Sample Size Calculations and Review Activity

Please use some color (other than black or yellow) for your answers. THANKS!!
Please place your name BEFORE your answer—for example:
Laura Le: ***This is my answer to this question.***

In the first part of the activity, you will be exploring how research design considerations affect the estimated sample size. Then, you will be applying what you have learned in the first half in this course to three scenarios.

ESTIMATE SAMPLE SIZE

Before researchers begin a study, they should obtain an estimate of the sample size. To obtain this value, one would need to set…

- in R, the power value, the significance level, and the estimated effect size, or
- in SAS, the power value, the significance level, difference in means, and estimated (pooled) standard deviation.

Let's explore how the values that are set by the researchers affect the estimated sample size that is calculated at the beginning of a study.

An initial pilot study was carried out to determine an estimate for the pooled standard deviation. From this pilot study, the researchers obtained an estimate of the pooled standard deviation of 4. The researchers would also like to be able to detect a difference in means of 1.6.

1. For R, what would the effect size be? (SAS users do not need to complete this question.)

2. Using your software of choice, calculate the estimated sample size for this study if the researchers set power = 0.9 and significance level =0.05. Share your code and output.

Let's change the power value and see how that affects the estimated sample size.

3. Using your software of choice, calculate the estimated sample size for this study if the researchers set **power=0.8** and significance level=0.05. Share your code and output.

4. What is the effect on the sample size if the power is decreased?

© 2026 Laura Le. Licensed under CC BY-NC-SA 4.0.

Now let's use the original power value and change the difference in means value (or effect size value) to see how that affects the estimated sample size.

5. Using your software of choice, calculate the estimated sample size for this study if the researchers are looking to detect a **difference in means of 3.2** and set power =0.9 and significance level=0.05. (Note: R users will need to calculate the effect size first. Use the same standard deviation as previously provided.) Share your code and output.

6. What is the effect on the sample size if the effect size or difference in means is increased?

Now let's use the original difference in means of 1.6 and change the significance level value to see how that affects the estimated sample size.

7. Using your software of choice, calculate the estimated sample size for this study if the researchers set power=0.9 and **significance level=0.01**. Share your code and output.

8. What is the effect on the sample size if the significance level is decreased?

REVIEW

Now that you are halfway through the course, let's practice integrating the topics from each of weeks into three different scenarios.

For each scenario, you will be **applying** the research design process that includes:

1. identifying the sampling method and study design,
2. identifying the population of interest,
3. performing exploratory data analysis (i.e., summary statistics and graphical displays),
4. identifying the statistical method of inference,
5. checking assumptions for the inferential method,
6. carrying out the inferential method,
7. making conclusions about the population based on the results from the inferential method, and
8. identifying whether the results can be applied/generalized to the population of interest, given the sample.

SCENARIO #1

In observational studies, maternal periodontal disease has been identified as a risk factor for preterm birth (< 37 weeks) and low birth weight (< 2500 g) outcomes. A multi-center randomized trial, called the Obstetrics and Periodontal Therapy study, evaluated the effect of nonsurgical periodontal treatment intervention on preterm birth, comparing outcomes of women treated before 21-weeks gestation (treatment) to those treated after delivery (control)[1]. The outcomes of interest for the study involved many variables related to maternal and infant health (e.g., gestational age at the end of pregnancy, birthweight, Apgar scores). One of the research questions of interest was:

Research Question: Is there a difference in average birthweight of infants between those in the treatment group (mothers who received periodontal treatments *before 21 weeks gestation*) and those in the control group (mothers who received periodontal treatments *after* delivery)?

The data and a description of the study and variables can be found on the course page, titled *OPT_Study_Person-level_Data.csv*, *OPT_Dataset_Introduction.pdf*, and *OPT_Data_Dictionary.pdf*[2]. [Note: Be sure to refer to these additional documents, the introduction and the dictionary, as you are working on the questions for this scenario.]

The variables of interest for the posed research question are *Group* (for intervention group) and *Birthweight* (for infant birth weight at time of birth, in grams).

9. What sampling method and study design was used for this study? (e.g., Was it observational or experimental? If observational, which design? Random sample? Control group? Randomized? Blinded?)

10. To what population do you think the researchers would like to make inferences?

11. What exploratory data analysis (i.e., summary statistics and graphical displays) techniques would you use to explore the variables of interest? Based on your answer, carry out the analyses using your software of choice. Share your code and output.

