4 Variations on the *t*-Test

Contents

What You Will Learn in This Chapter

In Chap. 3, we introduced the basic concept of statistics within the framework of SDT. Here, we present a classic introduction to hypothesis testing and present variations of the *t*-test.

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_4

4.1 A Bit of Terminology

Type of Experiment

- Experimental study: samples are randomly assigned to two groups. For example, patients are *randomly* assigned to an experimental group that takes a potentially potent drug and a control group that receives a placebo.
- Cohort study: the groups are defined by predefined labels, such as patients vs. controls, vegetarians vs. meat eaters, astronauts vs. earthlings. Cohort studies are common and useful, however, they also face severe problems as seen in Chap. 3, Implication 5a.

Type of Variables and Metrics In plots, usually, the x-axis represents the independent variable, the y-axis the dependent variable. For both variable types there are four main types of measurement scales.

- Nominal: there is no order. For example, blood pressure is determined for people from four different countries. On the x-axis, you can plot *any* order of the countries. As another example for a nominal scale: therapy A vs. B.
- Ordinal: just ranks. For example, a general has a higher rank than a lieutenant but the rank of the general is not, say, two times higher than the rank of the lieutenant. On the x-axis, you can plot the ranks in an *ascending* order. The distance between points on the x-axis has no meaning.
- Interval: values can be added or subtracted but not meaningfully divided or multiplied. A 30 °C day is 15 °C hotter than a 15 °C day but it is not twice as hot because 0 °C does not mean the absence of heat. For this reason, physicists use the Kelvin temperature scale, which sets 0 K as absolute zero.
- Ratio: values can be added, subtracted, multiplied and divided, in particular, ratios make sense. The classic example is a measurement of weight (e.g., kilograms). The value zero defines the origin of the scale and refers to “none” of whatever the variable describes, be it length, weight or whatever.

Type of Tests

- Parametric test: a test where a model of the data distribution is presupposed. For example in Chap. 3, we assumed that the population tree heights are Gaussian distributed. Parametric distributions can typically be described by a small number of parameters (e.g., the mean and the standard deviation for Gaussian distributions).
- A non-parametric test: a test that does not assume any specific distribution. Some non-parametric equivalents of the t-test are discussed below.

4.2 The Standard Approach: Null Hypothesis Testing

In Chap. 3, we explained statistics within the framework of SDT. Here, we describe the classic null hypothesis testing approach with the example of the two-sample t-test.

The steps for making a statistical decision for a two-sample t-test are:

1. State your alternative hypothesis, called H_1, such as Therapy A is better (or different) than Therapy B.
2. Assume the null Hypothesis H_0 is true: there is no difference between Therapy A and B.
3. From your data, compute the standard error:

$$s_{\overline{x}_A - \overline{x}_B} = s\sqrt{2/n}$$

4. Compute the test statistic as in Chap. 3:

$$t = \frac{\overline{x}_A - \overline{x}_B}{s_{\overline{x}_A - \overline{x}_B}}$$

and the corresponding p-value.

5. Make your decision. If $p \leq 0.05$, reject the H_0 hypothesis and accept H_1: call the effect significant. If $p > 0.05$, you cannot make a statement; in particular do not conclude that H_0 is true.

The above approach has the nice characteristic of setting a limit on the probability of making a Type I error (false positive). Suppose that the null hypothesis is actually true; meaning that we actually draw our samples from the noise-alone distribution. If we now draw many samples from the noise-alone distribution, we will find, on average, that p is less than 0.05 only 5% of the time. We could be more stringent and require $p < 0.01$, in which case p is less than 0.01 only 1% of the time. Of course, there is always a trade-off; the more stringent we are, the more Type II errors (misses) we make when the samples are actually from the alternative hypothesis distribution.

