

STAT 200 Assignment 2
Winter 2020/21 Term 1

Instructions:

- Typewritten answers are preferred. If your handwriting is illegible or if the answers are not presented in a neat and organized manner, with answers clearly numbered you may lose marks.
- You must show steps with proper justification in your solutions. Partial credits are given to intermediate steps and reasonings. At the same time please keep your answer concise with important key points/steps.
- Please submit only pdf or image files (either PNG or JPG/JPEG). If you submit OS specific files (such as .pages files) and/or corrupted files, your assignment will not be graded.
- If you use R Commander to compute probabilities for any of the questions, please state clearly that you have done so and provide parameter values that you input in R Commander.

1. 2019 saw the first reported case of COVID-19. Scientists from the World Health Organization are tasked with creating protocols for quarantining infected patients. In particular, they must outline the length of time that patients are to be kept in quarantine before they are considered safe to release and no longer contagious to the public.

Assume that it is known that the time to recovery has a mean of 14.36 days and variance of 6.82 days². Be sure to define clearly any variables and models you use.

(a) Assuming that the time to recovery follows a Normal distribution, find the following:

- i. An infected patient is selected at random. What is the minimum number of days that they should be quarantined such that there is a 99.9% chance that they will be recovered before release? [3 marks]
- ii. 100 infected patients are selected at random. What is the probability that their mean time to recovery is no more than 15 days? [3 marks]
- iii. Suppose 21 patients on a cruise ship simultaneously and suddenly become infected with the virus. With only 2 weeks remaining on their voyage, what is the probability that all patients will be recovered before they return to port, as to not infect anybody on shore when they return on the 15th day? Assume that the patients' recovery are independent of one another. [4 marks]

(b) Assume the true distribution of time to recovery is unknown. Given this, consider the following:

- i. Explain one reason why the Normal distribution might be a poor representation of the true distribution of time to recovery? Briefly explain in 1 - 4 sentences. [2 marks]
- ii. Is the probability you calculated in part (1(a)ii) still accurate? Briefly explain in 1 - 4 sentences. [2 marks]
- iii. A random sample of 144 infected patients reveals that their mean time to recovery is 15.09 days. Is the mean recovery time of these 144 patients unusual? Justify your answer using probabilistic evidence. [2 marks]

2. Suppose a medical doctor suspects that 98% cases of COVID-19 would show fever symptoms. In order to study whether fever is one of the most common characteristics of COVID-19, this doctor randomly collected data on 1,099 patients with laboratory-confirmed COVID-19 and the result showed that 966 patients have experienced the fever.

(a) Use the doctors data to construct a 99% confidence interval for the true proportion of confirmed COVID-19 patients who had fever. [3 marks]

- (b) Write out the assumptions and conditions necessary for the interval you constructed in part 2a in the context of the question. [3 marks]
- (c) If we say “we are 99% confident the true proportion of confirmed COVID-19 patients who had fever is between the range of the interval you calculated in part (2a)”. Provide an interpretation of that sentence in the context of the question. [2 marks]
- (d) When planning to estimate a population proportion, the doctor needs to determine the appropriate sample size. Suppose the doctor doesn’t have any prior information about the proportion and desires the estimate to be correct to within 0.05 with 99% confidence, how large a sample size is needed? [3 marks]
3. A certain airline guarantees customers that the airline rarely loses passengers’ baggage. The public relations department claims that on those occasions when luggage is lost, 92% is recovered and delivered to its owner within 24 hours. An independent consumer group surveyed travelers on this airline and found that 145 out of 165 people who lost luggage received their missing bags within 24 hours. You will conduct a hypothesis test to see whether the data collected by the independent consumer group is different from the airline’s claim.
- (a) Identify the population of interest in the context of the question. [1 mark]
- (b) State the null and alternative hypotheses in the context of the question in words. [1 mark]
- (c) Determine/compute any conditions that must be valid to carry out the test in the context of the question. [2 marks]
- (d) Compute the test statistic. [2 marks]
- (e) Find the exact value or provide a range of values for the P-value. Sketch your model, label and shade the corresponding region (a sketch by hand is fine). [2 marks]
- (f) State your conclusion in the context of the question with a 5% significance level. [1 mark]
- (g) Based on your conclusion above, what type of error are you at risk of making? Explain your answer in the context of this question in 1 - 4 sentences. [2 marks]
4. In this exercise, we will examine two different varieties of wheat seeds: (1) Kama; and (2) Canadian. We are interested in the *length* of the wheat kernel. We are going to use real data, available here: <https://archive.ics.uci.edu/ml/datasets/seeds>. Below is sample of size 30 of each wheat variety’s measured kernel length:
- Kama:
- 5.717 5.712 5.351 5.832 5.226 5.395 5.139 5.384 5.877 5.388
 5.386 5.008 5.397 4.902 5.789 5.662 5.656 5.585 5.609 5.504
 5.618 5.099 5.757 5.479 5.554 5.579 5.527 5.520 5.701 5.630
- Canadian:
- 5.236 5.180 5.472 5.136 5.240 5.220 5.394 5.410 5.046 5.088
 5.073 5.236 5.180 5.363 5.317 5.176 5.386 5.175 5.009 5.186
 5.011 5.160 4.899 5.413 5.140 5.320 5.088 5.091 5.090 5.451
- Before you answer the questions below, think about the following (no need to write it down): just by looking to these data, can you conclude which wheat variety has a longer kernel? Imagine if the sample were of size 1000? [No marks]
- (a) Find the average kernel length of each type of wheat. Based on these averages alone, can we claim that we know that Kama seeds have a longer true kernel average than Canadian seeds? Briefly explain in 1 - 4 sentences why or why not. [2 marks]

- (b) Next, check how much the kernels' lengths vary – obtain the standard deviation for each type of wheat. [1 mark]
- (c) Find the standard error of the sample averages of each wheat type. [1 mark]
- (d) Obtain 99.7% confidence intervals for the population mean length of each wheat type. [4 marks]
- (e) Obtain a 95% confidence interval for the difference in true mean lengths between the two types of wheat. [2 marks]
- (f) Based on the interval you obtained in part (e), do the data suggest the true mean seed length is different for the two types of wheat? State your hypotheses (in words or define your notation) and conclusion in the context of the question. [3 marks]