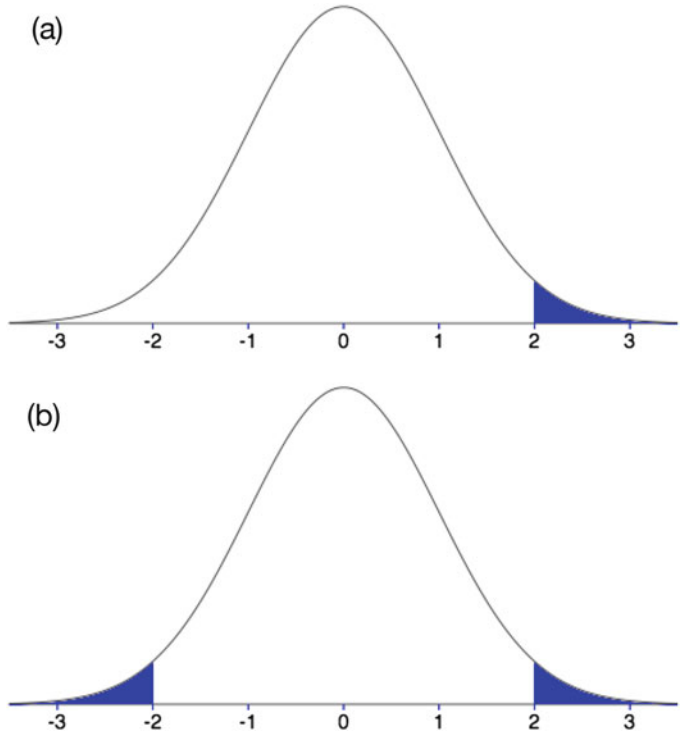

Fig. 3.5 The relation between criterion critical values and Type I error rates. The curve shows the noise alone distribution, i.e., when the Null hypothesis is true. The distribution is centered at zero and the variance is estimated from the data. (**a**) With a critical value of $t_{cv} = 2.0$, the Type I error rate is the area under the curve larger than t_{cv}. This test is called a one-sample t-test. (**b**) With a critical value of $t_{cv} = \pm 2.0$, the Type I error rate is the area under the curve for scores more extreme than ± 2.0

0.0228. If you followed the physics approach and used a 5σ rule, then $t_{cv} = 5$ and the Type I error rate is 0.000000287.

There is great flexibility in this approach. For example, you might suspect that Northern and Southern trees have different heights, but have no guess as to which would be larger. In this scenario, you might use a criterion $t_{cv} = \pm 2.0$, where a t-value more extreme (further from zero) than 2.0 would be taken as evidence that there is a difference in population means. With this approach, the Type I error rate would be 0.0456, which is twice as large as for the one-tailed test (see Fig. 3.5b). This test is called a "two-tailed t test".

In the above example, we set a criterion and computed the Type I error for this criterion. In statistics, usually, it is the other way around. We fix a Type I error rate and compute the corresponding criterion t-value. For example, if we accept a 5% Type I error rate, the corresponding criterion t-value is $t_{cv} = \pm 1.96$ for the two-tailed t test if n is large.[4] Rather than specify a certain t-value as the criterion, people compute the area under the curve beyond the t-value computed from the data. This area is called the p-value (see Fig. 3.6). Hence, we compute the t-value from the data and then the p-value. The p-value tells how likely it is that, if the Null hypothesis is true, we obtain our t-value or an even larger one. If the p-value is smaller than 0.05, we call the effect significant. Hence, to control the Type I error rate one simply requires that the computed p-value be less than the desired Type I error rate.

As mentioned, within medicine, psychology, neuroscience, and biology, a common desired rate is 0.05. For large sample sizes and for situations where one considers both positive and negative t-values (two-tailed t-test), a $p = 0.05$ corresponds to $t = \pm 1.96$. Thus, setting the Type I error rate to 0.05 corresponds to setting a criterion of $t_{cv} = \pm 1.96$. This relationship is why these fields follow an (approximate) 2σ rule. Whereas the t-value can be computed by hand, we need a statistics program to compute the p-value.

Summary If the t-value is larger than a certain value (which depends on the sample size n), we conclude that there is a significant effect.

3.2.5 Type II Error: Hits, Misses

In general, SDT tells us that for a given d', setting a criterion not only determines the Type I error rate, it also establishes the rate of Hits, Misses, and Correct Rejections. Indeed, it is easy to see that using a criterion that sets the Type I error rate to 0.05 also determines the Correct Rejection rate (the rate of concluding there is insufficient evidence for an effect when there really is no effect) to be $1.0 - 0.05 = 0.95$. As shown by the blue

[4] For small sample sizes, the t_{cv} criterion is larger because the sampling distributions are not quite Gaussian shaped. Statistical software that computes the p-value automatically adjusts for the deviation of sampling distributions from a Gaussian shape.

