

Name: _____

SID: _____

Collaborators: _____

Week 4 Problem Set

Confidence Intervals PHW142

You must put your name and SID at the top of the page.

Please submit your work in the Week 4 Problem Set dropbox as a PDF file.

Remember you can only upload once, so be sure your work is complete and correct, and that you upload the correct PDF file.

R Markdown files for two of the practice problems for means are in Unit 1 of the Week 4 bCourses site, and are also under the Week 4 More Practice tab. The keys for the practice problems contain examples of the R functions for confidence intervals for proportions.

This problem set is worth 20 points.

Part 1. Vitamin C loss in wheat soy blend

1. **(9 points)** These data come from earlier editions of Introduction to the Practice of Statistics, by David Moore and George McCabe. The data come from studies conducted for the US Agency for International Development (USAID). Wheat-soy blend (WSB) is prepared for emergency food relief. Vitamin C is added to the WSB when it is prepared in the US and shipped overseas for ready availability for disaster relief. There is a concern that vitamin C is lost in shipment and storage.

The researchers selected a simple random sample of 27 bags of freshly-prepared wheat-soy blend at the US preparation site, before shipment to Haiti. They took samples of the WSB, and assessed the Vitamin C content. The bags were specially marked so that they could be resampled 5 months later in Haiti.

The data consist of 27 pairs of measurements. The key variable for analysis is the difference between the factory measurement and the Haiti measurement. We'll create the difference as the first step in our data analysis. Theoretically, the value of Vitamin C in the Haiti sample can't be any larger than the US value. But because the vitamin C isn't completely uniformly mixed in, and there's some measurement error, some of the differences might be negative.

The units on the vitamin C measurements are milligrams of vitamin C per 100 grams of WSB.

The **comma separated values (.csv) file** `wsbhaiti.csv` has the measurements when the bags were freshly prepared and when they were assessed 5 month later in Haiti. The column names are factory and Haiti.

As usual, you will need to download this file to your working directory.

When you use the library function to load the packages you need, be sure to include `tidyverse`, `readxl`, and both `epiR` and `epitools`.

Please use the `read_csv` function to get the csv file into an R data frame named `wsb`.

You will need to create a new variable that is the difference. If the data frame name is `wsb`, here is how to do it. (You will need to type in this line, rather than copying and pasting it.)

Since the sample size is only 27 (the number of bags), we'll check the distribution assumptions carefully. You will discover that there is an outlier in the box plot of the differences. We will proceed with caution and calculate a confidence interval using the t distribution.

