Inf2 – Foundations of Data Science 2023

Task: Preparation for Semester 2 Week 4 Workshop

FOUNDATIONS OF DATA SCIENCE

24th January 2024

Please attempt the questions below. Bring your work on the questions to the workshop session. Don't worry if you get stuck – you can discuss why at the workshop session.

Acknowledgement: These questions are adapted from exercises in Devore & Berk (2012) *Modern Mathematical Statistics with Applications*, Springer.

Questions to attempt before the workshop

1. **Distribution of the sample mean**
   A retired statistician runs a cafe (called "The t shop"). She knows that the percentage of the bill that diners give as a tip has a mean value of $\mu = 9\%$ and a standard deviation of $\sigma = 6\%$.

   (a) Suppose a random sample of $n = 100$ bills is drawn at random. If $\overline{X}$ is the sample mean, what type of distribution to you expect the sampling distribution of $\overline{X}$ to be?

   (b) Where is the mode of the sampling distribution of $\overline{X}$?

   (c) What is the standard error in the mean?

   (d) What is the approximate probability that the sample mean tip is greater than 8%? Hint: you will need to find the area under a normal curve, and can do this by using a table such as Table A.3 in *Modern Mathematical Statistics with Applications*.

2. **Confidence interval calculation 1**
   A random sample of 110 lightning flashes in a region resulted in a sample mean radar echo duration of 0.81s and a sample standard deviation of 0.34s ("Lighting strikes to an Airplane in a Thunderstorm" *J. Aircraft* 21: 607–611, 1984). Calculate a 99% (two-sided) confidence interval for the true mean echo duration μ, and interpret the resulting interval.

3. **Confidence interval calculation 2**
   Here is a sample of ACT scores (mean of Maths, English, Social Science and Natural Science scores) for students taking a first year calculus course:

24.00	28.00	27.75	27.00	24.25	23.50	26.25
24.00	25.00	30.00	23.25	26.25	21.50	26.00
28.00	24.50	22.50	28.25	21.25	19.75	

   (a) Calculate a two-sided 95% confidence interval for the population mean.

   (b) The University ACT mean for first years starting that year was about 21. Are the calculus students better than average, as measured by the ACT?

Questions to discuss at the workshop

4. **Confidence intervals concepts**
   Suppose that a random sample of 50 bottles of a particular brand of cough syrup is selected and that the alcohol content of each bottle is determined. Let μ denote the mean alcohol content for the population of all bottles of the brand under study. Suppose that the resulting 95% confidence interval for μ is (7.8, 9.4).

   (a) Would a 90% confidence interval calculated from this sample have been narrower or wider than the given interval? Explain your reasoning.

   (b) Consider the following statement: There is a 95% chance that μ is between 7.8 and 9.4. Is that statement correct? Why or why not?

   (c) Consider the following statement: We can be highly confident that 95% of all bottles of this type of cough syrup have an alcohol content that is between 7.8 and 9.4. Is this statement correct? Why or why not?

   (d) Consider the following statement: If the process of selecting a sample of size 50 and then computing the corresponding 95% interval is repeated 100 times, exactly 95 of the resulting intervals will include μ. Is this statement correct? Why or why not?

   (e) In order to make the 95% confidence interval three times narrower, how many samples would we need to collect?

5. **Distribution of the sample mean**

   (a) In Question 1, suppose that there is now a random sample of 10 bills. Discuss whether you could repeat your work in 1(a)–1(d).

   (b) Looking further at the data, the owner of the *t*-shop notices that only about 69% of customers leave a tip at all – the other customers don't give a tip. The distribution of tips given by these customers is a uniform distribution between 10% and 16%. Does this information affect your answer to 5a? You may want to try running statistical simulations to help understand the problem.

6. **Checking for normality**
   In question 3, before applying the test, we should have checked that the distribution was approximately normal. Use a normal probability plot (*Modern Mathematical Statistics with Applications* p. 211), also known as a "Q-Q plot" to assess visually if the data is normally distributed.