CHAPTER SEVEN: HYPOTHESIS TESTING WITH *Z* AND *T* — Test: Tuesday, March 17

Roughly speaking, if a confidence interval is built around an assumed μ and a sample is collected, there is a high probability (equal to the confidence level) that the sample mean will fall within the confidence interval. If the sample mean falls outside the confidence interval, it could be a coincidence, or it could be that the population mean is not the value originally assumed. This general approach, called hypothesis testing, can be applied in many different testing situations, including comparing two populations to see if there is a significant difference between their parameters. It provides the foundation of social science research, but it is not without its drawbacks. In particular, since it is based on probability, any conclusion has a possibility of being incorrect.

7-A Introduction to Hypothesis Testing — Tuesday • 2/17

inferential statistics • hypothesis testing • null hypothesis • alternate hypothesis • left-tailed test • right-tailed test • two-tailed test • reject the null • statistically significant • type I error • type II error

1. State the null and alternate hypothesis for a test of a single mean or proportion.
2. Identify factors leading to uncertainties about conclusions from statistical inference.
3. State the meaning of a type I error and a type II error for a test.

7-B *P* Values — Monday • 2/23

p value • level of significance • alpha • power

1. Calculate *z* or *t* for a sample mean or proportion.
2. Explain the logic of a *p* value in a simple probability situation.
3. Calculate the *p* value of a *z* test of a single mean or proportion, and interpret its meaning.
4. Use a *p* value to make a statistical conclusion about a test.

7-C Critical Values — Thursday • 2/26

critical value • critical region

1. Find a critical value for a *z* test.
2. Find a critical value for a *t* test.
3. Use a critical value to make a statistical conclusion about a test.
4. Report the results of a study using a *p* value range.

7-D Within-Subjects Designs — Monday • 3/2

within-subjects design

1. Do a *t* test of a single mean difference.
2. Identify possible sequence effects with a given within-subjects design.
3. State how a given within-subjects design could make use of counterbalancing, and discuss how effective this would be.
4. Determine whether or not a within-subjects design is appropriate for a test of two means.

7-E Between-Subjects Designs — Thursday • 3/5

between-subjects design • sequence effects • counterbalancing

1. Calculate *t* for a test of two means.
2. Use a `Test` function on the calculator to do a test of two means.
3. Use a `Test` function on the calculator to do a test of two proportions.

7-A Introduction to Hypothesis Testing

The goal of INFERENTIAL Statistics is to identify if a population parameter is different from an established value or from another given parameter. This is called HYPOTHESIS TESTING.

A NULL Hypothesis, abbreviated H_0, states (roughly speaking) that there is nothing to find with respect to the population parameter being studied. For example, "average body temperature upon wakening is not lower than 98.6°" or "boys and girls are equally likely to apply for college" are null hypotheses because they are stating that there is no difference to be found.

An ALTERNATE Hypothesis, abbreviated H_1, states that there is a difference to be found in the population. This is the prediction of the researcher.

A ONE-Tailed H_1 can be either RIGHT-Tailed or Left-Tailed. It states that a given population parameter is different from a specified value or another given parameter in a specified direction. There are two possible outcomes from a one-tailed test: Either do or do not conclude that the population parameter is different from the specified value or other parameter in the predicted direction (higher for right-tailed, or lower for left-tailed). For example, a one-tailed test to see if women are heavier than men on average would not find that men are heavier than women on average, even though this is correct, because this was not being tested.

A TWO-Tailed H_1 states that a given population parameter is different from a specified value or another given parameter in either direction. There are three possible outcomes from a two-tailed test: Conclude that the population parameter is higher than the specified value or other parameter, conclude that it is lower, or do not conclude that it is different. Two-tailed tests are not ideal in scientific research because they require the researcher to simultaneously have two opposite predictions, and also because they are less powerful.

❶ State the null and alternate hypothesis for a test of a single mean or proportion.

1. The alternate hypothesis is the prediction.
2. Identify which tail(s) are to be used for the test:

   A left-tailed test is used when there is a reason to believe that the parameter being studied is lower than what it is being compared to.

   A right-tailed test is used when there is a reason to believe that the parameter being studied is higher than what it is being compared to.

