

2.3: Chapter 9- Related Samples

Key Terms

difference score

matched pairs

related samples

So far, we have dealt with data measured on a single variable at a single point in time, allowing us to gain an understanding of the logic and process behind statistics and hypothesis testing. Now, we will look at a slightly different type of data that has new information we couldn't get at before: change. Specifically, we will look at how the value of a variable, within people, changes across two time points. This is a very powerful thing to do, and, as we will see shortly, it involves only a very slight addition to our existing process and does not change the mechanics of hypothesis testing or formulas at all!

Change and Differences

Researchers are often interested in change over time. Sometimes we want to see if change occurs naturally, and other times we are hoping for change in response to some manipulation. In each of these cases, we measure a single variable at different times, and what we are looking for is whether or not we get the same score at Time 2 as we did at Time 1. The absolute value of our measurements does not matter—all that matters is the change, or the difference score. Let's look at an example.

Table 9.1 shows scores on a quiz that five employees received before they took a training course and after they took the course. The difference between these scores (i.e., the score after minus the score before) represents improvement in the employees' ability. The third column is what we look at when assessing whether our training was effective. We want to see positive scores, which indicate that the employees' performance went up. What we are not interested in is how good they were before the training or after the training. Notice that the lowest-scoring employee before the training (with a score of 1) improved just as much as the highest scoring employee before the training (with a score of 8), regardless of how far apart they were to begin with. There's also one improvement score of 0, meaning that the training did not help this employee. An important factor in this is that the participants received the same assessment at both time points. To calculate improvement or any other difference score, we must measure only a single variable.

Table 9.1. Raw and difference scores before and after training.

Before	After	Improvement
6	9	3
7	7	0
4	10	6
1	3	2
8	10	2

When looking at change scores like the ones in Table 9.1, we calculate our difference scores by taking the Time 2 score and subtracting the Time 1 score. That is:

$$X_D = X_{T2} - X_{T1}$$

Where X_D is the difference score, X_{T1} is the score on the variable at Time 1, and X_{T2} is the score on the variable at Time 2. The difference score, X_D , will be the data we use to test for improvement or change. We subtract Time 2 minus Time 1 for ease of interpretation; if scores get better, then the difference score will be positive. Similarly, if we're measuring something like reaction time or depression symptoms that we are trying to reduce, then better outcomes (lower scores) will yield negative difference scores.

While we frequently use difference scores for data that are collected from the same participants twice, we can also test to see if people who are matched or paired in some way agree on a specific topic. These are called matched pairs data. For example, we can see if a parent and a child agree on the quality of home life, or we can see if two romantic partners agree on how serious and

committed their relationship is. In these situations, we also subtract one score from the other to get a difference score. This time, however, it doesn't matter which score we subtract from the other because what we are concerned with is the agreement.

In both of these types of data, what we have are multiple scores on a single variable. That is, a single observation or data point is comprised of two measurements that are put together into one difference score. This is what makes the analysis of change unique—our ability to link these measurements in a meaningful way. This type of analysis would not work if we had two separate samples of people that weren't related at the individual level, such as samples of people from different states that we gathered independently. Such datasets and analyses are the subject of Chapter 10.

A Rose by Any Other Name . . .

It is important to point out that the related samplest test has been called many different things by many different people over the years: related samples, paired samples, matched pairs, repeated measures, dependent measures, dependent samples, and many others. What all of these names have in common is that they describe the analysis of two scores that are related in a systematic way within people or within pairs, which is what each of the datasets usable in this analysis have in common. As such, all of these names are equally appropriate, and the choice of which one to use comes down to preference. In this text, we will refer to related samples, though the appearance of any of the other names throughout this chapter should not be taken to refer to a different analysis; they are all the same thing.

Now that we have an understanding of what difference scores are and know how to calculate them, we can use them to test hypotheses. As we will see, this works exactly the same way as testing hypotheses about one sample mean with a t statistic. The only difference is in the format of the null and alternative hypotheses.