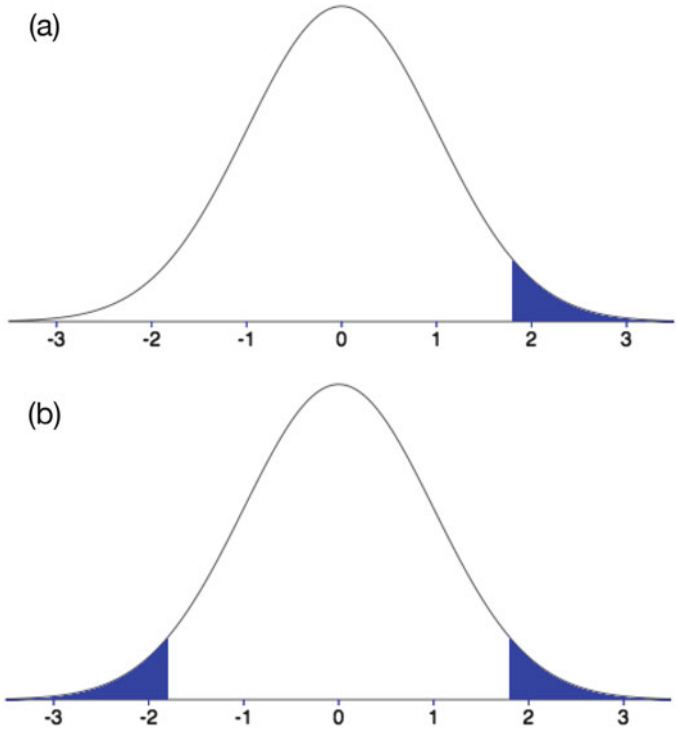

Fig. 3.6 The relation between t test statistics and p-values. (**a**) For a one-tailed test with $t = 1.8$, the p-value is the area under the curve larger than 1.8, $p = 0.0359$. (**b**) For a two-tailed test with $t = 1.8$, we compute the area more extreme than ± 1.8 in both tails. Here the p-value is 0.0718

area in Fig. 3.7, for a given criterion the area under this alternative sampling distribution to one side or the other corresponds to the probability of taking samples that produce a Hit (exceed the criterion and conclude evidence for a difference of population means) and Type II error (not satisfy the criterion and not conclude evidence for a difference of population means).

Hence, it seems that it would likewise be easy to compute the Type II error rate. However, this is not the case. When computing the Type I error, we know that the sampling distribution corresponding to the Null hypothesis is centered at one value, namely, 0. Hence, there is only one Null hypotheses. However, there are infinity many alternative hypotheses (see Implication 2e). But perhaps, we are only interested in substantial differences between the means of the North and South trees when the North trees are at least 1.2 m larger than the South trees. In this case, we know the minimal separation between the population distributions and can ask the question how large the sample size n must be to reach a significant result at least 80% of the time.

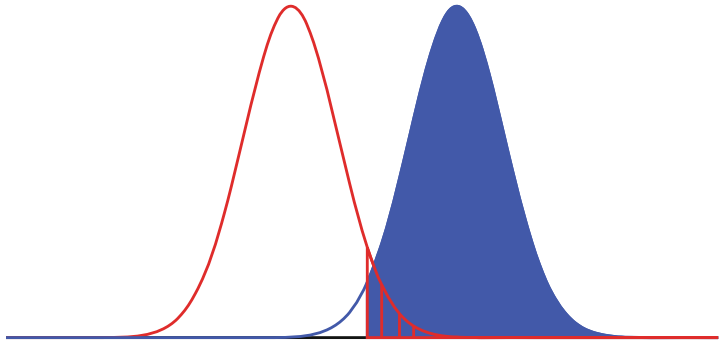

Fig. 3.7 Sampling distributions and the calculation of the Hit rate. The criterion corresponds to the lower end of the red hatched region and the lower end of the blue filled region. Any t-value that falls above this criterion will lead to a decision to reject the H_0 and conclude there is a difference in population means. The area in blue is the probability that the H_a sampling distribution will produce a t-value from this region. The Type I error rate is the red hatched area under the red curve

The Type II error plays an important role in terms of power. Power is the probability to obtain a significant result when indeed there is an effect, i.e., the Alternative hypothesis is true. Power is just another term for the Hit rate. The Hit rate is 1—Type II error rate. Power will be crucial in Part III and further explained in Chap. 7.

