Web Visualization: Power of a Two-Sample *t*-Test

LEARNING OUTCOMES:

- Explain the concept of power in the context of a two-sample *t*-test.
- Summarize how the power of a two-sample *t*-test is affected by the:
    - difference in means between the two populations;
    - standard deviations of the two populations; and
    - sizes of the samples obtained from the two populations.
- Develop an intuition for why the above three factors affect the difficulty of detecting a difference in the two population means.
- Use an online interactive resource to determine the required sample size for obtaining a desired value of the power.
- Describe the connection between the power and the type I error rate of a hypothesis test.

In this activity, we will develop an intuition about the *power* of a hypothesis test, in the context of a two-sample *t*-test:

Power is the probability of rejecting the null hypothesis when a specific alternative hypothesis is true.

In the context of a two-sample *t*-test, the power is the probability of rejecting the *null hypothesis of equal means*, given that the true means are different. The value of the power depends on the true difference between the means.

You will use an online interactive resource to visualize how the power of a two-sample *t*-test changes with respect to various conditions. The resource can be accessed at

`https://shiney.zoology.ubc.ca/whitlock/RobustnessOfT/.`

This resource was created by Michael Whitlock at the University of British Columbia. It allows you to modify the means of the two populations, and thus the difference of the two means. It also allows you to modify the standard deviations of the two populations, and the size of the sample obtained from each.

For all of the questions below, make sure you've selected the option "Classic 2-sample *t*-test" at the top-left of the resource.

Part I: Introducing the frequency distributions

The frequency curves in the top right of the resource show the distributions of the variable of interest in the two populations.

1. Using the controls on the left side of the resource, set the mean of population 1 to be equal to 10, and the mean of population 2 to be equal to 12. Set both standard deviations to 1. What do you notice about the shape, centre and spread of the frequency curves?

2. Increase both of the standard deviations to 2. How have the frequency curves changed?

3. In which of the two situations, Question 1 or Question 2, do you think the populations look more distinct? In which case do you think the *samples* will look more different?

Part II: Understanding the histogram

The histogram in the bottom right of the resource gives the results of 2,000 simulations. In each of the 2,000 simulations, a sample is taken from population 1, a sample is taken from population 2, and the two-sample *t*-statistic is calculated. The histogram at the bottom is the histogram of the resulting 2,000 two-sample *t*-statistics.

- The light blue part of the histogram shows the *t*-statistics that WOULD NOT lead to rejecting the null hypothesis that both population means are equal.
- The dark blue part of the histogram shows the *t*-statistics that WOULD lead to rejecting the null hypothesis that both population means are equal.

The histogram colours are based on using a two-sided two-sample *t*-test, with a significance level (alpha) set to 5%. The dark blue areas in the histogram correspond to *t*-statistics that are so far from zero that they lead to the rejection of the null hypothesis that the population means are equal.[1]

For the following three questions, set the mean of population 1 to be 10, and the mean of population 2 to be 12. Set both standard deviations to 2.5, and both sample sizes to 25.

1. Based on two samples, each of size 25, we reject the null hypothesis that the two population means are equal if the *t*-statistic is below approximately -2.0 or above approximately 2.0. Look at the histogram at the bottom right of the

[1] CAUTION: Since the histogram is based on simulated data, your histogram will differ slightly from a classmate's, or even from what you might see if you repeated the task. Try this! Make a choice of settings in the sidebar on the left, and look at the histogram. Change the settings, and then return them to their previous values. Is the histogram slightly different than before?

Note that there are exact distributions that can be used in place of these histograms. However, we don't cover this here.

resource; ignore its colours for now. Approximately what proportion of *t*-statistics were below -2.0 or above 2.0?

2. Look at the colours in the histogram. Approximately what proportion of the histogram is dark blue? (Recall that the dark blue part shows the *t*-statistics that lead to rejecting the null hypothesis.)

3. The value of the power is given in red letters at the top of the resource. Relate this value to your answers in Questions 1 and 2. Does this value match what you expected?

4. Describe, in your own words, what power means in the context of a two-sample *t*-test. What does the power value in Question 3 tell you about this specific two-sample *t*-test?