Hypotheses of Change and Differences

When we work with difference scores, our research questions have to do with change. Did scores improve? Did symptoms get better? Did prevalence go up or down? Our hypotheses will reflect this. Remember that the null hypothesis is the idea that there is nothing interesting, notable, or impactful represented in our dataset. In a related samples t test, that takes the form of "no change." There is no improvement in scores or decrease in symptoms. Thus, our null hypothesis is:

H_0 : There is no change or difference

$$H_0: \mu_D = 0$$

As with our other null hypotheses, we express the null hypothesis for related samples t tests in both words and mathematical notation. The exact wording of the written-out version should be changed to match whatever research question we are addressing (e.g., "There is no change in ability scores after training"). However, the mathematical version of the null hypothesis is always exactly the same: the average change score is equal to zero. Our population parameter for the average is still μ , but it now has a subscript D to denote the fact that it is the average change score and not the average raw observation before or after our manipulation. Obviously, individual difference scores can go up or down, but the null hypothesis states that these positive or negative change values are just random chance and that the true average change score across all people is 0.

Our alternative hypotheses will also follow the same format that they did before: they can be directional if we suspect a change or difference in a specific direction, or we can use an inequality sign to test for any change:

H_A : There is a change or difference H_A : The average score increases H_A : The average score decreases

$$H_A: \mu_D \neq 0$$

$$H_A: \mu_D > 0$$

$$H_A: \mu_D < 0$$

As before, your choice of which alternative hypothesis to use should be specified before you collect data based on your research question and any evidence you might have that would indicate a specific directional (or non-directional) change.

Critical Values and Decision Criteria

As with before, once we have our hypotheses laid out, we need to find the critical values that will serve as our decision criteria. This step has not changed at all from Chapter 8. Our critical values are based on our level of significance (still usually $\alpha = .05$), the directionality of our test (one-tailed or two-tailed), and the degrees of freedom, which are still calculated as $df = n - 1$. Because this is a t test like the last chapter, we will find our critical values on the same t table using the same process of identifying the correct column based on our significance level and directionality and the correct row based on our degrees of freedom or the next lowest

value if our exact degrees of freedom are not presented. After we calculate our test statistic, our decision criteria are the same as well: $p < \alpha$ or $t_{\text{obt}} > t^*$.

Test Statistic

Our test statistic for our change scores follows exactly the same format as it did for our one-sample t test. In fact, the only difference is in the data that we use. For our change test, we first calculate a difference score as shown above. Then, we use those scores as the raw data in the same mean calculation, standard error formula, and t statistic. Let's look at each of these.

The mean difference score is calculated in the same way as any other mean: sum each of the individual difference scores and divide by the sample size.

$$M_D = \frac{\sum X_D}{n}$$

Here we are using the subscript D to keep track of that fact that these are difference scores instead of raw scores; it has no actual effect on our calculation. Using this, we calculate the standard deviation of the difference scores the same way as well:

$$s_D = \sqrt{\frac{\sum (X_D - M_D)^2}{n - 1}} = \sqrt{\frac{SS}{df}}$$

We will find the numerator, the sum of squares, using the same table format that we learned in Chapter 3. Once we have our standard deviation, we can find the standard error:

$$s_{M_D} = \frac{s_D}{\sqrt{n}}$$

Finally, our test statistic t has the same structure as well:

$$t = \frac{M_D - \mu_D}{s_{M_D}}$$

As we can see, once we calculate our difference scores from our raw measurements, everything else is exactly the same. Let's see an example.

Example A Increasing Satisfaction at Work

Workers at a local company have been complaining that working conditions have gotten very poor, hours are too long, and they don't feel supported by the management. The company hires a consultant to come in and help fix the situation before it gets so bad that the employees start to quit. The consultant first assesses 40 of the employees' level of job satisfaction as part of focus groups used to identify specific changes that might help. The company institutes some of these changes, and six months later the consultant returns to measure job satisfaction again. Knowing that some interventions miss the mark and can actually make things worse, the consultant tests for a difference in either direction (i.e., and increase or a decrease in average job satisfaction) at the $\alpha = .05$ level of significance.

