STAT 270 Assignment 5

Upload to Crowdmark by 5pm on Monday, Aug. 10, 2020

Instructions: You must provide any code you use (final computed values are not suarfficient). **I strongly recommend that you create a Jupyter Notebook for each question that requires data analysis**, embedding your written comments in Markdown cells. Create separate file(s) for each question (in .jpg, .pdf, or .png format) for uploading to Crowdmark.

1. (8 marks) Consider the problem from Assignment 4 where researchers are interested in the diagonal length (in cm) of fish observed at local fish markets. Unknown to them, the true mean length of these fish is 4.45 cm. They want to know whether the mean length of fish is less than the historical mean of 4.5 cm. They are willing to assume that lengths are iid and normally distributed with SD 2 cm. They use a significance level of 10%.

   a. (5 marks) What is the power of the original study (with $n = 159$ fish)? **NOTE**: You must show your calculations starting from the mathematical definition of power, i.e., do **not** simply provide a formula.
   b. (3 marks) If they would like to achieve a power of 90%, how many fish will they need to sample in total?

2. (10 marks) A cereal manufacturer claims that 1 in 10 of their cereal boxes has a code hidden inside that buyers can redeem for a special prize. Rumours begin circulating, however, that prizes are rarer than advertised. To estimate the actual probability of winning a prize, 45 boxes are randomly selected. Of these, 0 have codes for prizes.

   a. (4 marks) Construct a 95% Wald confidence interval for the probability of winning a prize.
   b. (1 mark) Which general problem with the Wald confidence interval does your confidence interval in a) illustrate?
   c. (4 marks) Construct a 95% Agresti-Coull confidence interval for the probability of winning a prize.
   d. (1 mark) Based on your interval in c), is the probability of winning as advertised? Explain.

3. (7 marks) A city will create bike lanes in their downtown area if more than 60% of residents are in favour. They conduct a survey of 400 randomly selected residents to assess their opinions. Say 266 residents support the plan. Use a significance level of 1%.

   a. (2 marks) State the null and alternative hypotheses.
   b. (1 mark) What is the observed test statistic?
   c. (1 mark) What is the distribution of the test statistic under the null hypothesis?
   d. (1 mark) What is the p-value?
   e. (2 marks) State your conclusions in the language of the problem.

4. (4 marks) In a poll of whether Vancouver should bid to host another Olympics Games, 141 out of 210 randomly selected residents under the age of 40 were in favour, while 111 out of 190 randomly selected residents aged 40 and above were in favour. Construct a 90% confidence interval for the difference in proportion of supporters across the two age groups.

5. (7 marks) In a random sample of 100 men and 100 women from a population, 24 of the men and 32 of the women have a particular genetic variant. Does the prevalence of this variant differ across the sexes? Use a significance level of 5%.

   a. (2 marks) State the null and alternative hypotheses.
   b. (1 mark) What is the observed test statistic?
   c. (1 mark) What is the distribution of the test statistic under the null hypothesis?
   d. (1 mark) What is the p-value?
   e. (2 marks) State your conclusions in the language of the problem.

6. (9 marks) Students in a distance education course are randomly assigned to either Section A or Section B. Both sections are taught by the same instructor and the materials and assessments in each are identical except that students in Section B have access to supplementary video lectures. The final marks of students in the two sections are recorded in the file `marks.txt`. Of interest is whether the mean mark in the population of students who have access to the videos differs from that in the population of students who do not. Use a significance level of 5%.

   a. (2 marks) Which three assumptions do you need to make about students' marks to be able to conduct a valid *t*-test in this context?
   b. (2 marks) State the null and alternative hypotheses.
   c. (1 mark) What is the observed test statistic?
   d. (1 mark) What is the distribution of the test statistic under the null hypothesis?
   e. (1 mark) What is the p-value?
   f. (2 marks) State your conclusions in the language of the problem.

7. (6 marks) Consider a random selection of 100 patients who undergo treatment for high blood pressure. Let Y_{i1} be the systolic blood pressure (in mmHg) of the i^{th} patient prior to treatment, let Y_{i2} be the systolic blood pressure (in mmHg) of the i^{th} patient post treatment. Assume that the differences in blood pressure (post- minus pre-treatment) are independent and approximately normally distributed. The data are available in the file `bp.txt`.
   a. (4 marks) Construct a 95% confidence interval for the mean effect of treatment, i.e., the mean change in systolic blood pressure following treatment.
   b. (2 marks) Does your confidence interval in a) require the Y_{i1}'s and the Y_{i2}'s to be independent and approximately normally distributed? Explain.

8. (8 marks) SFU decides to conduct a poll of the SFU community to investigate the demand for a gondola from Production Way up to the Burnaby campus. They are interested in the proportion of community members who would regularly use the gondola, if constructed (p). They suspect that this proportion lies between 0.1 and 0.3.

a. (3 marks) How many community members would SFU need to survey to be able to construct a 95% confidence interval of maximum width 0.08?

b. (5 marks) Consider the hypothesis test of whether p is less than 0.2. How many community members would SFU need to survey to achieve a power of 85% if p is actually 0.1? They plan to use a significance level of 5%. **NOTE**: You must show your calculations starting from the mathematical definition of power, i.e., do **not** simply provide a formula.