   A two-tailed test is used when there is a reason to believe that the parameter being studied is lower than what it is being compared to, and there is also a reason to believe that it is higher. Two-tailed test is essentially two tests in one, and they are not as common as one-tailed tests.
3. The null hypothesis is the complement of the alternate hypothesis. It can typically can be stated by adding "not" to the alternate hypothesis.

❶ The average birthweight of a certain breed of cattle is known to be 29 kg. Reed is testing whether feeding the mother a protein-rich diet before birth causes increased birth weights.

1. H_1: The average birthweight of calves born to cows fed protein-rich diets is above 29 kg.
2. This is a right-tailed test because he is predicting an increase.
3. H_0: The average birthweight of calves born to cows fed protein-rich diets is not above 29 kg.

By definition (see 1-A), statistics are values calculated from samples, not from populations. Therefore, statistics cannot *prove* anything about a population. Any findings could be coincidental. It is invalid and misleading to use the word *prove* in a statistical conclusion.

❷ Identify factors leading to uncertainties about conclusions from statistical inference.

1. If there is not much difference in the results, we have little evidence.
2. If the sample size is small, we cannot rely on the law of large numbers.
3. If the standard deviation is high, this indicates that other samples may be very different from the one collected.

❷ Reed finds the weights (in kg) of six newborn calves with mothers fed protein-rich diets: 32.5, 30.1, 28.1, 27.4, 29.9, 29.6. The average weight, 29.6 kg, is more than 29 kg. Why should he be cautious in concluding that protein-rich diets increase birthweight?

1. $\bar{x} = 29.6$ is not much higher than $\mu = 29$ and could easily be a coincidence.
2. $n = 6$ is a small sample size, and he cannot rely on those six cattle being representative of the whole population.
3. $s = 1.8$ kg shows a lot of variation between the cattle, indicating that other samples could be much lower than $\bar{x} = 29.6$ kg.

In hypothesis testing, researchers can either REJECT or Fail to Reject the Null hypothesis. If the null hypothesis is rejected, the data are considered STATISTICALLY SIGNIFICANT. Failing to reject the null hypothesis does not mean that it is accepted; it simply means there is not enough evidence to claim otherwise. It is invalid and misleading to use the word *accept* in a statistical conclusion. For example, Mark hypothesizes that kids prefer ice cream over carrots. If he asks three kids their preference and two say ice cream, then the following are all valid:

- More kids said ice cream than carrots, so the sample results support the hypothesis.
- With only three kids in the sample, the sample results could have easily been a coincidence. They do not provide sufficient evidence for Mark to conclude that kids prefer ice cream over carrots.
- The fact that his sample did not provide enough evidence to conclude that kids prefer ice cream over carrots does *not* mean that he has any evidence that kids do *not* prefer ice cream over carrots.

Researchers never know whether their statistical conclusion is in fact correct. If this were known, the experiment would not have been conducted in the first place.

A TYPE I ERROR is rejecting the null hypothesis when it is actually true.

A TYPE II ERROR is not rejecting the null hypothesis when it is actually false.

❸ State the meaning of a type I error and a type II error for a test.

1. In a type I error, researchers conclude their alternate hypothesis is correct, but actually it is not. This results in the researchers reporting a result that was actually a coincidence, so they are making a false claim.
2. In a type II error, researchers do not conclude their alternate hypothesis is correct, but actually it is. This results in the researchers abandoning their claim rather than reporting it.

❸ State possible statistical conclusion errors for Reed's cows.

1. Type I error: Protein-rich diets do not increase cattle birth weights but Reed falsely claims that they do cause an increase.
2. Type II error: Protein-rich diets do increase cattle birth weights but Reed fails to claim that they cause an increase.

7-B *P* Values

As in previous chapters, the *z* or *t* score of a sample mean is the number of standard errors it is above the mean: $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ or $t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$.

A *z* score can also be calculated for a sample proportion: $z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$.

❶ Calculate *z* or *t* for a sample mean or proportion.

1. For a sample mean $\bar{x}$, calculate *z* or *t* in a mean formula above. Use *t* unless σ is known.
   For a sample proportion $\hat{p}$, calculate *z* in the proportion formula above.