Step 1: State the Hypotheses

First, we state our null and alternative hypotheses:

H_0 : There is no change in average job satisfaction H_A : There is an increase in average job satisfaction

$$H_0: \mu_D = 0$$

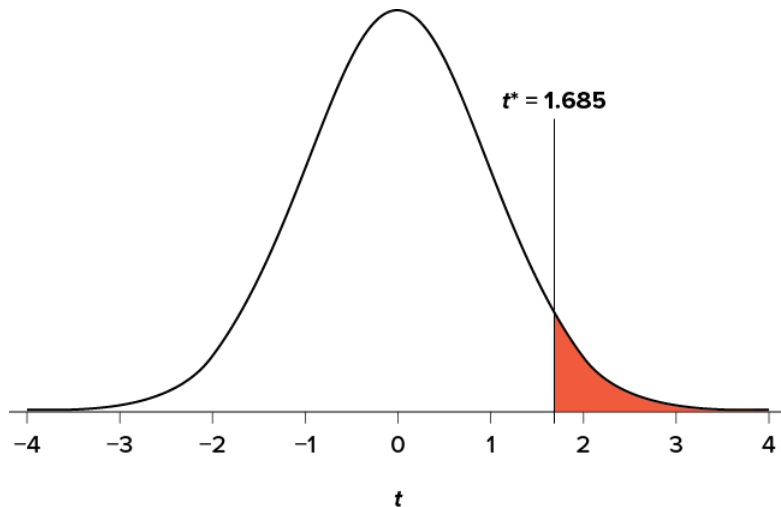
$$H_A: \mu_D > 0$$

In this case, we are hoping that the changes we made will improve employee satisfaction, and, because we based the changes on employee recommendations, we have good reason to believe that they will. Thus, we will use a one-directional alternative hypothesis.

Step 2: Find the Critical Values

Our critical values will once again be based on our level of significance, which we know is $\alpha = .05$, the directionality of our test, which is one-tailed to the right, and our degrees of freedom. For our related-samples t test, the degrees of freedom are still given as $df = n - 1$. For this problem, we have 40 people, so our degrees of freedom are 39. Going to our t table, we find that the critical value is $t^* = 1.685$ as shown in [Figure 9.1](#).

Figure 9.1. Critical region for one-tailed t test at $\alpha = .05$. (“Critical Region $t_{1.685}$ ” by Judy Schmitt is licensed under [CC BY-NC-SA 4.0](#).)



Step 3: Calculate the Test Statistic and Effect Size

Now that the criteria are set, it is time to calculate the test statistic. The data obtained by the consultant found that the difference scores from Time 1 to Time 2 had a mean of $M_D = 2.96$ and a standard deviation of $s_D = 2.85$. Using this information, plus the size of the sample ($n = 40$), we first calculate the standard error:

$$s_{M_D} = \frac{s_D}{\sqrt{n}} = \frac{2.85}{\sqrt{40}} = \frac{2.85}{6.32} = 0.46$$

Now, we can put that value, along with our sample mean and null hypothesis value, into the formula for t and calculate the test statistic:

$$t = \frac{M_D - \mu_D}{s_{M_D}} = \frac{2.96 - 0}{0.46} = 6.43$$

Notice that, because the null hypothesis value of a related samples t test is always 0, we can simply divide our obtained sample mean by the standard error.

Next, we will calculate Cohen’s d as an effect size using the same format as we did for the last t test:

$$d = \frac{M_D - \mu_D}{s_D} = \frac{2.96}{2.85} = 1.04$$

This is a large effect size. Notice again that we can omit the null hypothesis value here because it is always equal to 0.

Step 4: Make the Decision

We have obtained a test statistic of $t = 6.43$ that we can compare to our previously established critical value of $t^* = 1.685$. The number 6.43 is larger than 1.685, so $t > t^*$ and we reject the null hypothesis:

Reject H_0 . Based on the sample data from 40 workers, we can say that the intervention statistically significantly improved job satisfaction ($M_D = 2.96$, $SDD = 2.85$) among the workers, $t(39) = 6.43$, $p < .05$, $d = 1.04$.

Hopefully, the above example made it clear that running a related samples t test to look for differences before and after some treatment works exactly the same way as a regular one-sample t test does, which was just a small change in how z tests were performed in [Chapter 7](#). At this point, this process should feel familiar, and we will continue to make small adjustments to this familiar process as we encounter new types of data to test new types of research questions.