4.3 Other t-Tests

4.3.1 One-Sample t-Test

Sometimes, one wants to compare a single mean with a fixed value. This test is called a one-sample t-test. For example, sometimes a researcher wants to show that a therapy increases IQ, which is 100 on average. We assume that without the therapy the distribution has a mean score of $\mu_0 = 100$. Hence, if the null hypothesis is true and the therapy has no effect, we get the standardized distribution of the IQ in the population. The sampling distribution of the mean has a standard deviation of:

$$s_{\overline{x}} = \frac{s}{\sqrt{n}} \tag{4.1}$$

which is the standard error of the mean. We compute a t-value as:

$$t = \frac{\bar{x} - \mu_0}{s_{\overline{x}}} \tag{4.2}$$

and the degrees of freedom is:

$$df = n - 1 \tag{4.3}$$

With this information we can compute a p-value and make decisions just as for the two-sample t-test.

4.3.2 Dependent Samples t-Test

Often there are two samples of data, but they are related in some specified way. For example, a researcher wants to test whether a therapy increases the level of red blood cells by comparing cell concentration before and after the therapy in the *same* participants. As another example, one might measure preferences for action movies among couples in a relationship. The key characteristic is that every score measured in one sample can be uniquely tied to a score in the other sample. Hence, we can create a *difference score* for each pair. Thus, the two measures of the red blood cell concentration before (x) and after (y) therapy for a given patient would produce a single difference score for that patient:

$$d = y - x \tag{4.4}$$

We now have a single sample of a set of difference scores, and we can run a one-sample t-test on those difference scores just as above. The standard error of the difference scores is:

$$s_{\overline{d}} = \frac{s_d}{\sqrt{n}} \tag{4.5}$$

where s_d is the standard deviation of the sample difference scores d (pay attention to the context of the discussion so as to not confuse this variable with Cohen's d). Like before, we can compare the sample mean to a hypothesized difference of population means (being sure to subtract the population means in the same way that was done for individual scores):

$$t = \frac{\overline{d} - (\mu_y - \mu_x)}{s_{\overline{d}}} \tag{4.6}$$

We compute the p-value in the same way as before by using:

$$df = n - 1 \tag{4.7}$$

This test is also called a repeated measures t-test, a paired t-test, or a within-subjects t-test.

4.3.3 One-Tailed and Two-Tailed Tests

In the above examples, we implicitly considered whether the means of two groups are different from each other ($\mu_A \neq \mu_B$), i.e., we did not specify whether Therapy A is better than Therapy B. These t-tests are called *two-tailed* t-tests, because a large t-value could happen in either tail of the null sampling distribution. One could also pose that if there is a difference then Therapy A is better than Therapy B ($\mu_A > \mu_B$). Alternatively, one could pose that if there is a difference then Therapy B is better than Therapy A ($\mu_B > \mu_A$). In these cases, the t-value can only be in one tail (see Chap. 3, Fig. 3.5). Thus, for a one-tailed test you only have one criterion to satisfy and hence a smaller criterion is required to maintain the desired false positive rate of 0.05. For this reason, the one-tailed test has a higher power than a corresponding two-tailed test.

However, the use of the one-tailed t-test is controversial. In our tree example, the North trees could be either larger or smaller than the South trees. Thus, a two-tailed t-test is more appropriate unless one has a good argument that the North trees are larger than the South trees. One cannot use a one-tailed test when the two-tailed test has led to a non-significant result (see Sect. 11.3.5)! One also cannot decide to use a one-tailed test just because you observe that the data produces one mean larger than the other. Decisions about whether to use a one-tailed or a two-tailed test must be made based on theoretical justification; it cannot be determined by the data or the outcome.

4.4 Assumptions and Violations of the t-Test

As traditionally taught, the main point of the t-test is to control the Type I error rate (false alarm rate). Deviations from the below assumptions almost always alter the corresponding Type I error rate; sometimes in a big way and sometimes in a small way.