❶ The average weight of a a certain breed of cat is 4.13 kg. In a sample of 20 such cats, $\bar{x} = 3.81$ kg and $s = 0.52$ kg.

$t = \frac{3.81 - 4.13}{0.52 / \sqrt{20}} = -2.75$

❶ 65% of this breed of cat are black. Of the 20 cats in the sample, 14 were black.

$z = \frac{14/20 - .65}{\sqrt{((.65)(.35)/20)}} = 0.47$

The *P* VALUE of a sample statistic is the probability that the sample statistic of another random sample of the same size would support the prediction at least as well, given the null hypothesis is actually true. For example, if the prediction is $\bar{x} > 80$ and $\bar{x}$ turns out to be 93, the *p* value would be $P(\bar{x} > 93)$ using $\mu = 80$.

If *p* is small, it indicates that the null hypothesis is likely to be false. However, *P* values are only meaningful for data that were observed after a justified hypothesis was made. Testing a hypothesis by using the same data that were used to develop the hypothesis in the first place is circular reasoning and frequently leads to faulty conclusions.

❷ Explain the logic of a *p* value in a simple probability situation.

1. Establish a specific prediction, and justify the prediction.
2. If the predicted event does happen, find the probability *p* of the predicted event happening again by chance if tested a second time. In other words, if the justification were meritless but the event coincidentally occurred anyway, how likely is it to coincidentally occur the next time as well?
3. Use the *p* value to explain how confident you are that it was not a coincidence.

❷ Charlie asks Claire to pull a card from a deck. It is the five of hearts, but she puts it back in the deck without showing him. He then shuffles the deck and draws the top card, which he correctly predicts is her card.

1. Charlie predicted that he could find Claire's card. The justification for this prediction is that he implies, in advance, that he has learned how to do this magic trick.
2. $p = \frac{1}{52} \approx 2\%$
3. 2% is very low. It's possible that Charlie was just lucky, but it is likely that he knew what he was doing and could do this trick again.

A low *p* value indicates that the researcher's claim is likely to be true, but it does not necessarily indicate that the researcher's justification for the claim is valid. It is possible that the claim is correct because of confounding variables rather than because of the reasons given by the researcher (see chapter 1).

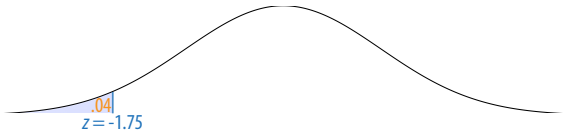
For a *z* test, *p* can be found by calculating *z* and using the methods in chapter 5.

❸ Calculate the *p* value of a *z* test of a single mean or proportion, and interpret its meaning.

1. Identify whether the test is left-tailed, right-tailed, or two-tailed.
2. Calculate *z* (see ❶).
3. Sketch a normal curve, and label the *z* value. If the test is two-tailed, label -*z* as well.
4. Shade from *z* to the left end of the curve for a left-tailed test.
   Shade from *z* to the right end of the curve for a right-tailed test.
   Shade from each *z* outward to the ends of the curve for a two-tailed test.
5. Use the *z* table to find the total shaded area.
6. Given the null hypothesis is true, there is a probability of *p* that results would turn out as far away as they did in the hypothesized direction.

❸ IQ scores are normally distributed with $\mu = 100$ and $\sigma = 15$. Logan hypothesizes that children living within 2000 meters of a coal power plant have lower IQs. In his sample of 120 such children, he finds $\bar{x} = 97.6$.

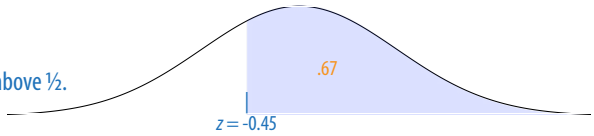
1. $z = \frac{97.6 - 100}{15/\sqrt{120}} = -1.75$
2. The test is left-tailed, because he is testing to see if IQ's are lower than 100.
5. $P(z < -1.75) = 4.0\%$



6. If the average IQ of children living within 2000 meters of the coal plant is really not below 100, there is only a 4% chance that another sample of 120 such children would also have a mean as low as 97.6. Since this is so unlikely, it is likely that the original sample mean was not a coincidence, and that the average IQ of children living within 2000 meters of the coal plant really is below 100.