1.1 What are the sample average and standard deviation for the vitamin C loss?

1 point

1.2 What is the estimated standard error for the vitamin C loss? (Show details of the calculation.)

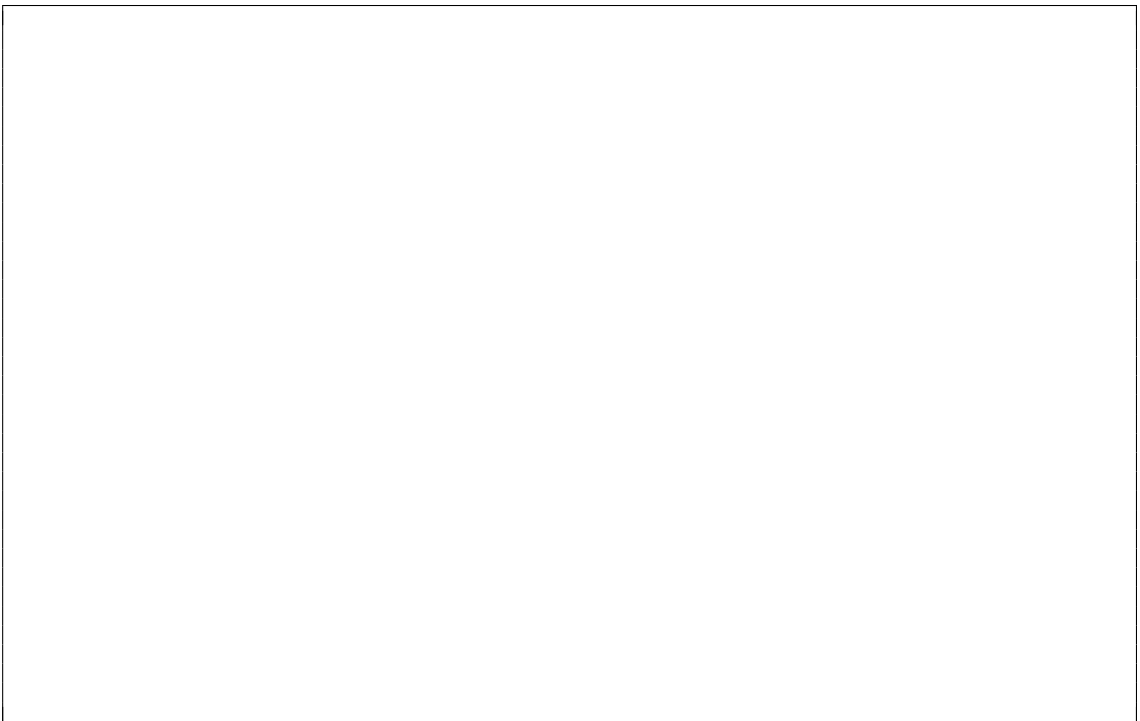
1 point

Include your box plot of the difference values here:



1.3 Use the boxplot to assess the data for the presence of outliers and for symmetry.

Discuss all the details of the distribution. (You do not need to find the values of the fences.) **1 point**



Include your normal quantile plot of the differences here.