4.4.1 The Data Need to be Independent and Identically Distributed

Sampled data need to be Independent and Identically distributed (IID). This is a requirement for many statistical tests. For example, you want to test whether a pain killer reduces not only headaches but also body temperature. You can collect a sample of participants and measure their body temperature before and after the intake of the drug and compute a paired t-test. It is important that a person only participates one time. If you would like to make a hypothesis about the general population, you cannot do the experiment 10 times on 10 days only on yourself because this data is not independent. Maybe you are the only person on the planet for whom the painkiller works.

Here is another example. You are measuring visual acuity at eight locations in the visual field for three patients. Thus, you have 24 data points but they are not independent, so you cannot subject the 24 data points to a t-test. You could average the eight data points for each patient and compute a t-test. Hence, your sample size is only 3, not 24.

Data need to be identically distributed, i.e., they need to come from the same population probability distribution. For example, the height of different types of plants cannot be mixed in a sample. Even if both distributions are Gaussians, the variances may be very different for oaks and edelweiss. If you, for example, measure the heights in a sample of plants collected at both the North and South rims, there might be large differences just because you included more oaks and fewer edelweiss in the North than South sample.

4.4.2 Population Distributions are Gaussian Distributed

The t-test requires that the population distributions are Gaussians[1] or sample sizes are large (often a value of $n = 30$ suffices). However, the t-test is rather robust with respect to populations that are not too different from a Gaussian shape. By robust, we mean that the Type I error rate is close to what is intended (e.g., 5% when using $p < 0.05$ to reject H_0). As long as the distribution is unimodal,[2] even a high amount of skew has only a little effect on the Type I error rate of the t-test (a skewed distribution is not symmetric and has a longer tail on one side of the distribution than the other).

4.4.3 Ratio Scale Dependent Variable

Since the t-test compares means, it requires the dependent variable to be on a ratio scale of measurement. Computing a mean does not make sense for nominal data. Computing variance (or standard deviation) does not make sense for nominal or ordinal data. Since the t-test uses both the sample mean and the sample standard deviation, neither nominal nor ordinal data should be analyzed with a t-test.

There are different opinions about whether a t-test should be used for data on an interval scale. Technically, the properties of the t-test require ratio scale data, but in many cases the t-test behaves rather reasonably for interval data.

[1] Whether the Gaussian assumption is met can be tested by the Kolomogorov-Smirnov test.

[2] A unimodal distribution has only one peak. For example, the Gaussian has only one peak. Bimodal distributions have two peaks.

Table 4.1 Type I error rates for 10,000 simulated t-tests with different population standard deviations and sample sizes

	$n_1 = n_2 = 5$		$n_1 = 5,\ n_2 = 25$	
	$\sigma_2 = 1$	$\sigma_2 = 5$	$\sigma_2 = 1$	$\sigma_2 = 5$
$\sigma_1 = 1$	0.050	0.074	0.052	0.000
$\sigma_1 = 5$	0.073	0.051	0.383	0.047

4.4.4 Equal Population Variances

The standard two-sample t-test assumes that each population has the same variance. Unequal standard deviations, especially combined with unequal sample sizes, can dramatically affect the Type I error rate. Table 4.1 reports the Type I error rate for 10,000 simulated t-tests where the null hypothesis was actually true. For each simulated test, a computer program generated "data" from population distributions and ran the t-test on that generated data. Across different simulations, the *population* standard deviations were either equal (e.g., $\sigma_1 = \sigma_2 = 1$) or unequal (e.g., $\sigma_1 = 5$, $\sigma_2 = 1$) and the sample sizes were either equal (e.g., $n_1 = n_2 = 5$) or unequal (e.g., $n_1 = 5$, $n_2 = 25$).

