

STAT 477/577 - Homework Assignment 4

Due Tuesday, February 15, by 11:59 PM

General homework guidelines: All homework assignments should be submitted using Canvas. Please submit your answers separately, for each of the problems, as set-up in the Canvas submission portal. You are allowed to either type in your answers, as well as submit graphs directly, within each submission item. You are also allowed to submit a scanned copy of your answer, as long as the answers are submitted separately for each question, as instructed. You can either scan your answers (ask if you don't have access to a scanner or do not know how to use your phone to do so), or just submit a picture of your answer. Please note that if we can't read your answer, we won't be able to award any (partial) credit.

You have one attempt to submit your answers. If technical issues appear and your submission portal has closed for some reason, please email Prof. Caragea explaining the situation and requesting permission to resubmit. Please note that such requests must be made before the deadline.

For full credit, please make sure you submit your HW by the deadline. A late submission is possible, with a 20% penalty, as long as it is turned in before the end of the day on Thursday following the due day (for this HW, it is February 17-th). No submissions will be accepted past this date.

Homework problems.

1. According to a Gallup poll, of 1,000 randomly selected adults aged 18 or older in the United States, 65% believe that global warming is more a result of human actions than natural causes.
 - (a) Use R to obtain a summary table and bar graph of the sample data.
 - (b) Describe the population proportion of interest p in words.
 - (c) Use R to calculate a 95% confidence interval for the population proportion p using the normal approximation method.
 - (d) Give the interpretation of the 95% confidence interval you calculated in part (c) in the context of the problem.
 - (e) Use R to calculate a 95% confidence interval for the population proportion p using the Wilson's score method. Compare the center and width of this interval to the one you calculated in part (c).
 - (f) Gallup is planning to conduct another poll on global warming. They would like to have a 95% confidence interval with a margin of error of no more than 2.5%. What sample size do they need to obtain this margin of error? Make sure to specify how you are calculating the sample size.
2. Unlike confidence intervals for other parameters, several methods have been developed for the confidence interval for the population proportion p . For all confidence intervals, you want to have a coverage rate, the percentage of confidence intervals containing the true population parameter calculated from a large number of samples, close to the stated confidence level for the confidence intervals. For example, if you generate 100,000 samples from a population and calculate 100,000 confidence intervals using each sample's data, you want approximately 95,000 or 95% of these confidence intervals to contain the population parameter.

In this problem, we will study the coverage rate for each method from lecture: the normal approximation method and the Wilson's score method using simulation. Since the binomial distribution depends on the sample size n and the population proportion p , we will simulate 100,000 samples from the binomial distribution with each combination of the values of n and p (9 total) in the table below.

n	25	250	1000
p	0.5	0.75	0.9

For each of the 100,000 simulated samples, we will determine if the value of p is located within the 95% confidence interval and use this information to calculate the coverage rate for each method. R code for running this simulation study is located in the file **HW3.Template.R** in Canvas. You will

need to enter the function in R before you can use the function. You will need to run the function 18 times, 9 times for each method and record the coverage rates for each run in a table similar to the one below. What do you notice about the coverage rates of the two methods? How do these results depend on the values of n and p ?

	Normal			Wilson's		
p	$n = 25$	$n = 250$	$n = 1000$	$n = 25$	$n = 250$	$n = 1000$
0.5						
0.75						
0.9						

3. Offspring of certain fruit flies may have either yellow or black bodies and either normal or short wings. Genetic theory predicts that these traits will appear with the following probabilities:

Yellow, Normal	Yellow, Short	Black, Normal	Black, Short
9/16	3/16	3/16	1/16

A researcher examines 200 flies and identifies the traits for each fly. These data can be found in the file **flies.csv**.

- Use R to give the summary table and a bar graph of the sample data.
- Give the null and alternative hypotheses for the goodness of fit test for the correctness of the genetic theory.
- Calculate the expected number of fruit flies from the "Yellow, Normal" category under the assumption the genetic theory is true. Only calculate this expected value here, and show your work. You may use R to verify your answer.
- Calculate the contribution of the "Yellow, Normal" category to the test statistic X^2 . Only calculate this contribution here, and show your work. You may use R to verify your answer.
- Use R to find the test statistic and p -value for this hypothesis test.
- Write a conclusion about the correctness of the genetic theory.