3.3 Summary

The above considerations are fundamental for understanding statistics. For this reason, we spell out the main steps once more and highlight the most important points. Even if you did not go through the above subsections, the main ideas should be understandable here.

We were interested in whether the *mean* height of oak trees of the North rim of the Alps is the same as for the South rim. The question is easy to answer. We just measure all trees, compute the two means, and know whether or not there is difference. If we miss a few trees, we obtain estimates, which are likely not too different from the true means. The fewer trees we measure, the larger is the *sampling error*, i.e., the more likely it is that our two sample means differ substantially from the two true means. As we have shown, this sampling error can be quantified by the standard error $s_{\overline{x}}$, which depends on the population standard deviation σ and the sample size n. If σ is small, we need only to sample a few trees to get a good estimate of the mean. For example, if $\sigma = 0$ we need only to sample one tree from each population because all trees have the same height in each population. If σ is large, we need to sample many trees to get a good estimate of the mean.

Let us now collect a sample of n trees from both the North and the South rim of the Alps with the samples being much smaller than the number of all trees in the two populations. We compute the mean heights for both samples. Because of undersampling,

almost surely there is some difference between the two sample means. However, we cannot know whether the observed difference indicates that there is a real difference in the means of the populations or whether the observed difference occurs because of undersampling while the population means are actually identical. If the true means are identical but we conclude, based on the sample mean difference, that the true means are different, we are making a Type I error (Fig. 3.4). Scientists generally want to avoid making a Type I error because their default position is that there is no effect until the data suggest otherwise. No decision making process can avoid sometimes making a Type I error, but we can control how often we make such an error. The important value is the t-value, which we can easily compute by hand:

$$t = \frac{(\overline{x}_{North} - \overline{x}_{South})}{s}\sqrt{\frac{n}{2}} = d\sqrt{\frac{n}{2}} \tag{3.12}$$

We compute the sample means $\overline{x}_{North}$ and $\overline{x}_{South}$ from the n trees x_i we collected, we estimate the standard deviation s of the trees (see grey box), and multiply by a function (the square root) of the sample size $n/2$. The right hand side shows that the t-value is nothing else as an estimate of the effect size d multiplied with a function of the sample size. The t-value tells us how easily we can discriminate whether or not a difference in the sample means comes from a real difference of the population means. The situation is exactly as in Chap. 2. The t-value is nothing else as a d' where, instead of dividing by the standard deviation, we divide by the standard error, which is a measure of the sampling error, taking both sources of noise, population variance and undersampling, into account. A large t-value means we can easily discriminate between means and a small t-value suggests the decision is hard. Note that a large t-value can occur because the estimated effect size d is large, n is large, or both are large.

Assume that there is no effect, i.e., the mean height of the North and South trees is identical ($\delta = 0$), then the p-value tells us how likely it is that a random sample would produce a t-value at least as big as the t-value we just computed. Thus, if we are happy with a 5% Type I error rate and the p-value is smaller than 0.05, we call our mean difference "significant".

The p-value is fully determined by the t-value and is computed by statistics programs. Most importantly, the t-value combines an estimate of the effect size d with the sample size ($\sqrt{\frac{n}{2}}$), which is why the t-value, and hence the p-value, confounds effect and sample size and, therefore, represents only partial information! This insight will be important to understand several implications, which we present after the following example.

3.4 An Example

Computing the p-value is simple, as the following short example will show. Understanding the implications of the t-test is more complicated.

Let us assume that we collected the heights of five trees from the North and five trees from the South. The data are presented in the first column of Fig. 3.8. The computations for the two-tailed t-test are also presented in the figure. For the given sample sizes and the computed t-value, our statistical software program tells us that the corresponding p-value is 0.045. Since this p-value is smaller than 0.05, we conclude that the data indicates a significant difference between the mean heights of Northern and Southern trees.[5]

Results from tests like this are often summarized in a table as presented in Table 3.1. The p-value is in the column marked "Sig. (2-tailed)." In the table, degrees of freedom (df) are mentioned. The degrees of freedom are important for the computation of the p-value because the shape of the sampling distribution is slightly different from a Gaussian for small sample sizes. In addition, one can compute the df from the sample size and vice-versa. In the case of the t-test, $df = n_1 + n_2 - 2 = 5 + 5 - 2 = 8$.

As mentioned, significance does not tell you too much about your results. It is important to look and report the effect size. Cohen proposed guidelines for effects sizes for a t-test, which are shown in Table 3.2.