Table 4.1 demonstrates that if the sample sizes are equal then a difference in the population standard deviations somewhat increases the Type I error rate. Around 7% of the samples rejected the null hypothesis. However, if the samples sizes are unequal and the variances are different, then the Type I error rate is much smaller or larger. When the small sample is paired with the small population standard deviation, then the Type I error rate is much smaller than the intended criterion, 0.05. In this particular set of simulations, not a single t-test rejected the null hypothesis. On the other hand, if the small sample is paired with the large population standard deviation then the Type I error is nearly 40%, which is nearly eight times larger than the intended 5% criterion! The problem is that the default t-test pools the standard deviation from each sample to produce a single estimate of the population standard deviation. If the small sample is paired with the small population standard deviation, then the pooled estimate is too large and the test is unlikely to reject the null. If the small sample is paired with the large population standard deviation, then the pooled estimate is too small and the test is too likely to reject the null.

These problems can be addressed by using a variation of the t-test called the Welch test. However, there is a cost; if the population standard deviations are actually equal, then the Welch test has smaller power than the standard t-test (it is less likely to reject the null hypothesis when there really is a difference).

4.4.5 Fixed Sample Size

Before the experiment, one needs to fix the sample sizes for both groups. One cannot change the sample sizes during the ongoing experiment. This requirement is more difficult

Table 4.2 Parametric tests and their corresponding non-parametric tests

Parametric	Non-parametric
One sample t-test	Sign test
Two-sample t-test	Wilcoxon rank sum test
Repeated measures t-test	Man-Whitney U-test

to satisfy than you might suppose. We discuss a type of violation and its impact in Sect. 10.4.[3]

4.5 The Non-parametric Approach

If your data are not Gaussian distributed you might consider using a non-parametric test. For each t-test described above, there is a non-parametric test, as summarized in Table 4.2.

Non-parametric tests have less power because they cannot exploit a model, i.e., non-parametric tests usually need larger sample sizes for significant results.

The calculations for non-parametric tests look rather different than the calculations of a t-test, however, the non-parametric tests follow the same basic principles of SDT.

Take Home Messages

1. For a t-test, make sure your data are iid distributed and the y-axis is a ratio-scale.
2. Data should be Gaussian distributed or n should be large.

4.6 The Essentials of Statistical Tests

Let us return to the t-test. We had a research question about *mean* differences of trees. Then, we assumed a statistical model, namely, that trees are Gaussian distributed. From the model we derived the equation for the t-value, which is called a *test statistic* and allowed us to compute the p-value and thereby control the Type I error rate. This principle can be applied to many statistical questions. For example, we may ask questions about whether the *variances* of two population distributions differ, the shapes of the population distributions differ (χ^2 test), or the ratio of two means is different from 1 (z-test). More complex tests compute, for example, the means depending on other variables and even

[3] One can fix the sample size n and apply additional criteria such as: the total sample comprises 20 participants, however, if a participant has reduced visual acuity, as determined by an eye test before the experiment, this person can be excluded at this stage and can be replaced by another participant.

more complex tests assume much more complicated models, for example, hierarchical probability distributions.

The principle for all tests is always the same and all cases can be understood in the framework of SDT following the rationale that explains the t-test. The only difference is that a different statistical model than the t-distribution is used. How exactly the test statistics are computed for the various tests is of less importance for understanding statistics because these calculations are done by the computer. For all parametric tests, the p-value confounds effect and sample size.

4.7 What Comes Next?

It is always a good idea to keep our experimental design as simple as possible so you can apply a t-test or a corresponding non-parametric test. However, maximum simplicity is not always possible. For example, we may want to study more than two tree populations, and then a t-test cannot be applied. More variables, e.g., more tree populations, come with a multiple testing problem, which we describe in the next part of this book. The multiple testing problem can be addressed with either statistical methods or clever experimental designs (Chap. 7). We will portray the most common methods because they include an approach that is not evident in the t-test. Although there exist other tests, we do not explain them because this book is about the essentials of statistics and not a compendium.

In Part I of this book, we have laid out many fundamental terms of statistics, as they are needed for statistics users. Readers who are not interested in the specific tests of Part II can proceed directly to Part III, where these key terms are used to explain why we currently have a science and statistics crisis.