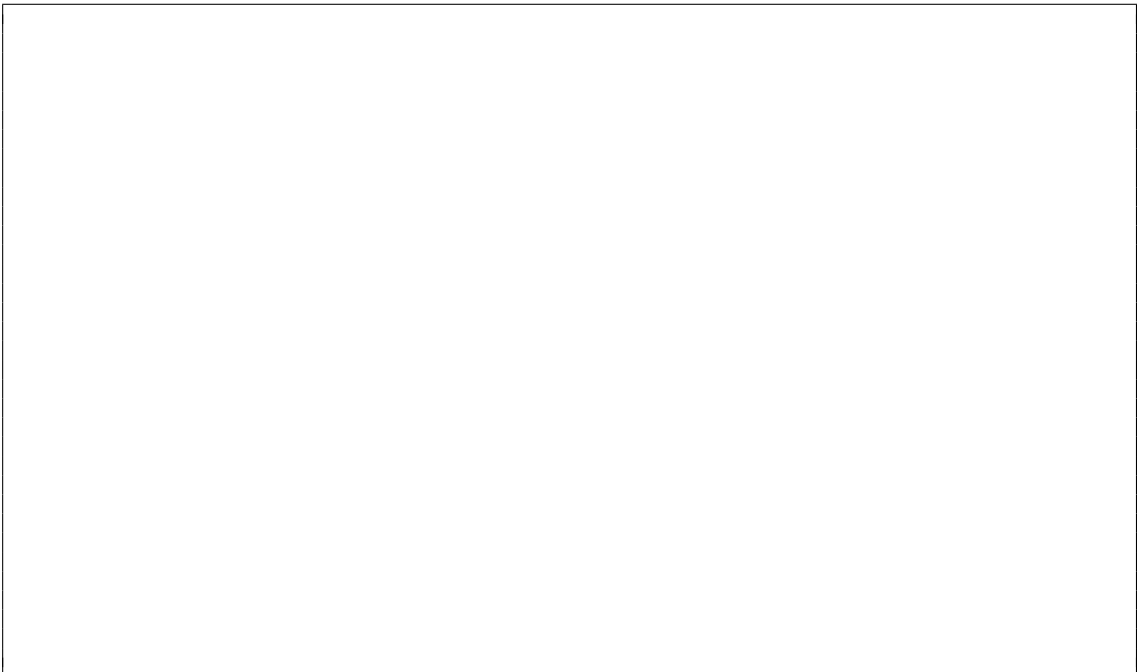


1.4 Interpret the normal quantile plot.

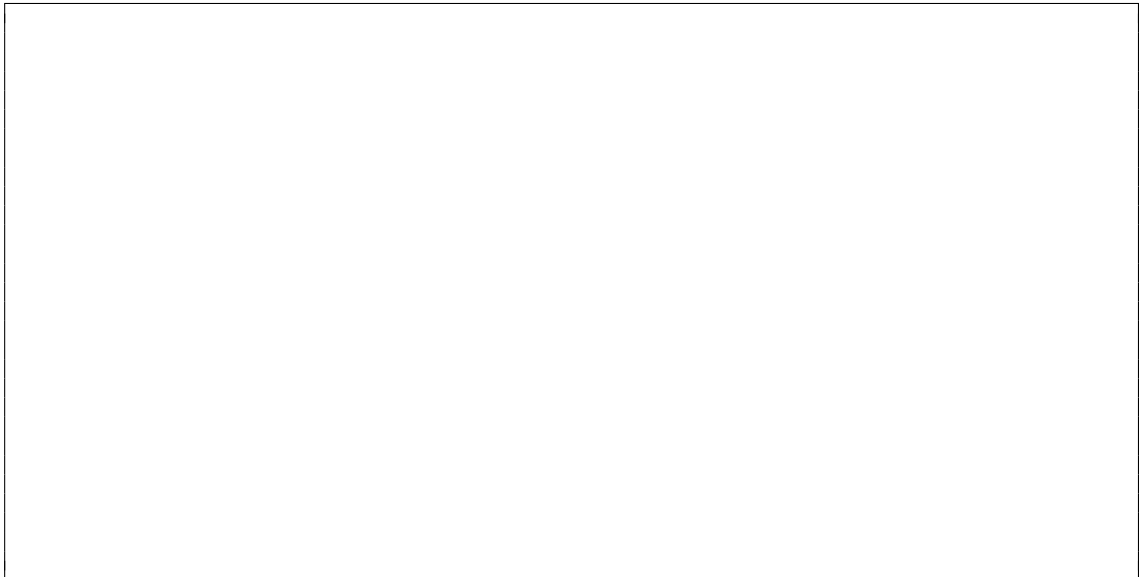
1 point

Do the data contradict the assumption that the population of the differences have a distribution that follows a normal curve?

Discuss the closeness of the points to the "gold standard line" and state an overall conclusion.

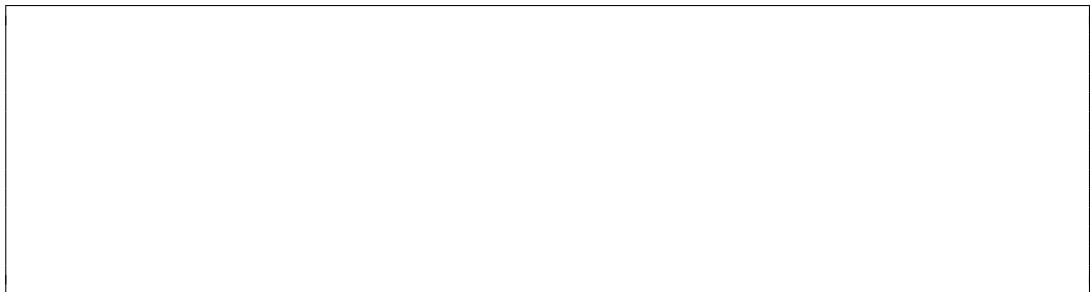


- 1.5 Based on the box plot, the normal quantile plot, and the criteria given by Baldi and Moore for using the t distribution, explain why we should be cautious about using the t distribution to calculate a confidence interval for the population mean vitamin C loss.

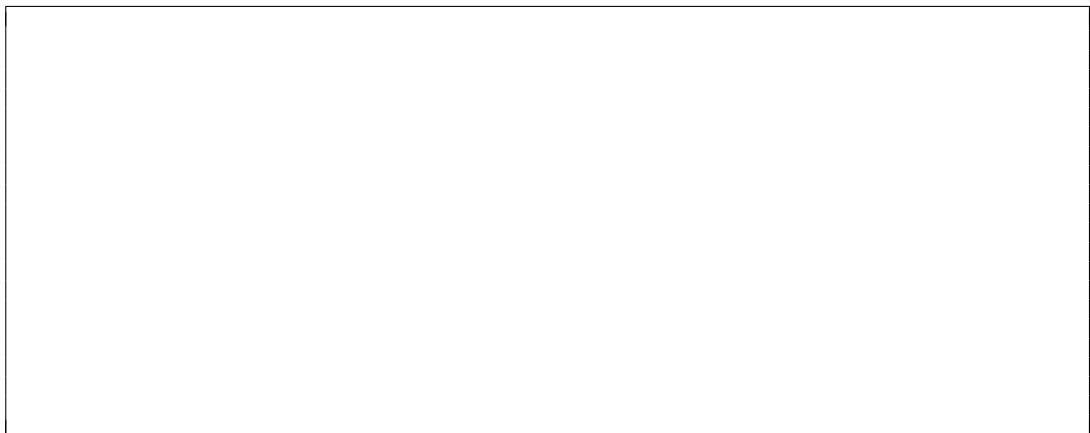
1 point

- 1.6 Complete the following questions to calculate the t-value.