❸ Caden hypothesizes that Measure B will pass. (In other words, H_1 is "the proportion of voters who support Measure B is more than ½.") In a survey of 80 people, 38 say they plan to vote for it, and 42 say they plan to vote against it.

1. $z = \frac{38/80 - ½}{\sqrt{((½)(½)/\sqrt{80})}} = -0.45$
2. The test is right-tailed, because he is seeing if the proportion supporting Measure B is above ½.
5. $P(z > -0.45) = 1 - .326 = 67.4\%$



6. If the population is actually split 50-50 on Measure B, there is a 67% chance than another sample of 80 people would also include at least 38 who support Measure B. Since this could so easily happen by chance, we do not have evidence to conclude that Measure B will pass.

In this sample, the results turned out opposite what was predicted (the majority of people planned to vote against it), making the *p* value over 50%. This is uncommon.

The LEVEL OF SIGNIFICANCE is how low *p* has to be in order to assume that a difference is too big to be a coincidence and that the null hypothesis should be rejected. It is the probability of making a type I error, given the null hypothesis is true. It is labeled as α, the Greek letter ALPHA.

α can be set to any low probability, such as 5% or 10%, but it cannot be set or changed after any sample statistics or data are known. In this class, we will always use the accepted standard for scientific research in most contexts, which is $\alpha = .05$, even when the book says to use a different value.

Based on the above, 5% is the probability that the null will be rejected, given it is true. This is not the same as the probability that the null is true, given it is rejected. Before the test is conducted, it is correct that there is a 5% chance of a type I error if the null is actually true. But if the null is rejected, the chance that this was a type I error is not 5%.

The POWER of a Test is the probability of rejecting a false H_0. The power of a test is calculated based on an estimate of what the real parameter value is, which is beyond the scope of this course. The Greek letter BETA, written β, is the probability of failing to reject H_0, and is equal to 1 minus power.

❹ Use a *p* value to make a statistical conclusion about a test.

1. Choose the test's tail(s) before the data and statistics are known.
2. Calculate *p* (see ❸).
3. If $p < .05$, reject H_0.
4. Interpret your conclusion:

   If H_0 is not rejected, state that you cannot conclude the alternate hypothesis. Do not state anything about the null hypothesis.*

   If H_0 is rejected, conclude the alternate hypothesis, including the direction of the rejection. Keep in mind that this does not mean that the alternate hypothesis has been proven true, but rather that you have sufficient evidence to suppose that is true.

❹ Make a conclusion from Caden's data on Measure B.

1. The test was one-tailed.
2. $p = .674$ (see ❸)
3. $.674 > .05$ so do not reject H_0.
4. We cannot conclude that Measure B will pass.

* Make sure not to conclude that Measure B will not pass. He was doing a one-tailed test, so he cannot conclude anything about the opposite of his prediction.

Hypothesis testing uses statistics (see chapter 1): Sample data are used to draw a conclusion about the population. Therefore, the conclusion of a hypothesis test is always a claim about the population if the null hypothesis is rejected, or the lack of such a claim in the null hypothesis is not rejected.

In scientific journals, such claims about populations are typically implied by making claims about the samples. For example, "Lower-achieving adolescents earned higher GPAs in core classes at the end of ninth grade when assigned to the growth mindset intervention, $n = 6{,}320$, $t = 3.51$, $p = 0.001$." (Yeager et al., 2019, p. 366). The stated claim refers to the 6,320 freshmen in the study, not all freshmen, but because the finding was statistically significant (*p* was below .05), the implication is that the claim applies to freshmen in general.

When referring to samples in a conclusion, the word *significantly* can be included ("earned significantly higher GPAs") if the null is rejected, or it can be omitted since it is implicitly stated by *p* being below .05. But when not rejecting the null, inclusion of the word can be critical ("did not earn significantly higher GPAs"), indicating that the difference in the sample was not large enough to make a claim about the population, rather than there being no difference in the sample at all.

7-C Critical Values

A CRITICAL VALUE, such as z_0 or t_0, is the value that the calculated statistic must reach in order to reject H_0.

