Working with *t* Statistics

One of the themes of statistics is that there are some numbers that we never get to know. This is true for μ: we can estimate it with a confidence interval, but there's always uncertainty. We can test it, but with Type I and Type II errors, we can never be certain of the right answer.

When working with a proportion, with an estimate of p that we trust, we can calculate σ, the population standard deviation for a given study, and so we can use the idealized normal curve, the Z curve, to perform calculations. In most real-world situations involving quantitative data, it isn't reasonable to assume σ can be determined. Some things, like an IQ test that's been tried and calibrated on thousands of subjects, do have known standard deviations, but in most cases σ is as hard to pin down as μ.

If we want to get more realistic results, we will have to rely on the standard deviation for the sample, s, which we *can* know. The calculation is very similar to the z-score calculation, but it uses s. It's called a **one-sample t statistic**:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

The big difference between z scores and t statistics is the table used to convert them into probabilities. Before we look at that table, let's examine another detail of z score calculations.

THE ACCURACY OF P-VALUES AND TABLE A-3

At the end of a z-score calculation in a test or hypotheses for a proportion, we use the table to tell us a P-value. Then what do we do with it? We compare it to 0.05 (or 0.10, or…) and make our decision based on that. We don't use the number that the table provides for us. In fact, if the table would simply tell us whether the result was higher or lower than 0.05 (or 0.10, or…), that would be enough.

The table for t statistics works that way: it doesn't give you a specific value for P, but it does allow you to determine what range the P-value falls into, and this is enough information to decide whether P is higher or lower than α.

The table also requires the **degrees of freedom** for the study. For a one-sample t statistic, this is simply one less than the sample size, n − 1. (It's not important to know what "degrees of freedom" really means. For right now, it's enough to understand that it's a measure of how variable the observations in the sample could be.) The degrees of freedom change the shape of the probability density curve for t, so the areas under the curve change slightly too. Providing the detail that Table A-2 allows for t statistics would mean printing a two-page table for every row in Table A-3. It's just not necessary.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

Exercise 1: A physical therapist is testing a new exercise to see if it reduces recovery time after a back injury. Typically, recovery takes 7.2 weeks. With the new exercise, the therapist has assisted 26 patients, and their recovery times had an approximately normal distribution with a mean of 6.8 weeks and a standard deviation of 1.2 weeks. The therapist wants to test the claim that the new exercise reduces recovery time.

(a) If σ is approximated with the sample s.d., are the results statistically significant?

(b) If the therapist admits that σ is unknown, and applies t statistics instead, are the results statistically significant?

Solution: (a) This is a simple case for a test of significance using z-scores.

$$H_0 : \mu = 7.2 \qquad z = \frac{6.8 - 7.2}{1.2/\sqrt{26}} = -1.6997\ldots$$
$$H_a : \mu < 7.2$$

If we look up z < −1.70 in Table A-2 in the textbook, we get .0446 for a P-value. This is less than 5%, so the result is statistically significant. (Whether the observed reduction of about 3 days out of 7 weeks with a standard deviation of over a week is *practically* significant is another question.)

(b) If we don't assume that the standard deviation from the sample is the same as σ (and we shouldn't) then we use a t statistic. The null and alternative hypotheses are the same as before, and in fact the calculation for t is the same:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{6.8 - 7.2}{1.2/\sqrt{26}} = -1.6997\ldots$$

Now we need to look this value up in Table A-3 from the textbook. The degrees of freedom for the study will tell us which row in the table to pay attention to. The formula is df = n − 1, so the degrees of freedom is 26 − 1 = 25. (If we didn't have this line in the table, we'd round to a number that is in the table. For df = 123, for example, we'd round to 100.) The values in the 25 row look like this:

	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
	Area in Two Tails				
	0.01	0.02	0.05	0.10	0.20
25	2.787	2.485	2.060	1.708	1.314

The numbers on row 25 are the **critical t-values** for values commonly used as α, which appear as column headings in the top row. So in a right-tailed test, a t of 1.708 with df = 25 would result in a P-value of *exactly* 0.05. For values of t that aren't in the table, we can't determine P exactly, but we can find a range of values that must contain the correct P-value. To do this, find the two adjacent numbers in the row that t falls between.

Our t is −1.6997, which is negative, and we're interested in the left-hand tail of the normal curve. Table A-3 only deals with positive values of t, so we take the absolute value of t: 1.6997. (This has the same effect that subtracting from 1 has for z-scores: it compensates for wanting information about the "other" tail of the curve.) This number falls between 1.314 and 1.708. If we look up from these numbers to the column headings for one tail, we get .10 and .05. The P-value that goes with the t statistic is between 0.05 and 0.10. Whatever the P-value actually is, it must be higher than 0.05. That's enough to make a decision: this result is not significant at the 0.05 level, and does not constitute good evidence that the exercise decreases recovery time.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

COMBINING TWO DATA SETS

In some statistical studies, we're not interested in finding the mean for a particular set of data, but in the differences between two means. A pre-existing claim may not be available to use for a test of hypotheses, and two studies are done — in effect, estimating a mean for a control group as well. There are two methods for doing this, depending on the design of the experiment.