Take Home Messages

- Since the p-value is determined by the t-value, it confounds effect size (d) and sample size (n). The original idea behind the t-test was to provide tools to understand to what extent a significant result is a matter of random sampling, given a certain effect size d. Nowadays, the p-value is often mistaken as a measure of effect size, which was never intended and is simply wrong!
- Partial information: proper conclusions can only be based on both the estimated population effect size, d, and the sample size, n. Hence, it is important to report both values, to take both values into account for conclusions, and to understand whether a significant result is driven by the estimated effect size d, the sample size, or both.

[5] Alternatively, one could identify a critical value criterion, $t_{cv} = \pm 2.306$ and note that t is farther from zero than this critical value.

North	$(x_{i,North} - \bar{x}_{North})^2$
20.80	(20.80 - 18.00)^2 = 7.840
17.81	0.036
17.92	0.006
18.30	0.090
15.17	8.009
$\bar{x}_{North} = \sum_{i=1}^{n} \frac{x_{i,North}}{n} = 18.00$	$s^2_{North} = \frac{\sum_{i=1}^{n} (x_{i,North} - \bar{x}_{North})^2}{n-1} = 3.995$
South	$(x_{i,South} - \bar{x}_{South})^2$
16.91	3.648
15.28	0.078
13.70	1.690
16.81	3.276
12.30	7.290
$\bar{x}_{South} = \sum_{i=1}^{n} \frac{x_{i,South}}{n} = 15.00$	$s^2_{South} = \frac{\sum_{i=1}^{n} (x_{i,South} - \bar{x}_{South})^2}{n-1} = 3.996$

$$df = (5-1) + (5-1) = 4+4 = 8$$

$$s_p^2 = \frac{s^2_{North}(n-1) + s^2_{South}(n-1)}{(n-1)+(n-1)} = \frac{3.995(4) + 3.996(4)}{4+4} = 3.996$$

$$s_{\bar{x}_{North}-\bar{x}_{South}} = \sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{n}\right)} = \sqrt{3.996\left(\frac{1}{5} + \frac{1}{5}\right)} = 1.264$$

$$t = \frac{\bar{x}_{North} - \bar{x}_{South}}{s_{\bar{x}_{North}-\bar{x}_{South}}} = \frac{18.00 - 15.00}{1.264} = 2.373$$

Fig. 3.8 The computations for the independent samples t-test start with computing the means for each data column ($\bar{x}_{North}$ and $\bar{x}_{South}$ shown at the bottoms of the first column). Using these, the variances can be computed (s^2_{North} and s^2_{South} at the bottoms of the second column). The degrees of freedom (df) for each column is computed by taking the number of scores in the column minus one. The pooled variance (s_p^2) is then computed by taking a weighted sum of the two variance terms (weighting by the degrees of freedom). We then substitute the pooled variance into the formula for the standard error $s_{\bar{x}_{North}-\bar{x}_{South}}$ and use this in the denominator of our t formula. The numerator is simply the difference between the means we calculated in our first step

Table 3.1 Output from a typical statistical software package

	t	df	Sig. (2-tailed)	Cohen' s d
Tree height	2.373	8	0.045	1.5 (large effect)

The columns labeled *t*, *df*, and *Sig. (two-tailed)* yield the t-value, its corresponding degrees of freedom, and $p-value$ respectively. The *df* value reported here is the sum of df_N and df_S (i.e., $4 + 4 = 8$)

Table 3.2 Effect size guidelines for d according to Cohen

	Small	Medium	Large
Effect Size	0.2	0.5	0.8

3.5 Implications, Comments and Paradoxes

For the following implications, Eq. 3.12 is crucial, because it tells us that the t-value and, thus the p-value, are determined by the estimated d and the sample size n.

Implications 1 Sample Size

Implication 1a According to Eq. 3.12, if the estimated $d \neq 0$, then there is always an n for which the t-test is significant. Hence, even very small effect sizes can produce a significant result when the sample size is sufficiently large. Hence, not only large effect sizes lead to significant results as one might expect, any non-zero effect size leads to significant results when n is large enough.[6]

Implication 1b If the estimated $d \neq 0$ (and $d < 4.31$), then there are sample sizes $n < m$, for which the t-test is not significant for n but is significant for m.[7] This pattern may seem paradoxical if you read it as: there is no effect for n but there is an effect for m. However, this is not the correct reading. We can only conclude that for m we have sufficient evidence for a significant result but insufficient evidence for n. From a null result (when we do not reject the null hypothesis) we cannot draw any conclusions (see Implication 3). We will see in Part III that this seeming paradox points to a core problem of hypothesis testing.

Implication 1c. Provocative Question Isn't there always a difference between two conditions, even if it is just very tiny? It seems that, except for a few cases, the difference between population means $\mu_1 - \mu_2$ is never really zero. How likely is it that the North and South tree means are both exactly 20.2567891119 m? Hence, we can always find a sample size n such that the experiment is significant. Why then do experiments at all?