Example B Bad Press

Let's say that a bank wants to make sure that their new commercial will make them look good to the public, so they recruit 7 people to view the commercial as a focus group. The focus group members fill out a short questionnaire about how they view the company, then watch the commercial and fill out the same questionnaire a second time. The bank really wants to find significant results, so they test for a change at $\alpha = .10$. However, they use a two-tailed test since they know that past commercials have not gone over well with the public, and they want to make sure the new one does not backfire. They decide to test their hypothesis using a confidence interval to see just how spread-out the opinions are. As we will see, confidence intervals work the same way as they did before, just like with the test statistic.

Step 1: State the Hypotheses

As always, we start with hypotheses:

$$H_0: \text{There is no change in how people view the bank} \quad H_A: \text{There is a change in how people view the bank}$$

$$H_0: \mu_D = 0 \quad H_A: \mu_D \neq 0$$

Step 2: Find the Critical Values

Just like with our regular hypothesis testing procedure, we will need critical values from the appropriate level of significance and degrees of freedom in order to form our confidence interval. Because we have 7 participants, our degrees of freedom are $df = 6$. From our t table, we find that the critical value corresponding to this df at this level of significance is $t^* = 1.943$.

Step 3: Calculate the Confidence Interval

The data collected before (Time 1) and after (Time 2) the participants viewed the commercial is presented in [Table 9.2](#). In order to build our confidence interval, we will first have to calculate the mean and standard deviation of the difference scores, which are also in [Table 9.2](#). As a reminder, the difference scores are calculated as $\text{Time 2} - \text{Time 1}$.

Table 9.2. Opinions about the bank.

Time 1	Time 2	XD
3	2	-1
3	6	3
5	3	-2
8	4	-4
3	9	6
1	2	1
4	5	1

The mean of the difference scores is:

$$M_D = \frac{\sum X_D}{n} = \frac{4}{7} = 0.57$$

The standard deviation will be solved by first using the sum of squares table:

XD	XD - MD	(XD - MD) ²
-1	-1.57	2.46

XD	XD - MD	(XD - MD) ²
3	2.43	5.90
-2	-2.57	6.60
-4	-4.57	20.88
6	5.43	29.48
1	0.43	0.18
1	0.43	0.18
$\Sigma = 4$	$\Sigma = 0$	$\Sigma = 65.68$

$$s_D = \sqrt{\frac{SS}{df}} = \sqrt{\frac{65.68}{6}} = \sqrt{10.95} = 3.31$$

Finally, we find the standard error:

$$s_{M_D} = \frac{s_D}{\sqrt{n}} = \frac{3.31}{\sqrt{7}} = 1.25$$

We now have all the pieces needed to compute our confidence interval:

$$90\% \text{ CI} = M_D \pm t^*(s_{M_D})$$

$$= 0.57 \pm 1.943(1.25)$$

$$\text{Upper bound} = 0.57 + 1.943(1.25)$$

$$= 0.57 + 2.43$$

$$= 3.00$$

$$\text{Lower bound} = 0.57 - 1.943(1.25)$$

$$= 0.57 - 2.43$$

$$= -1.86$$

$$90\% \text{ CI} = (-1.86, 3.00)$$

Step 4: Make the Decision

Remember that the confidence interval represents a range of values that seem plausible or reasonable based on our observed data. The interval spans -1.86 to 3.00, which includes 0, our null hypothesis value. Because the null hypothesis value is in the interval, it is considered a reasonable value, and because it is a reasonable value, we have no evidence against it. We fail to reject the null hypothesis.

Fail to reject H_0 . Based on our focus group of 7 people, we cannot say that the average change in opinion ($M_D = 0.57$, $SDD = 3.31$) was any better or worse after viewing the commercial, 90% CI: (-1.86, 3.00).

As with before, we only report the confidence interval to indicate how we performed the test.