For two-tailed tests, the critical values are the same as in a confidence interval with confidence level $c = .95$, because this includes all but 5% of the curve.

For one-tailed tests, the critical value is the same as in a confidence interval with confidence level $c = .90$, because this includes all but 5% of the curve on either side.

As with for confidence intervals, use t for means (unless σ is known), and use z for proportions.

❶ Find a critical value for a z test.

1. Choose the test's tail(s) before the data and statistics are known.
2. For a two-tailed test, $z_0 = \pm 1.96$.
   For a right-tailed test, $z_0 = 1.64$.
   For a left-tailed test, $z_0 = -1.64$.

❷ Find a critical value for a t test.

1. Choose the test's tail(s) before the data and statistics are known.
2. Identify the degrees of freedom for the test.
3. In the t chart, cross reference the degrees of freedom with $\alpha' = .05$ for a one-tailed test or with $\alpha'' = .05$ for a two-tailed test. Make t_0 negative for a left-tailed test, or positive and negative for a two-tailed test.

❷ $n = 20$, two-tailed

$df = 19$, $t_0 = \pm 2.093$

The CRITICAL REGION is the region(s) of the tail(s) that are beyond the critical value(s). It is the region in which the calculated statistic must fall in order for H_0 to be rejected, and it has an area of 5%.

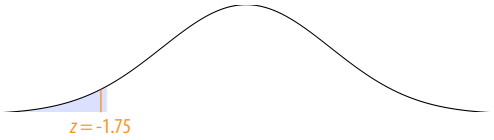
For a two-tailed test, the critical region is the part of the curve not in a 95% confidence interval.

❸ Use a critical value to make a statistical conclusion about a test.

1. Choose the test's tail(s) before the data and statistics are known.
2. Identify the critical value and label it on a curve. For a two-tailed test, put both the positive and negative critical value.
3. Shade from the critical value(s) (see ❶ or ❷) to the end(s) of the tail(s).
4. Calculate z or t (see ❶ in 7-B), and plot it on the curve.
5. For a right-tailed or two-tailed test, if $z > z_0$ or $t > t_0$, then z or t is in the critical region and H_0 is rejected.
   For a left-tailed or two-tailed test, if $z < z_0$ or $t < t_0$, then z or t is in the critical region and H_0 is rejected.
6. Interpret your conclusion.

❸ Make a conclusion about Logan's sample of IQ's of children living near a coal plant in 7-B.

1. The test was left-tailed.
2. $z_0 = -1.64$
1. $z = \frac{97.6 - 100}{15 / \sqrt{120}} = -1.75$
5. $z = -1.75$ is in the critical region because $-1.75 < -1.64$, so reject H_0.
4. We can conclude that the average IQ of children who live within 2000 meters of a coal plant is below 100.



When researchers report their results, they state the calculated statistic with degrees of freedom in parentheses if applicable. They also state the *p* value or, in some cases, one of the following four ranges for it: $p > .05$, $p < .05$, $p < .01$, and $p < .001$. $p > .05$ is sometimes replaced with *ns*, meaning "nonsignificant."

❹ Report the results of a study using a *p* value range.

1. If H_0 is rejected, this means that $p < .05$. In this case, find what the critical value z_0 or t_0 would have been if you had used $\alpha = .01$ instead of $\alpha = .05$. For *z*, this is $z_0 = 2.33$ for one-tailed or $z_0 = \pm 2.58$ for two-tailed. For *t*, this can be found in the *t* table using degrees of freedom and $\alpha' = .01$ for one tailed or $\alpha'' = .01$ for two-tailed. If the calculated value *z* or *t* reaches the new critical value z_0 or t_0, this means you can claim "$p < .01$" rather than "$p < .05$."
2. Repeat step 1 using $\alpha = .001$ instead of $\alpha = .01$. For *z*, the new critical value is $z_0 = 3.09$ for one-tailed or $z_0 = \pm 3.29$ for two-tailed.
3. State the conclusion, in the context of what was being tested, followed by the calculated *z* or *t* value and then by "$p > .05$" if H_0 is not rejected, or by "$p < .05$," "$p < .01$," or "$p < .001$" if H_0 is rejected. If using *t*, include the degrees of freedom in parentheses immediately after *t*.