Implications 2 Effect Size

Implication 2a As mentioned, the p-value is not a measure of the population effect size δ and, for each $d \neq 0$, there is a n for which there is a significant outcome. Thus, small effects can be significant. According to a study, consuming fish oil daily may significantly prolong your life. However, it may prolong your life by only 2 min. Do you bother?

Implication 2b By itself, the p-value does not tell us about the effect size. For example, when the sample size increases (everything else equal), the p-value decreases because the variance of the sampling distribution becomes smaller (see Fig. 3.3). Thus, if the effect size d is exactly the same, the p-value changes with sample size.

[6] We can describe the situation also as following. If there is a real effect $d \neq 0$, e.g., between the tree means, we can find a sample size n, for which we obtain a significant result in almost all cases (or with a certain high probability).

[7] If $d > 4.31$ you do not need to compute statistics because the difference is so large. In this case, even $n = 2$ leads to a significant result.

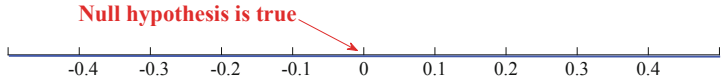

Fig. 3.9 In terms of effect size, the null hypothesis is represented by exactly one point, the null point. All other, infinitely many, points belong to the H_1 hypothesis

Implication 2c The p-values of two experiments A and B may be exactly the same. However, one cannot draw conclusions from this fact. For example, it may be that experiment A has a large effect size d and a small sample size and vice versa for experiment B. Hence, one can never compare the p-values of two experiments when the sample size is not the same. Likewise, a lower p-value in experiment A than B does not imply a larger effect size. The sample size might just be higher.

Implication 2d A study with a small sample size leading to a small p-value indicates a higher estimated effect size than a study with a larger sample size and the same p-value.

Implication 2e Just to reiterate and to visualize the basic situation. The worst case scenario is when the null hypothesis is true, i.e., there is no difference between means, and we conclude there is one: making a Type I error. In this case $\mu_1 - \mu_2 = 0$. If the null hypothesis is not true, $\mu_1 - \mu_2$ is not 0 and can be a value from $-\infty$ to $+\infty$ in principle. All of these values are part of the alternative hypothesis that there is a difference between the North and South trees. Hence, when we are worrying about the Type error I and the null hypothesis, we are worrying about only one single point embedded in infinitely many other points (see Fig. 3.9).

Implications 3 Null results

Implication 3a Absence of proof is not proof of absence: one can never conclude that there is *no* effect in an experiment ($d = 0$) when there was no significant result. A non-significant p-value indicates either that there is no difference or a real difference that is too small to reach significance for the given sample size n.

Implication 3b A difference of significance is not the same as a significant difference. Consider a study measuring the effect of a cream containing Aloe Vera on skin eczema. Patients with eczema are randomly assigned to two groups: one receiving the cream with Aloe Vera and one receiving a placebo cream. After 4 weeks, the size of the eczema is measured again. There was a significant reduction in the Aloe Vera group but not in the placebo group (Fig. 3.10). It might be tempting to conclude that the study demonstrates that Aloe Vera cures eczema. However, there is a reduction in the placebo group too—just smaller (which may be due to self-healing). In fact, when we compute the difference in eczema reduction for each participant in both groups and compute a two-tailed t-test between the two groups, the difference is not significant.

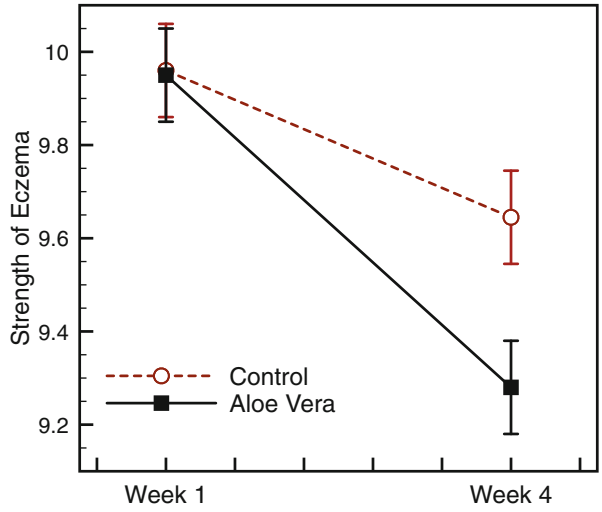

Fig. 3.10 An Aloe Vera cream was given to an experimental group and a placebo to a control group to study whether Aloe Vera can reduce the effects of eczema. The strength of eczema was tested at the beginning of the experiment and after 4 weeks. The plot shows the mean value of eczema strength for each group at each measurement. The error bars indicate the standard error for each mean value. Eczema strength significantly reduced in the experimental group but not in the control group. Can we conclude that Aloe Vera reduces eczema? We cannot because eczema strength reduced also in the control group potentially because of self-healing. Actually, there was no significant effect when the improvements in both groups were compared with a two-tailed t-test. A difference of significance is not the same as a significant difference. The graph shows the mean values and corresponding standard errors (see grey box "Some definitions")