1.6.1. What are the degrees of freedom for the t-distribution?

 $\frac{1}{2}$ point

- 1.6.2. Find the t^* critical value for the 95% confidence interval.
If you use OOMPH Stat, paste a screenshot here.

 $\frac{1}{2}$ point

If you use the `qt` function in R, copy and paste the results here.

1.7 Show the calculation of the margin of error and the 95% confidence interval.

1 point

Use the function `epi.conf` to find the 95% confidence interval and paste your output here.

1.8 Write a short summary that interprets the confidence interval.

1 point

Include an explanation of exactly what a 95% confidence level means.

- 1.9 Does the confidence interval provide evidence that vitamin C is being lost between the preparation of the WSB blend and 5 months later? Explain.

1 point

Part 2. confidence intervals for a proportion

fast food for baby

adapted from exercise 19.31 (3rd edition), page 480

2. **(4 points)** The Gerber Products Company sponsored a large survey of the eating habits of American infants and toddlers. Among the many questions parents were asked was whether their child had eaten fried potatoes on the day before the interview. Among 679 infants 9 to 11 months old, 61 [Baldi and Moore say 9%] had eaten fried potatoes that day.

This problem asks for a 95% confidence interval for a population proportion.

- 2.1 Explain why the conditions to use the normal approximation are satisfied.

1 point

- 2.2 Calculate the standard error for the confidence interval, using $\hat{p} = 0.09$.

1 point

- 2.3 Calculate the 95% confidence interval from "first principles," showing all the details. You already have \hat{p} and the standard error. Now use the OOMPH normal calculator, or the `qnorm` function in R to get the z^* critical value, and calculate the upper and lower confidence limits.

1 point

Use the `binom.approx` function to repeat the calculations for the normal approximation, and paste your results here.

Use `binom.wilson` function with 61 Yes to calculate the 95% Wilson confidence interval and paste your results here:

2.4 Are the intervals very similar, or are there marked differences? Explain.

clinical progression of SARS

adapted from Baldi and Moore exercise 19.10, 3rd edition; 19.7, 4th edition

3. **(2 points)** Severe acute respiratory syndrome (SARS) is a viral respiratory illness that was first reported in Asia and then spread worldwide to near pandemic levels before eventually being contained. A study examined a random sample of 75 patients diagnosed with SARS to describe the disease's clinical progression. One finding was that, while fever and pneumonia initially improved after treatment in all 75 patients, 64 patients developed a recurring fever.

- 3.1 Explain why the conditions to use the normal approximation to find a confidence interval for the population proportion of patients treated for SARS who develop recurring fever are not satisfied.

1 point

Use the function `binom.wilson` find the Wilson 90% confidence interval for the population proportion of patients treated for SARS who develop recurring fever.

Paste your results here:

- 3.2 Write a summary sentence, interpreting the 90% confidence interval for an audience of clinicians who are users rather than creators of statistics.