12. What statistical method of inference would you use to answer the research question? Explain your reasoning.

---

[1] Michalowicz, B. S., Hodges, J. S., DiAngelis, A. J., Lupo, V. R., Novak, M. J., Ferguson, J. E., ... & Tschida, P. A. (2006). Treatment of periodontal disease and the risk of preterm birth. New England Journal of Medicine, 355(18), 1885-1894.

[2] Retrieved from TSHS Resource Portal. https://www.causeweb.org/tshs/obstetrics-and-periodontal-therapy/

13. Check the assumptions for the inferential method you identified in the previous question. If you use software of choice, share your code and output.

14. Using your software of choice, carry out the inferential method you identified in a previous question. Share your code and output.

15. Using the inferential results, what conclusion can be made about the population?

16. Given the sample, can the results of this study be applied/generalized to the population that you listed in a previous question? Why or why not?

SCENARIO #2

Vaping has become a public health concern, especially among teens. Researchers at the University of Michigan have conducted an annual survey, titled *Monitoring The Future*, to measure drug and alcohol use and related attitudes among students nationwide (e.g., 392 public and private schools in 2018). Of the high school seniors (n=4,839) that received and responded to the vaping question in 2018, 37.3% of them reported vaping within the past 12 months (which, by the way, is up 9% from 2017).

Research Question: What is an estimate of the proportion of high school seniors who reported vaping within the past 12 months in 2018?

17. What sampling method and study design was used for this study? (e.g., Was it observational or experimental? If observational, which design? Random sample? Control group? Randomized? Blinded?)

18. To what population do you think the researchers would *like* to make inferences?

19. What statistical method of inference would you use to answer the research question? Explain your reasoning.

20. Check the assumptions for the inferential method you identified in the previous question. If you use software of choice, share your code and output.

21. Using your software of choice, carry out the inferential method you identified in a previous question. Share your code and output.

22. Using the inferential results, what conclusion can be made about the population?

23. Given the sample, can the results of this study be applied/generalized to the population that you listed in a previous question? Why or why not?

SCENARIO #3

Some swimmers in the 2008 summer Olympics used new swimsuits (wetsuits) that possibly provided unfair advantages, which caused controversy in the sport. Can a swimsuit's design really make a swimmer faster? A study investigated whether wearing wetsuits influences swimming velocity compared to wearing regular bathing suits[3]. Twelve competitive swimmers and triathletes were recruited to swim 1500 m at maximum speed twice each, once wearing a wetsuit and once wearing a regular bathing suit. The order of the trials was randomized. The outcome of interest was the maximum velocity in meters/seconds. The research question of interest was:

Research Question: Is there a mean difference in maximum velocity (in m/s) between wearing a wetsuit and wearing a regular bathing suit for the professional swimmers?

The data[4] can be found on the course page, titled *Wetsuits.csv*.

The variables of interest for the posed research question are *Wetsuit* (for max velocity (in ms/s) when wearing a wetsuit) and *NoWetsuit* (for max velocity (in ms/s) when not wearing a wetsuit).

[3] De Lucas, R. D., Balikian, P., Neiva, C. M., Greco, C. C., & Denadai, B. S. (2000). The effects of wet suits on physiological and biomechanical indices during swimming. *Journal of Science and Medicine in Sport*, *3*(1), 1-8.

[4] Data retrieved from Lock, et al. (2017). Statistics: Unlocking the Power of Data.

24. What sampling method and study design was used for this study? (e.g., Was it observational or experimental? If observational, which design? Random sample? Control group? Randomized? Blinded?)

25. To what population do you think the researchers would like to make inferences?

26. What exploratory data analysis (i.e., summary statistics and graphical displays) techniques would you use to explore the variables of interest? Based on your answer, carry out the analyses using your software of choice. Share your code and output.

27. What statistical method of inference would you use to answer the research question? Explain your reasoning.

28. Check the assumptions for the inferential method you identified in the previous question. If you use software of choice, share your code and output.

29. Using your software of choice, carry out the inferential method you identified in a previous question. Share your code and output.

30. Using the inferential results, what conclusion can be made about the population?

31. Given the sample, can the results of this study be applied/generalized to the population that you listed in a previous question? Why or why not?

EXTENSION

32. Which study design(s) was/were NOT used in one of the three scenarios above? For the particular study design, find a medical or public health article that uses that study design. Provide a reference and abstract of the article and describe how the particular study design was used.