One might argue that with a larger sample size, the difference between the two groups might become significant. This may indeed be true. However, also the effect in the placebo group may become significant. What should we conclude? Contrary to intuition, this situation does not constitute a problem because we can ask whether or not there is a stronger effect in the Aloe Vera than the placebo condition (thus, discounting for the self-healing).

Importantly, the example shows that it often makes very little sense to compare statements such as "there was an effect in condition A but not in condition B". Such conclusions are ubiquitous in science and should be treated with great care (see also Chap. 7). The classic example is as in the Aloe Vera case to compare an intervention condition, where a significant results is aimed to occur, with a control condition, where a Null result is aimed for.

Implications 4 Truth, Noise and Variability

Implication 4a Why do we compute statistics? Often it is implicitly assumed that statistics "clears off" the noise that is inevitable in complex systems. In the submarine example

measurements are corrupted by changes in the water, such as fish, algae, or the device itself, which randomly fluctuate. This kind of noise is called measurement noise. All these noise sources compromise the true signal from the rock and the true signal when there is no rock. The situation can be described with the following equation:

$$x_j = \mu + \epsilon_j$$

where x_j is the outcome in the jth measurement, μ is the true signal, and ϵ_j is the noise, which changes from trial to trial. Usually it is assumed that ϵ_j is Gaussian distributed with a mean of zero. Thus, basing decisions on one trial is not a good idea. As mentioned, averaging across many measurements can clear off the noise. That's why it is better to compare mean values than single measurements. As we have seen, the larger the n, the better the measurement of the mean.

This kind of model is appropriate in many fields, such as physics. However, in biology, medicine and many other fields, the situation is often quite different. For example, we might determine the strength of headaches before and after a pain killer was consumed. We find that the drug decreases headache strength *on average*. However, as with most drugs, there may be people who do not benefit from the drug at all. In addition, some people receive more benefit than others, i.e., for some people headaches may almost disappear while for others there is only a small (or even opposite) effect. These person-specific effects might hold for all days and situations.

Person-specific effects can be described with the following equation:

$$x_{ij} = \mu + v_i + \epsilon_{ij}$$

where x_{ij} is one measurement, e.g., person i taking the pain killer at day j. μ is the mean value of the entire population, e.g., to what extent the drug decreases headaches on average. v_i is the sensitivity of the person i for the pain killer. As mentioned, some people always benefit strongly from the drug, while for others there is no effect, and for even others headaches always increase when taking the pain killer. Hence, v_i determines how much one person differs from other persons—and the mean μ. ϵ_{ij} is measurement noise and reflects, for example, to what extent the pain killer leads to different effects from day to day in the very same person. In some way, ϵ_{ij} indicates unsystematic variability, whereas v_i captures systematic variability. Just as another example. A person may have a higher blood pressure than someone else, and this difference is reflected by the inter-participant variability v_i. At the same time, blood pressure varies greatly from minute to minute in the very same person, and this difference is reflected by ϵ_{ij}, the intra-participant variability.

In many experiments, one cannot easily disentangle v_i and ϵ_{ij}. Both terms contribute to the estimated standard deviation of the population distribution, s. From a mathematical point it does not matter whether there is strong inter-participant variability or strong measurement noise. However, for interpreting the statistical analysis, the distinction is crucial. Assume there is a strong beneficial effect of the pain killer for half of the

population whereas there is a smaller detrimental effect for the other half of the population. On average, the drug has a positive effect and this effect may turn out to be significant. Importantly, whereas the drug is beneficial *on average*, this is not true individually. For half of the population, the effect is detrimental and it is not a good idea to use the pain killer. Hence, when v_i is not zero, significant results do not allow conclusions on the individual level. A study may show that carrots are good for eye sight on average. Whether this is true for you is unclear. Carrots may actually deteriorate your vision, even though they help the vision of other people. These considerations do not imply that such studies are wrong, they just show the limitations of studies where $v_i \neq 0$ for some i. For an international comparison of blood pressure values, average values are good. However, it is usually not a good idea to compare yourself to such a large group, whatever is being measured. Such a sample is not only heterogeneous across the regions but also contains people of different ages. A body mass index of 27 may be an issue for children below 5 years but not necessarily for people older than 70 years. Hence, it depends very much on the research question to what extent a mean comparison makes sense. It is a matter of interpreting statistics, not of computing statistics.