❹ Giuliana tests 25 200mg Advil tablets in a two-tailed test to see if μ is not in fact equal to 200. The critical value for 24 degrees of freedom is $t_0 = 2.064$. She finds $\bar{x} = 202.9$ with $s = 4.9$, and she calculates $t = \frac{202.9 - 200}{4.9 / \sqrt{25}} = 2.96$. This is greater than $t_0 = 2.064$, so she rejects H_0.

1. $t_0 = 2.797$ for $df = 24$ and $\alpha'' = .01$
   $2.96 > 2.797$, so $p < .01$
2. $t_0 = 3.745$ for $df = 24$ and $\alpha'' = .001$
   $2.96 < 3.745$, so *p* is not less than .001.
3. The average dosage per capsule is more than 200 mg, $t(24) = 2.96$, $p < .01$.

The current convention is to state the exact *p* value, rather than a range, except when the results are nonsignifcant ($p > .05$), extremely significant ($p < .001$), or presented in a table. Older studies will typically report $p < .05$ and $p < .01$ without stating exact *p* values.

7-D Within-Subjects Designs

In a basic BETWEEN-SUBJECTS Design, also known as a Between-Participants Design, each participant takes part in only one of the two conditions. Two separate means (or proportions) are calculated.

In a basic WITHIN-SUBJECTS Design, also known as a Within-Participants Design, each participant takes part in both conditions of the experiment. Only one mean is calculated, the average difference, rather than a mean for each condition.

Some designs use paired data. For between-subjects designs, this means pairing up each participant with their closest match in the sample based on factors relevant to the hypothesis, such as age, sex, or socioeconomic status, and randomly assigning one participant from each pair to each condition. When not using matched pairs, between-subjects designs are sometimes referred to as Independent Samples designs. Within-subjects designs typically involve collecting data from each participant twice, sometimes referred to as a Repeated Measures design, but in some cases each "participant" is actually a pair of participants, such as a husband and wife or an oldest and youngest sibling within a family that each provide data once.

A within-subjects study is tested using a *t* test of a single mean difference. This is identical to a *t* test of a single mean like in 7-C, except that $\bar{x}$ and *s* are calculated from participants' differences, not from their original data. For example, if one participant jumped 218 cm with shoes and 200 cm without, the only value used would be 18, not 218 or 200. The null hypothesis is typically $\mu = 0$, that is, that there is zero difference between the two conditions.

Within-subjects designs are more powerful than between-subjects designs because they reduce the effects of extraneous variables: A participant who is an outlier for an extraneous variable in one condition will likely be an outlier in the other condition as well, making the outlying values cancel each other out. For example, Hazel's data below yield a significant result of $t(6) = 2.43$ when tested in a within-subjects design, but a nonsignificant result of $t(12) = 1.37$ when tested with a between-subjects design (see 7-E ❷).

❶ Do a *t* test of a single mean difference.

1. Subtract the score of each participant in one condition from that same participant's score in the other condition. (In some cases, a "participant" is a pair of people.) Be sure to subtract always in the same direction, even if some differences are negative.
2. Calculate the mean $\bar{x}$ and standard deviation *s* of the differences.
3. Make a statistical conclusion about the differences (see 7-C).

❶ Hazel is testing to see if people can balance better with their eyes open than closed. She has seven participants each see how long they can balance on a balance board with eyes open and also with eyes closed.

Participant #:	1	2	3	4	5	6	7
Eyes open (seconds):	24	50	18	29	24	36	40
Eyes closed (seconds):	13	38	25	19	24	21	31
1. Difference (seconds):	11	12	-7	10	0	15	9

2. $\bar{x} = 7.14$, $s = 7.78$
3. Hazel will use a right-tailed test because she is doing seconds with open eyes minus seconds with closed eyes and she is predicting that this will be positive (that is, that open eyes will have higher times).

   $t(6) = \frac{7.14 - 0}{7.78 / \sqrt{7}} = 2.43$

   $t_0 = 1.943$ for $df = 6$

   $2.43 > 1.943$ so reject H_0

   $t_0 = 3.143$ for $df = 6$ and $\alpha' = .01$

   $2.43 < 3.143$, so *p* is not less than .01

   People can balance longer with their eyes open, $t(6) = 2.43$, $p < .05$.

