CHAPTER EIGHT: HYPOTHESIS TESTING WITH *Z* AND *T*

Review March 22 ~ Test March 31

Roughly speaking, if a confidence interval is built around a given 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.

8-A Introduction to Hypothesis Testing — Monday • 2/27

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 • power

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.

8-B *P* Values — Monday • 3/6

p value • level of significance • alpha

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.

8-C Critical Values — Friday • 3/10

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.

8-D Within-Participants Designs — Wednesday • 3/15

within-participants design

1. Do a statistical test for a within-participants design.

8-E Between-Participants Designs — Monday • 3/20

between-participants design • sequence effects • counterbalancing

1. Use a `Test` function on the calculator to do a statistical test of two means using a between-participants design.
2. Identify possible sequence effects with a given within-participants design.
3. State how a given within-participants design could make use of counterbalancing, and discuss how effective this would be.
4. Determine whether or not a within-participants design is appropriate for a test of two means.
5. Use a `Test` function on the calculator to do a statistical test of two proportions.

8-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 LEFT-Tailed H_1 states that a given population parameter is lower than a specified value or another given parameter.

A RIGHT-Tailed H_1 states that a given population parameter is higher than a specified value or another given parameter.

A TWO-Tailed H_1 states that a given population parameter is different from a specified value or another given parameter.

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

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 the parameter being studied is predicted to be lower than what it is being compared to.
   A right-tailed test is used when the parameter being studied is predicted to be higher than what it is being compared to.
   A two-tailed test is used when the parameter being studied is predicted to be different from what it is being compared to, but there is not a specific prediction as to which direction. Two-tailed tests are not common in scientific research, because they require the researcher to simultaneously have two opposite predictions.
3. The null hypothesis is the complement of the alternate hypothesis, and 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. Ryan 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 can never *prove* anything about populations. Any findings could be coincidental. Do not use the word *prove* with statistics.

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 Ryan 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 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 (in fact much more than the 0.6 kg difference between $\bar{x}$ and μ), indicating that other samples may vary widely from $\bar{x} = 29.6$ 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.

Only the null hypothesis is tested, not the alternate hypothesis. Failing to reject the null hypothesis does not mean the alternate hypothesis is accepted; it simply means there is not enough evidence to claim otherwise. For example, if the first two people you meet in Brazil are very nice, this is not enough information for you to conclude that Brazilians are nicer than Americans, but this lack of information should not be interpreted as indicating that Brazilians are not nicer than Americans.

Researchers never know whether their statistical conclusion is in fact correct. If it 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. The probability of a type I error (given H_0 is true) is called α, the Greek letter ALPHA.

A TYPE II ERROR is not rejecting the null hypothesis when it is actually false. The probability of a type II error (given a specified H_1 is true) is called β, the Greek letter BETA.

The POWER of a Test, $1 - \beta$, is the probability of rejecting H_0 (given a specified H_1 is true).

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

1. In a type I error, researchers believe their alternate hypothesis was correct but actually it is not, causing them to report a result that was actually a coincidence and thus make a false claim.
2. In a type II error, researchers believe their alternate hypothesis was incorrect but actually it is not, causing them to report that their results were coincidental and thus disregard a correct claim.

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

1. Type I error: Protein-rich diets do not increase cattle birth wieghts but Ryan falsely claims that they do cause an increase.
2. Type II error: Protein-rich diets do increase cattle birth wieghts but Ryan does not claim that they cause an increase.

8-B *P* Values

As in chapter 6, 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 Johnston cat is 4.13 kg. In a sample of 20 Johnston 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 Johnston cats 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 a second random sample of the same size would support the prediction at least as well, given the null hypothesis is actually true.

If p is very small, it indicates that the null hypothesis is likely to be untrue.

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

1. Find the probability p of the predicted event happening a second time, given the first time was a coincidence.
2. Based on how low p is, state how confident you are that the prediction coming true was not a coincidence.

❷ Jianna pulls a card from a deck. Nathan guesses that it is the nine of diamonds, and he is correct.

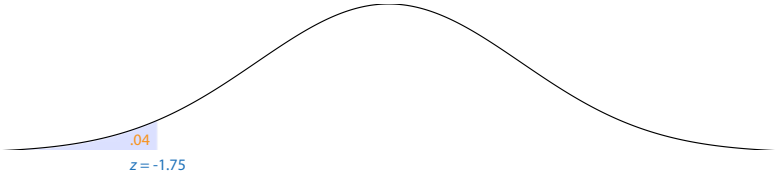
1. $p = \frac{1}{52} \approx 2\%$
2. 2% is very low. It's possible that Nathan just made a lucky guess, but it is likely that he actually knew her card.