Implication 4b The above considerations have philosophical implications. Usually, we assume that something is either the case or it is not the case. Either gravity acts on all matter in the entire universe, or it does not. Either oxygen is necessary for humans, or it is not. All of these facts hold true for each individual, i.e., for each element in the universe, for all humans, etc. If a fact has been proven by methods involving statistics, this conclusion is not necessarily justified when v_i is different from 0 because the results hold true only on average, not necessarily for all individuals.

Implication 4c The variability vs. noise problem becomes even more serious when the study contains a non-homogeneous sample differing systematically in a feature that is not explicitly considered. For example, based on how often they go to the doctor, it seems that shorter students are ill more often than taller students. However, this fact has nothing to do with body size. It is simply the case that female students are shorter than male students on average *and* see the gynecologist much more often than male students see the urologist. However, females see the gynecologist mainly for preventive medical checkups and are by no means more often ill than male students. Since students generally see doctors very infrequently, visits to the gynecologist weigh strongly in the statistics. It is obvious how mis-interpretations can occur even in such simple examples. In more complex situations such mis-interpretations are less easy to spot. By the way, one should question whether it is good idea to make conclusions about illness frequency based on doctor visits.

Implication 4d One can also consider the variability vs. noise problem the other way around. When you are planning an experiment, you need to specify whom to include. To be representative, it is good to sample from the entire population, e.g., from all people in a country or even world wide. However, with this procedure, you may include

a heterogeneous population, which makes conclusions difficult. Should you include astronauts or coma patients? What about ill people? The large portion of people with too high blood pressure? The more subpopulations you exclude, the less representative is your sample. Eventually, your sample may include only you.

Implication 4e As a last point. Effects often depend on dosage, i.e., different people may respond differently to different dosages. A pain killer may have beneficial effects for some people in a low dosage but be detrimental for a higher dosage. Hence, there is not only systematic inter-person variability but also systematic intra-person variability in addition to the unsystematic noise ϵ_{ij}. In many experiments, there are many sources involved, i.e., effects depend on dosage, inter-individual differences, and noise—limiting conclusions strongly. As we will see, dosage dependent effects are best described by correlations (Chap. 8) rather than by t-tests.

Implications 5a The Statistics Paradox and the Dangers of Cohort Studies

For large effect sizes, as they occur for example in physics, we often do not need to compute statistics. Likewise, the hypothesis that elephants are on average larger than ants does not need statistics because any living elephant is larger than any ant, δ is extremely large. The original idea of statistics was to determine whether a "bit noisy effect" really exists and to determine the sample sizes n needed to show that indeed the effect is real. We may say that statistics was developed for medium effect sizes and medium sample sizes. In the past it was usually impossible to obtain significant results with small effect sizes because data were scarce and handling large sample sizes was cumbersome. Hence, n was usually small and only experiments with large effect sizes produced significant results. This has changed largely because data collection has become cheap, and it is possible to combine and handle millions of samples as, for example, in genetics. For this reason, nowadays statistics is widely used not only for medium effects but also for very small effect sizes. However, this development is not free of danger. First of all, large sample sizes should not be confused with large effect sizes (Implication 2a). Second, conclusions are often very difficult to draw, particularly, in so called cohort studies. In cohort studies, for example, patients are compared with controls, or vegetarians are compared with meat eaters. The two groups are defined by a given label.

Here is an example. Starting in 1948, blood pressure was measured for 5209 participants in the small city of Framingham in Massachusetts. Participant age is plotted in Fig. 3.11 on the x-axis and the systolic blood pressure on the y-axis. Data were split depending on the level of education. First, there is a clear effect of age. Second, more education is associated with lower blood pressure. Using statistical techniques described in Chap. 6, the effect of education turned out to be significant. Does this mean that prolonged education *causes* lower blood pressure? Likely not. Maybe people with more education smoke less. Maybe, maybe not. They may smoke fewer cigarettes per day (dosage dependent). Maybe, maybe not. They may have started smoking later or quit