In some cases, taking part in one condition will influence the results of other conditions. These influences are called SEQUENCE EFFECTS.

❷ Identify possible sequence effects with a given within-subjects design.

1. Participants may tend to improve after the first condition such as due to practice.
2. Participants may tend to do worse after the first condition such as due to fatigue, boredom, or interference.
3. Participants may act differently once they know what is involved in the experiment.
4. Outside factors may change over time.

❷ Caleb is testing if eating a PowerBar 15 minutes before a run increases speed.

1. Participants may do better the second time because they are warmed up.
2. If the second run is done soon afterward, participants may run slower because they are tired.
3. There may be a placebo effect that influences participants to run faster with a PowerBar if they know they are being compared to times without a PowerBar.
4. If the second run is done much later or on a different day, the weather or other factors may be confounds.

Some sequence effects can be canceled by COUNTERBALANCING, in which some participants start with one condition and others start with the other condition. If the order of conditions is randomly assigned, then the within-participant design is a true experiment rather than a quasi-experiment.

❸ State how a given within-subjects design could make use of counterbalancing, and discuss how effective this would be.

1. In a study with two conditions, counterbalancing can be achieved by having half the participants take part in condition A first and the other half take part in condition B first. The order should be assigned randomly in order to make it a true experiment.
2. Weak to moderate order effects due to practice, fatigue, boredom, interference, or outside factors generally will mostly cancel out.

❸ Caleb uses counterbalancing in his PowerBar study.

1. He randomly assigns half of the participants to eat the PowerBar on Monday, and the other half to eat it on Thursday.
2. As long as Monday's and Thursday's conditions are comparable, his counterbalancing should be fairly effective.

❹ Determine whether or not a within-subjects design is appropriate for a test of two means.

1. If participating in the first condition is likely to help the participant do significantly better in the second condition, such as due to practice or finding out about the experiment, do not use a within-subjects design.
2. If participating in the first condition is likely to cause the participant to do significantly worse in the second condition, such as due to tiredness or boredom, do not use a within-subjects design.
3. If neither of the above apply, a within-subjects design should usually be used, especially if counterbalancing and random assignment can be applied.

❹ Is Hazel's balance study better done with a within-subjects design or a between-subjects design?

1. People might do better the second time because they had a little practice, but one quick practice probably will make minimal difference.
2. There is no reason to expect people to do worse the second time.
3. A within-subjects design would be appropriate, especially since it would be easy to counterbalance with randomly assigned order (some people do eyes-open first and some people do eyes-closed first).

7-E Between-Subjects Designs

A two-condition between-subjects study is tested using a *t* test of two means or a *z* test of two proportions.

A right-tailed test of two means tests if μ_1 is greater than μ_2. A left-tailed test of two means tests if μ_1 is less than μ_2. A two-tailed test of two means tests if μ_1 is different from μ_2 in either direction.

For a *t* test of a single mean (see 7-B), $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$, which is the same as $t = \frac{\bar{x}-\mu}{s\sqrt{\frac{1}{n}}}$. For a *t* test of two means, $t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$. This formula requires a single estimate *s* of the population standard deviation, but each sample has its own *s*. To get a single value, use the POOLED Standard Deviation, which is the square root of the total sum of squares divided by total degrees of freedom: $s = \sqrt{\frac{df_1(s_1^2)+df_2(s_2^2)}{df_1+df_2}}$. Alternate methods, not used in this course, do not use a pooled standard deviation.

A pooled standard deviation assumes that the two population variances are equal, but this assumption is unnecessary if the sample sizes are approximately equal.

❶ Calculate *t* for a test of two means.

1. Identify the sample mean, sample standard deviation, sample size, and degrees of freedom for each sample.
2. Calculate the pooled standard deviation *s* (see above).
3. Calculate *t* (see above).

❶ In a sample of nine games, the Falcons had possession of the ball for 38, 30, 31, 41, and 35 minutes at home, and for 32, 28, 25, and 27 minutes away.