Exercises

1. What is the difference between a one-sample t test and a related-samples t test? How are they alike?
2. Name three research questions that could be addressed using a related-samples t test.
3. What are difference scores and why do we calculate them?
4. Why is the null hypothesis for a related-samples t test always $\mu_D = 0$?
5. A researcher is interested in testing whether explaining the processes of statistics helps increase trust in computer algorithms. He wants to test for a difference at the $\alpha = .05$ level and knows that some people may trust the algorithms less after the training, so he uses a two-tailed test. He gathers pre-post data from 35 people and finds that the average difference score is $M_D = 12.10$ with a standard deviation of $sD = 17.39$. Conduct a hypothesis test to answer the research question.
6. Decide whether you would reject or fail to reject the null hypothesis in the following situations:
 1. $M_D = 3.50$, $sD = 1.10$, $n = 12$, $\alpha = .05$, two-tailed test

2. 95% CI = (0.20, 1.85)

3. $t = 2.98$, $t^* = -2.36$, one-tailed test to the left

4. 90% CI = (-1.12, 4.36)

7. Calculate difference scores for the following data:

Time 1	Time 2	XD
61	83	
75	89	
91	98	
83	92	
74	80	
82	88	
98	98	
82	77	
69	88	
76	79	
91	91	
70	80	

8. Researchers investigated the extent to which Japanese immigrant mothers encouraged their infants to interact with them or with objects in their environment (such as toys). The researchers observed 8 mothers with their infants. Data are the number of times that mothers encouraged their infants to engage with them or with objects during the observation period. Test the hypothesis at the $\alpha = .05$ level using the four-step hypothesis testing procedure.

With Mom	With Objects
15	40
12	47
14	32
10	50
20	35
28	45
12	42
15	32

9. Construct confidence intervals from a mean of $M_D = 1.25$, standard error of $s_{M_D} = 0.45$, and $df = 10$ at the 90%, 95%, and 99% confidence levels. Describe what happens as confidence changes and whether to reject H_0 .

10. A professor wants to see how much students learn over the course of a semester. A pre-test is given before the class begins to see what students know ahead of time, and the same test is given at the end of the semester to see what students know at the end. The data are below. Test for an improvement at the $\alpha = .05$ level. Did scores increase? How much did scores increase?

Pre-test	Post-test	XD
90	89	
60	66	

Pre-test	Post-test	XD
95	99	
93	91	
95	100	
67	64	
89	91	
90	95	
94	95	
83	89	
75	82	
87	92	
82	83	
82	85	
88	93	
66	69	
90	90	
93	100	
86	95	
91	96	

Answers to Odd-Numbered Exercises

1)

A one-sample t test uses raw scores to compare an average to a specific value. A related-samples t test uses two raw scores from each person to calculate difference scores and test for an average difference score that is equal to zero. The calculations, steps, and interpretation are exactly the same for each.

3)

Difference scores indicate change or discrepancy relative to a single person or pair of people. We calculate them to eliminate individual differences in our study of change or agreement.

5)

Step 1: $H_0: \mu = 0$ "The average change in trust of algorithms is 0," $H_A: \mu \neq 0$ "People's opinions of how much they trust algorithms changes."

Step 2: Two-tailed test, $df = 34$, $t^* = 2.032$

Step 3: $M_D = 12.10$, $s_{M_D} = 2.94$, $t = 4.12$

Step 4: $t > t^*$, Reject H_0 . Based on opinions from 35 people, we can conclude that people trust algorithms more ($M_D = 12.10$) after learning statistics, $t(34) = 4.12$, $p < .05$, $d = 0.70$, and the effect was moderate to large.

7)

Time 1	Time 2	XD
61	83	22

Time 1	Time 2	XD
75	89	14
91	98	7
83	92	9
74	80	6
82	88	6
98	98	0
82	77	-5
69	88	19
76	79	3
91	91	0
70	80	10

9)

At the 90% confidence level, $t^* = 1.812$ and $CI = (0.43, 2.07)$ so we reject H_0 . At the 95% confidence level, $t^* = 2.228$ and $CI = (0.25, 2.25)$ so we reject H_0 . At the 99% confidence level, $t^* = 3.169$ and $CI = (-0.18, 2.68)$ so we fail to reject H_0 . As the confidence level goes up, our interval gets wider (which is why we have higher confidence), and eventually we do not reject the null hypothesis because the interval is so wide that it contains 0.

This page titled [2.3: Chapter 9- Related Samples](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Linda R. Cote](#), [Rupa G. Gordon](#), [Chrislyn E. Randell](#), [Judy Schmitt](#), and [Helena Marvin](#).