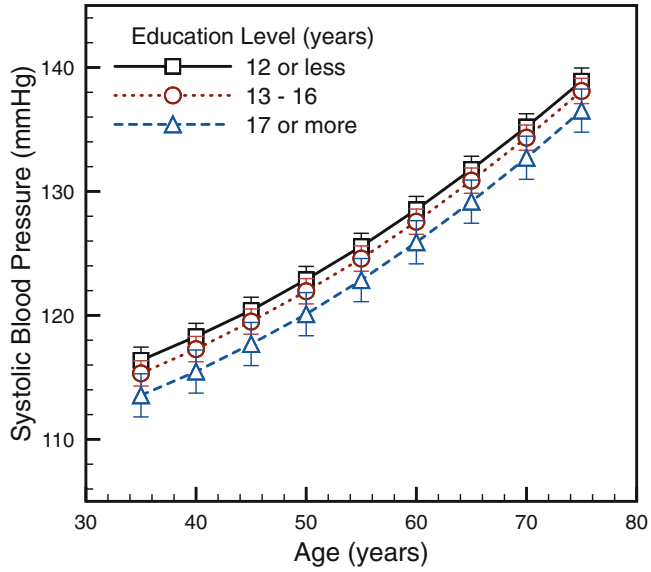

Fig. 3.11 Mean systolic blood pressure for three populations of people from the city of Framingham in Massachusetts, USA, as a function of age. The error bars indicate the standard error for each mean. The three populations differ in the number of years of education. Blood pressure increases with age. In addition, blood pressure is lowest for the group with the most years of education and is highest for the group with the fewest years of education. What can we conclude? As we argue below: not too much. Data are re-plotted from Loucks et al. [1]

earlier. Nutrition may play a role. Sports. The work environment. Genetics. Maybe there is a small subgroup, who works under very unhealthy conditions, which alone causes the higher blood pressure. There are so many potential factors, many of which are unknown today and will be discovered in the future: tangerines in your diet may lower blood pressure. Moreover, combinations of factors may play a role. Perhaps nutrition only plays a role when people do no sports.

The difference in blood pressure between the different education groups is only about 2 mm Hg. To put this effect size into perspective, measure your blood pressure and repeat 5 min later. You will see that 2 mm Hg is very small compared to your intra-person variance (ϵ_{ij}) and very low compared to the large range of inter-person variability (v_i). In addition, blood pressure strongly changes during activity. Maybe there is only a difference when the blood pressure is measured during rest. Maybe, maybe not. The main problem with these, so called, cohort studies is that there are too many factors that are causally relevant, but cannot be controlled for. To control for all these effects and the combinations, sample sizes may need to be larger than the number of people on the planet. In addition, is it really worth investigating 2 mm Hg? If you want to lower your blood pressure, a little bit of sport might do a good job and is much cheaper than paying thousands of dollars for education.

Implications 5b. Small Effects Sizes As shown, studies with small effect sizes require extra care. However, small effect size are not always problematic. First, it is good to reduce the side effects of a drug that is consumed by millions of people, even if it is only by 1%. Second, many important discoveries started off with small effects; but subsequent investigations refined the methods and produced bigger effects.

Implications 5c. Conclusions Importantly, both small and large sample sizes can be problematic. It is well known that *small* sample sizes are a problem because of undersampling. It is less well understood that *large* sample sizes may be as problematic when effect sizes are small because even tiny differences may become significant. In particular, cohort studies with small effects sizes and large sample sizes are often useless because small correlations between the investigated factor and unrelated factors can create significant results. For this reason, it is important to look at both the effect size and the sample size. Whereas the sample size n is usually mentioned, this is not always true for effect sizes. For the t-test, the effect size is often expressed as Cohen's d (see also Chap. 4). In the following chapters, we will introduce effect sizes for other tests.

How to Read Statistics? For different samples, the estimate of the effect d' may vary strongly. The larger the sample size n the less variance is there and the better is the estimate. Hence, first, look whether n is sufficiently large. If so, decide whether the effect size is appropriate for your research question. Tiny effect sizes are only in some cases important and may come from confounding, unidentifiable factors. In Part III, we will see that the combination of sample size and effect size can give interesting insights into the "believability" of a study. For example, we will ask how likely it is that four experiments, each with a small sample and effect size, all lead to significant results with p-values just below 0.05.

Take Home Messages

1. Even small effect sizes lead to significant results when the sample size is sufficiently large.
2. Do not compare the p-value of two experiments if n is not identical: a smaller p does not imply more significance.
3. Statistical significance is not practical significance.
4. Absence of proof is not proof of absence: avoid conclusions from a Null result.
5. Do not pit a significant experiment against a non-significant control experiment.
6. Cohort studies with small effects are usually useless.
7. A statement like "X is true" can only be true for sure if inter-subject variability is zero.

Reference

1. Loucks EB, Abrahamowicz M, Xiao Y, Lynch JW. Associations of education with 30 year life course blood pressure trajectories: Framingham Offspring Study. BMC Public Health. 2011;28(11):139. https://doi.org/10.1186/1471-2458-11-139.