1. $\bar{x}_1 = 35, s_1 = 4.64, n_1 = 5, df_1 = 4$
   $\bar{x}_2 = 28, s_2 = 2.94, n_2 = 4, df_2 = 3$
2. $s = \sqrt{\frac{4(4.64)^2 + 3(2.94)^2}{4+3}} = 4.00$
3. $t = \frac{35-28}{\sqrt{\frac{1}{5}+\frac{1}{4}}} = 2.61$

These statistics can also be calculated on a calculator, along with the *p* value.

❷ Use a `Test` function on the calculator to do a test of two means.

1. Choose the test's tail(s) before the data and statistics are known.
2. If you only have the raw data, type them into `L1` and `L2` (see ❺ in 2-B).
3. Push [STAT] and choose `TESTS`.
4. Choose `2-SampZTest...` if both σ's are known. (This is not common.)
   Choose `2-SampTTest...` if one or both σ's are not known.
5. Choose `Data` if you typed in the raw data into lists. Otherwise, choose `Stats`.
6. If entering stats, enter the sample mean `x̄1`, standard deviation `Sx1` or `σ1`, and sample size `n1` for the first sample, and enter `x̄2`, `Sx2` or `σ2`, and `n2` for the second sample.
7. For a left-tailed test, choose `μ1<μ2`.
   For a right-tailed test, choose `μ1>μ2`.
   For a two-tailed test, choose `μ1≠μ2`.
8. For a *t* test, highlight `Yes` for `Pooled` unless one sample size is much larger than the other and you have reason to believe that the two populations do not have approximately equal variance. (This exception is not common and will not apply to this course.)
9. Choose `Calculate`.
10. If $p < .05$, reject H_0.
11. Find the total degrees of freedom: $df = df_1 + df_2$.
12. Write the conclusion in words, followed by *z* or *t*(*df*). If *p* is below .05 and above .001, state *p* rounded to the nearest thousandth. Otherwise, state "$p > .05$" or "$p < .001$."

❷ Hazel is testing to see if people can balance better with their eyes open. Use the data from 7-D.

1. Do a right-tailed test because she is predicting that μ_1 (with eyes open) is greater than μ_2 (with eyes closed).
4. Choose `2-SampTTest...` because the population standard deviations are not known.
7. `μ1>μ2`
9. `t=1.37, p=.098`
10. .098 > .05 so do not reject H_0.
11. $df = 6 + 6 = 12$
12. She cannot conclude that people can balance longer with their eyes open, $t(12) = 1.38, p > .05$.

Between-subjects designs can be done for tests of two proportions, either by hand using $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$, or on the calculator.

❸ Use a `Test` function on the calculator to do a test of two proportions.

1. Choose the test's tail(s) before the data and statistics are known.
2. Push [STAT] and choose `TESTS`.
3. Choose `2-PropZTest`....
4. Enter the first sample's size as `n1` and its number of successes as `x1`. (That is, $\hat{p}_1 = $ `x1` ÷ `n1`.)
5. Enter the other sample's size as `n2` and its number of successes as `x2`. (That is, $\hat{p}_2 = $ `x2` ÷ `n2`.)
6. For a left-tailed test, choose `p1<p2`.
   For a right-tailed test, choose `p1>p2`.
   For a two-tailed test, choose `p1≠p2`.
7. Choose `Calculate`.
8. If $p < .05$, reject H_0.
9. Write the conclusion in words, followed by z. If p is below .05 and above .001, state p rounded to the nearest thousandth. Otherwise, state "$p > .05$" or "$p < .001$."

❸ Mueller & Dweck (1998) had fifth-graders do puzzles. They told all of them that they did well, and they told some of them that they must be smart. Then they gave all of them more difficult puzzles and told them that they did not do well. When asked later about how well they had done, 11 of the 30 who had been told they were smart and 8 of the 58 of the others lied about their score.

1. They did a right-tailed test because they predicted that kids who were told they were smart would be more likely to lie due to feeling that they had to live up to their reputation of being smart.
4. `x1=11, n1=30`
5. `x2=8, n2=58`
6. `p1>p2`
7. `z=2.47, p=.0067`
8. reject H_0
9. Being told they are smart makes kids more likely to lie about their scores, $z = 2.47$, $p = .007$.