For a z test, p can be found by calculating z and using the methods in chapter 6.

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

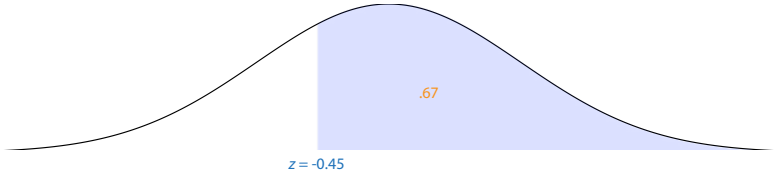
1. Identify whether the test is left-tailed, right-tailed, or two-tailed.
2. Calculate z (see ❶).
3. Sketch a normal curve and label the z value. If the test is two-tailed, label $-z$ as well.
4. Shade from z to the left end of the curve for a left-tailed test.
   Shade from z to the right end of the 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 the results would turn out as far away as they did in the hypothesized direction.

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



1. $z = \frac{97.6 - 100}{15/\sqrt{120}} = -1.75$
2. The test is left-tailed, because she 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 Molly hypothesizes that people prefer Big Macs over Whoppers. (That is, she wants to find out if the proportion of people who prefer Big Macs over Whoppers is more than ½.) In a survey of 80 people, 38 say they prefer Big Macs and 42 say they prefer Whoppers.



1. $z = \frac{{}^{38}/_{80} - {}^{1}/_{2}}{\sqrt{({}^{1}/_{2})({}^{1}/_{2})/80}} = -0.45$
2. The test is right-tailed, because she is seeing if the proportion preferring Big Macs is above ½.
5. $P(z > -0.45) = 1 - .326 = 67.4\%$
6. If people actually do not prefer Big Macs over Whoppers, there is a 67% chance than another sample of 80 people would also include at least 38 who prefer Big Macs.

Note that in this sample the results turned out opposite what was predicted (more people actually preferred Whoppers), making the p value over 50%. This is uncommon.

The LEVEL OF SIGNIFICANCE α 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.

The accepted standard for scientific research is $\alpha = .05$. Always use $\alpha = .05$ in this class.

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. It is correct to say, before a test is conducted, that it has a 5% chance of a type I error if the null is true. It is not correct to conclude that a null that has been rejected has a 5% chance of being a type I error.

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 (but do not state that you "accept" the null hypothesis).
   If H_0 is rejected, state the alternate hypothesis. If it was two-tailed, specifiy the direction of the rejection.

❹ Make a conclusion from Molly's data on Big Macs and Whoppers.

1. The test was two tailed with $\alpha = .05$.
2. $p = .674$ (see ❸)
3. $.674 > .05$ so do not reject H_0.
4. We cannot conclude that more people like Big Macs than Whoppers.

P values are only meaningful for data that were observed after the 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.

8-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 each side.

Recall from Chapter 7 that means use t (unless σ is known) and proportions use z.

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, $z = 1.64$.
   For a left-tailed test, $z = -1.64$.

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.

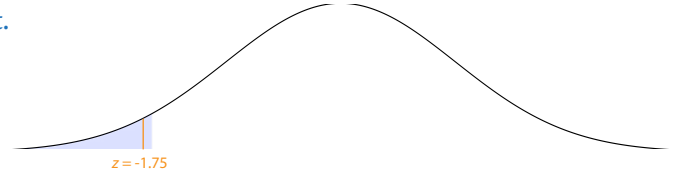
For a two-tailed test, the critical region is the part of the curve not in a 95% confidence interval.

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) to the end(s) of the tail(s).
4. Calculate z or t (see ❶ in 8-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, so reject H_0.
   For a left-tailed or two-tailed test, if $z < z_0$ or $t < t_0$, then z or t will be in the critical region, so reject H_0.
6. Interpret your conclusion.

❸ Make a conclusion about Emma's sample of IQ's of children living near the coal plant.

1. The test was left-tailed.
2. $z_0 = -1.64$
1. $z = \frac{97.6 - 100}{15/\sqrt{120}} = -1.75$
5. -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 the coal plant is below 100.



When researchers report their results, they state the calculated statistic with degrees of freedom in parentheses if applicable. In some cases they do not state the actual p value, but instead state 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 $z_0 = 2.58$ 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 $z_0 = 3.29$ 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 Melina tests 25 200mg Advil tablets in a two-tailed test to see if μ is not in fact equal to 200. The crtical value for 24 degrees of freedom is $t_0 = 2.064$. She finds $\bar{x} = 197.1$ with $s = 4.9$, and she calculates $t = \frac{197.1 - 200}{4.9/\sqrt{25}} = 2.96$. 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 less than 200 mg, $t(24) = 2.96$, $p < .01$.