5. Using the controls on the left, set the mean of *both* populations to 10. Answer the following questions.

   a. Looking at the population curves, do you think we should reject the null hypothesis under these conditions?

   b. Looking at the histogram, how often do you think we would reject the null hypothesis under these conditions?

   c. Notice that the text at the top now says “Type I error rate,” rather than “power.” Explain why. *(Hint: How do we define the Type I error rate?)*

Part III: Understanding power

Recall that we are testing the null hypothesis that the two population means are equal. The power, displayed in red letters at the top of the resource, is calculated as the proportion of the simulations that resulted in a rejected null hypothesis. In other words, it is the estimated probability of rejecting the null hypothesis when the two populations and the two sample sizes are as set in the left side-bar.[2]

For the following questions, we will see how the value of the power changes in various scenarios.

1. ***The effect of the difference in means.***

   We will begin by taking a look at what happens to the power for several different values of the two population means. For this question, set both standard deviations to 2, and both sample sizes to 25.

[2] CAUTION: Similarly to the histogram, since the value of the power is based on simulated data, it will also differ slightly from a classmate’s, or even from what you might see if you repeated the task. Try this! Make a choice of settings in the sidebar on the left, and look at the power value. Change the settings, and then return them to their previous values. Is the power value slightly different than before?

a. Set the mean of population 1 to be 10, and the mean of population 2 to be 9. What is the power of the test under these conditions? Looking at the histogram at the bottom, does this value for the power match what you see?

b. If you were to decrease the mean of population 2 to be equal to 8, predict whether the value of the power would increase or decrease. Consider what you think would happen to the frequency curves, the histogram, and the amount of dark blue relative to light blue.

c. Decrease the mean of population 2 to be equal to 8. What do you see happening to the frequency curves, histogram and power value? Explain whether this matches what you expected to see in part b. Is this case more or less likely to reject the null hypothesis, compared to the case in part a?

d. Gradually decrease the mean of population 2 even more, in increments of 0.5. What do you see happening as you do this? Note the frequency curves, histogram, and the power.

e. Based on your results in parts a-d, answer the following TRUE/FALSE questions. (For these questions, assume that everything apart from the two means is fixed.)

   i. The farther apart the population means, the farther the *t*-statistic value is from zero. TRUE/FALSE

   ii. The farther apart the population means, the smaller the proportion of the dark blue in the histogram. TRUE/FALSE

   iii. The farther apart the population means, the more likely it is to reject the null hypothesis that the population means are equal. TRUE/FALSE

   iv. The farther apart the population means, the smaller the power of the *t*-test. TRUE/FALSE

f. What do you predict will happen to the value of the power if the means differ by 2, but take on various different values (e.g. 5 & 7, 8 & 10, 13 & 15, etc.)? *(Multiple choice; circle one.)*

   A) The power will increase as both population means increase.

   B) The power will decrease as both population means increase.

   C) The power will stay approximately the same.

g. Try the following combinations of means for populations 1 and 2: 5 & 7, 8 & 10, 13 & 15, 10 & 8, and 7 & 5. Write down the power of each (rounding to two decimal places). What do you notice about these powers? Are they close, or are they very different? Does this match what you expected to see in part f?

h. Using your results in part g, would you say that the power depends on the individual values of the means themselves, or only on the difference between the means?

2. ***The effect of the population standard deviations.***

   In this question, we will explore the effect of the standard deviations of the two populations on the power. Set the mean of population 1 to be 13 and the mean of population 2 to be 11. Set both samples sizes to 25.

   a. The difference in the two population means is equal to 2. Consider two situations: one where both population standard deviations are large (e.g. 5) and one where both population standard deviations are small (e.g. 1). In which case do you think we are more likely to reject the null hypothesis that the two population means are equal? In which case will the power be larger? *(You may find it helpful to visualize this by considering the frequency curves in the resource for very large and for small values of the population standard deviations.)*

   b. Set both standard deviations to 2. What is the power of the two-sample *t*-test in this scenario?

   c. What do you think will happen to the power as we gradually increase the two standard deviations, in increments of 0.5? Try it. Does your guess agree with the power values that you see?

   d. Based on your results in parts a-c, answer the following TRUE/FALSE questions. (For these questions, assume that everything apart from the two standard deviations is being kept fixed.)