If a study can be done as a matched pairs experiment, we would subtract the observations within the pairs, and do a one-sample t statistic on the differences. Consider a study of a drug to lower cholesterol. We find a group of people with high cholesterol and measure their cholesterol after treatment, but also before treatment, since we have no claim for μ to compare against. Each test subject's levels, before and after, form a matched pair. The differences in their cholesterol levels are calculated, and these numbers form a sample. We perform a test of significance on the differences. Our null hypothesis is that the μ of the differences is 0, since if the drug has no effect the cholesterol levels before and after treatment should be the same.

On the other hand, if the therapist from Example 1 didn't have a mean recovery time for patients, there would be no way to do a matched pairs experiment to see if the exercise is effective. One person's recovery does not affect someone else's, and the therapist can't help his clients recover, then injure them again to see if they recover faster with a new treatment! He will have to do a **two-sample t statistic** calculation instead:

Example 2: The physical therapist used a control group of 38 patients who did not use the new exercise to compare the previous group of 26 patients against. The control group's recovery time averaged 7.4 weeks (s = 0.9). Is there evidence that the new exercise reduces recovery time?

Solution: To compare the two data sets, we need a different form for the t statistic calculation and for our hypotheses. Call the control "Group 1" and the therapists' clients "Group 2". If the exercise is effective, the mean for Group 2 should be lower. If not, the two means should be identical. The hypotheses are:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

The calculation for the two-sample t-statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(7.4 - 6.8) - (0)}{\sqrt{\frac{0.9^2}{38} + \frac{1.2^2}{26}}} = 2.1665\ldots$$

To look this t statistic up in Table A-3, we need the degrees of freedom. If we had stats software, it could tell us what the df for this situation is. Without it, we use the **conservative value** for df, which is one less than the smaller sample size. Here, that's 26 − 1 = 25. Looking at this row in Table A-3, 2.1665 falls between 2.060 and 2.167. The column headings for those numbers are .025 and .02, so the P-value must be below 0.05. This study does constitute good evidence in favour of the new exercise.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

EXERCISES

A. Given the degrees of freedom and t statistic for a right-tailed study, use Table A-3 to determine the range the P-value for the study must lie in. If α is 0.05, are these results significant?

1) df = 15, t = 2.034
2) df = 23, t = 0.774
3) df = 112, t = 1.473
4) df = 72, t = 2.113

B. Given the following values for a right-tailed study, evaluate the one-sample t statistic and whether the results are significant.

1) $n = 52$, $\bar{x} = 102$, $s = 7$, $\mu_0 = 104$, $\alpha = 0.10$
2) $n = 65$, $\bar{x} = 38.4$, $s = 7.8$, $\mu_0 = 36.9$, $\alpha = 0.05$
3) $n = 125$, $\bar{x} = 557$, $s = 26$, $\mu_0 = 561$, $\alpha = 0.01$

C. Each situation below involves comparing the means of two data sets. Decide whether a one-sample or a two-sample t statistic is better suited to the situations.

1) An orange grower is testing a strain of oranges to see if they will grow in a milder climate. She uses the diameters of the fruit as a measure of how well the plants grow.

2) A researcher recommends that office workers suffering from stress change to a four-day work week to relieve symptoms such as high blood pressure. An HR department gets a pool of employees who have complained of stress to determine if they should adapt policy to accommodate a four-day work week.

3) A T-shirt manufacturer wants to determine if bleach weakens the stitching on sleeves. They plan to use a machine which measures how much force is required to tear a sleeve off a shirt on a random selection of shirts from the assembly line.

D. Perform a two-sample t test for the following problems. Use $\alpha = 0.05$:

1) Does sleep deprivation hinder brain function? Two sets of test subjects were given word puzzles to solve. The control group got normal sleep, and their scores were approximately normally distributed ($n = 51$, $\bar{x} = 82$, $s = 11$), as were the scores from the group who got up three hours earlier each day for a week ($n = 51$, $\bar{x} = 76$, $s = 15$).

2) A report claims that women's reaction times behind the wheel are slower than men's by 1.6 seconds. You think this is an overstatement. You test drivers in a simulator to see how fast they press the brake pedal when a sudden obstruction appears. The study involved 43 men and 62 women. The women had a mean reaction time of 7.8 seconds ($s = 1.5$) and men had a mean reaction time of 6.6 seconds ($s = 1.2$).

SOLUTIONS

A: (1) 0.025 – 0.05, yes (2) > 0.10, no (3) 0.05 – 0.10, no (4) 0.01 – 0.025, yes
B: (1) $t \approx -2.060$, significant (2) $t \approx 1.550$, not significant (3) $t \approx -1.720$, not significant
C: (1) two-sample (2) one-sample (before/after) (3) one-sample (left/right sleeve)
D: (1) H_0: $\mu_1 = \mu_2$; H_a: $\mu_1 > \mu_2$, $t \approx 2.304$, df = 50, $0.01 < P < 0.025$, significant
(2) H_0: $\mu_f - \mu_m = 1.6$, H_a: $\mu_f - \mu_m < 1.6$; $t \approx -1.514$; df = 42 (40), $0.05 < P < 0.10$, not significant

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License