8-D Within-Participants Designs

In a simple WITHIN-PARTICIPANTS Design, each participant takes part in both conditions of the experiment, making $\bar{x}$ the mean difference between scores. The null hypothesis for within-participants designs is typically $\mu = 0$, that is, that there is zero difference between the two conditions.

❶ Do a statistical test for a within-participants design.

1. Subtract the score of each participant in one condition from that same participant's score in the other condition. 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 8-C).

❶ Julissa 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. Julissa 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, $t(6) = 2.43$, $p < .05$.

8-E Between-Participants Designs

In a simple BETWEEN-PARTICIPANTS Design, each participant takes part in only one of the two conditions. Then two separate means and standard deviations or two separate proportions are calculated.

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 test of two means, $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ or $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$. For a test of two proportions, $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}}$.

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

❶ Use a `Test` function on the calculator to do a statistical test of two means using a between-participants design.

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 ❷ in 3-B).
3. Push [STAT] and choose `TESTS`.
4. Choose `2-SampZTest...` if both σ's are known. This is uncommon.
   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, type the sample mean $\bar{x}$`1`, standard deviation `Sx1` or σ`1`, and sample size `n1` for the first sample, and $\bar{x}$`2`, `Sx2` or σ`2`, and `n2` for the second sample.
7. For a two-tailed test, choose μ`1`$\neq\mu$`2`.
   For a left-tailed test, choose μ`1`$<\mu$`2`.
   For a right-tailed test, choose μ`1`$>\mu$`2`.
8. For a t test, highlight `Yes` for `Pooled` unless the two populations are not likely to have approximately equal variance.
9. Choose `Calculate`.
10. If $p < .05$, reject H_0 and check if p is also less than .01 or .001.
11. Find the total degrees of freedom: $df = df_1 + df_2$.
12. Write the conclusion in words, followed by the calculated statistic and a p range ("$p > .05$," "$p < .05$," "$p < .01$," or "$p < .001$").

❶ Julissa is testing to see if people can balance better with their eyes open. Use the same data as in 9-E.

1. Julissa will use 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. μ`1`$>\mu$`2`
9. `t=1.37, p=.098`
10. .098 > .05 so do not reject H_0.
11. $df = 6 + 6 = 12$
12. Julissa's data do not indicate that people can balance longer with their eyes open, $t(12) = 1.38$, $p > .05$.

Within-participants designs are more powerful than between-participants designs (they are more likely to result in rejecting H_0) 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 conditions as well, making the outlying values cancel each other out. Note that Julissa's data yielded a significant result of $t = 2.43$ when tested in a within participants design, but a nonsignificant result of $t = 1.37$ when tested with a between-participants design.

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

❷ Identify possible order effects with a given within-participants design.

1. Participants may tend to improve after the first condition due to practice.
2. Participants may tend to do worse after the first condition 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 Alondra 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 participants are randomly assigned to the order in which they will take part in the different conditions. Within-participants designs that are not counterbalanced are quasi-experiments rather than true experiments.

❸ State how a given within-participants 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 randomly assigned to take part in condition A first and the other half take part in condition B first.
2. Weak and moderate order effects due to practice, fatigue, boredom, interference, or outside factors will be practically eliminated.

❸ Alondra uses counterbalancing in her PowerBar study.

1. She 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 fairly similar, her counterbalancing should be somewhat effective.

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

1. If participating in the first condition could 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-participants design.
2. If participating in the first condition could cause the participant to do significantly worse in the second condition, such as due to tiredness or boredom, do not use a within-participants design.
3. If neither of the above apply, a within-participants design should usually be used, especially if counterbalancing can be used.

❹ Is Julissa's balance study better off done with a within-participants design or a between-participants 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 think that people would do significantly worse the second time.
3. A within-participants design would be appropriate, especially since it would be easy to counterbalance (some people do eyes-open first and some people do eyes-closed first).

Between-participants designs can also be done for tests of two proportions.

❺ Use a `Test` function on the calculator to do a statistical 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 two-tailed test, choose `p1≠p2`.
   For a left-tailed test, choose `p1<p2`.
   For a right-tailed test, choose `p1>p2`.
7. Choose `Calculate`.
8. The calculator will state the p value. If $p < .05$, reject H_0 and check if p is also less than .01 or .001.
9. Write the conclusion in words.
10. Write the calculated statistic z or t.
11. If p is below .05 but above .001, state p rounded to the nearest thousandth. Otherwise, state "$p > .05$" or "$p < .001$."

❺ 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, $z = 2.47$, $p = .007$.