1 point

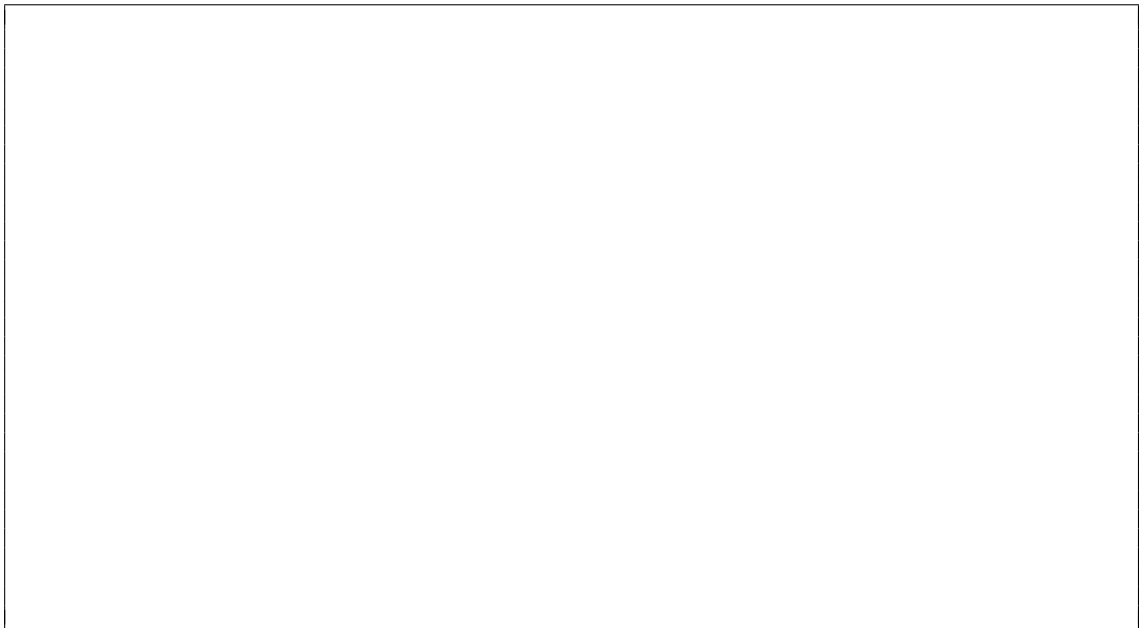
Part 3. Finding sample sizes

**estimating how much a drug will lower heart rate in patients with tachycardia
(fast heart rates)**

4. **(3 points)** Suppose that a pilot study indicated that the standard deviation of the reduction in heart rate with a high dose of the drug propranolol is 10 beats per minute.

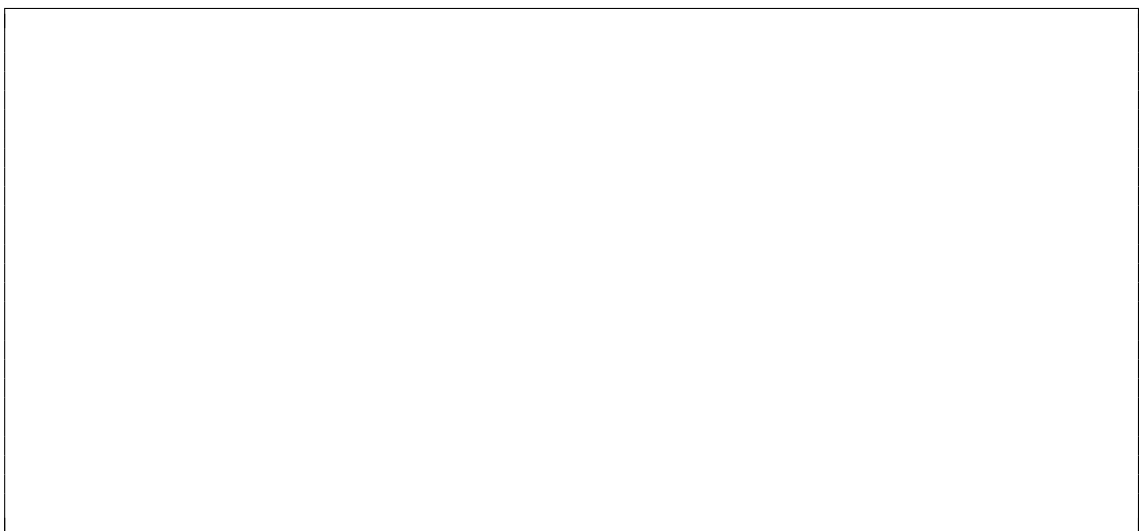
4.1 What size sample would be needed if we want the margin of error for a 95% confidence interval to be 2 beats per minute?

1 point



4.2 What size sample would be needed if we want the margin of error for a 99% confidence interval to be 2 beats per minute?

1 point



4.3 Compare the sample sizes for the 95% and 99% intervals.

1 point

planning a study of influenza

adapted from Baldi and Moore, exercise 19.11 in both editions

5. **(2 points)** According to the CDC, each year in the United States, on average, 5% to 20% of the population get the flu. Using 20% as our best guess of the population proportion, how large a sample would be needed to get a margin of error of ± 2 percent for a 95% confidence interval for the population percent of Americans who get the flu in a given year?

2 points