      i. The larger the two population standard deviations, the closer the *t*-statistic is to zero. TRUE/FALSE

      ii. The larger the two population standard deviations, the larger the proportion of the dark blue in the histogram. TRUE/FALSE

      iii. The larger the two population standard deviations, the more likely it is to reject the null hypothesis that the population means are equal. TRUE/FALSE

      iv. The larger the two population standard deviations, the smaller the power of the *t*-test. TRUE/FALSE

3. ***The effect of the sample sizes.***

   How do the sample sizes of the two populations affect the power? To answer this question, begin by setting the mean of population 1 to be 9 and the mean of population 2 to be 7. Set both standard deviations to 5.

   a. Consider two situations: one where both sample sizes are large (e.g. 90) and one where both sample sizes are small (e.g. 15). In which case do you think we are more likely to reject the null hypothesis that the two population means are equal? In which case will the power be larger? Explain your reasoning.

b. Set both sample sizes to 25. What is the power of the two-sample *t*-test in this scenario?

c. What do you think will happen to the power as we gradually increase the two sample sizes? Try it. Does your guess agree with the power values that you see?

d. Based on your results in parts a-c, answer the following TRUE/FALSE questions. (For these questions, assume that everything apart from the two sample sizes is fixed, that the mean of population 1 is still 9, and that the mean of population 2 is still 7.)

    i. The larger the two sample sizes, the farther the *t*-statistic is from zero. TRUE/FALSE

    ii. The larger the two sample sizes, the smaller the proportion of the dark blue in the histogram. TRUE/FALSE

    iii. The larger the two sample sizes, the more likely it is to reject the null hypothesis that the population means are equal. TRUE/ FALSE

    iv. The larger the two sample sizes, the larger the power of the *t*-test. TRUE/FALSE

4. ***Deciding on sample sizes: An application.***

[Motivated by Example 12.3 on pg. 335 of Whitlock & Schluter (2020)[3]; based on a study by Young & Brodie (2004).[4]] The horned lizard *Phrynosoma mcalli* has a fringe of horns surrounding its head. Researchers wish to test whether the horn length makes a difference on whether or not a lizard will be eaten by a predator. To accomplish this, they will compare the horn lengths between a random sample of n_1 lizards that have been found killed by a natural predator (Population 1), and a random sample of n_2 lizards that are still alive (Population 2) in the same area.

The researchers do not want n_1 and n_2 to be too large, since obtaining the observations can be time-consuming. However, they also want n_1 and n_2 to be large enough so that they can detect a difference between the two populations, if a difference exists. Specifically, for a significance level (alpha) taken to be 5%, if the difference in mean horn length is 2 mm, they want to reject the null hypothesis with a probability of at least 0.90.

[3] Whitlock, M. C. & Schluter, D. (2020). The Analysis of Biological Data (3rd ed.). Macmillan.

[4] Young, K. V. & Brodie, E. D. (2004). How the horned lizard got its horns. *Science, 304*(5667), 65-65.

For the questions below, suppose that the biologists know that from previous studies the standard deviations of the horn lengths for both populations are approximately 2.5 mm.

a. State the null and alternative hypotheses in words, *in the context of this study*.

b. What is the value of the power that the researchers wish to obtain, for the case that the difference in mean horn length is 2 mm?

c. Suppose you were to use the online resource, with the true difference in mean horn lengths being 2 mm. State one possible combination of values you could set for "population 1 mean" and "population 2 mean." *(Hint: Do the actual values of the means matter, or is the difference in means the only important factor? Think back to Questions 1g and 1h.)*

d. Suppose that the researchers were to sample 20 lizards killed by a natural predator, and 30 lizards that are still alive. From data collected, do you think this will allow them to detect a difference of 2 mm in population means with a power of at least 0.90? To answer this, use the resource to find the probability that they will reject the null hypothesis that the two means are equal.

e. Suppose that the researchers can only obtain a sample of 20 killed lizards. How many living lizards should they sample, in order to obtain the desired power of at least 0.90? Use the online resource to find your answer.

f. If the researchers only needed a power of at least 0.70, would the sample sizes in part e work for them? Why might the researchers want to re-do their power calculations to get different sample sizes?

g. If, instead, the researchers wanted a power of around 0.90 when the difference of the means is 1 mm instead of 2 mm, should they increase or decrease the sample size of living lizards from the value found in part e? (Suppose the restriction that only 20 killed lizards can be found is still imposed.)