Stat 22000 Summer 2020 Homework 9 Solutions

Problems to Turn In: due **midnight of Thursday, July 16, on Canvas**.

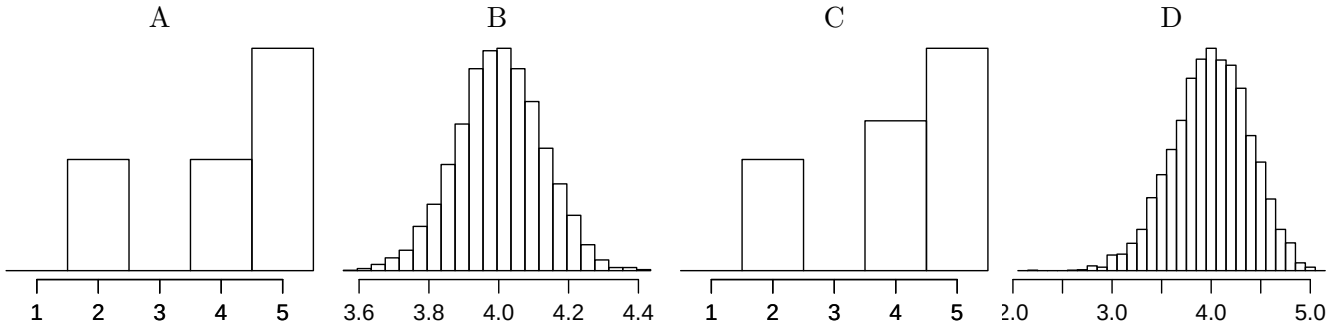
1. Suppose if an introductory statistics textbook was given to all potential readers for review in 1-to-5 stars rating system, with 5 stars being the best, the population distribution of the ratings would be as follows:

   50% of the readers give 5 stars, 25% give 4 stars, 25% of the readers give 2 stars.

   Of course it is impossible to ask all potential readers to rate the textbook. So the book seller samples potential readers from the population and ask them to rate the book.

   The following graphs are

   (1) Histogram of the ratings given by all potential readers
   (2) Histogram of the ratings given by 100 randomly sampled readers
   (3) Histogram of 5000 sample means of the ratings from random samples of each size 10
   (4) Histogram of 5000 sample means of the ratings from random samples of each size 100



   Determine which graph (A, B, C, or D) is which and explain your reasoning.

Answer: *[4pts in total = 2pts for which is which + 2pts for the reasons]*

A — (1), B — (4), C — (2), D — (3).

Plot A is (1) the histogram of the population because the first two bars are equally tall, and the last bar is twice as tall as the first two, which matches with the makeup of the population (25% are 2 stars, 25% are 4 stars, and 50% are 5 stars).

Plot C is (2) the histogram of the sample because it looks like the histogram of the population (3 bars at 2, 4, and 5) but not exactly identical).

Plot B and D are both histograms of the sample means because they are not always integers and looks more normal than the histogram of the population. Plot B is more normal and is less variable than Plot D. So Plot B is the histogram for the sample means of a larger sample.

---

2. The 2008 General Social Survey asked, "*What do you think is the ideal number of children for a family to have?*" The 678 females who responded had a median of 2, mean of 3.22, and standard deviation of 1.99.

   (a) What is the point estimate of the population mean (mean of ideal number of children answered by all U.S. women)?

(b) Find the standard error of the sample mean.

(c) Verify that the 99% confidence interval for the population mean is $(3.02, 3.42)$.

(d) Is it plausible that the population mean is 2? Explain.

Answer: *[7 points in total]*

(a) *[1pt]* The point estimate of the population mean is the sample mean 3.22.

(b) *[2pts]* SE $= s/\sqrt{n} = 1.99/\sqrt{678} \approx 0.076$.

(c) *[2pts]* 99% CI is (point estimate) $\pm\, 2.58\times$ SE $= 3.22 \pm 2.58 \times 0.076 \approx 3.22 \pm 0.20 = (3.02, 3.42)$.

(d) *[2pts]* Not plausible because 2 is not in the 99% CI for the population mean.

---

3. Continue the previous problem. Determine whether the following statements are true or false, and explain your reasoning.

(a) From the data summary, the distribution of the sample is not normal.

(b) This confidence interval $(3.02, 3.42)$ is not valid since the distribution of the sample is not normal

(c) We are 99% confident that the average of the ideal number of children in a family answered by the 678 females in the sample is between 3.02 and 3.42.

(d) 99% the women in the samples thought that the ideal number of children is between 3.02 and 3.42.

(e) If a new sample was taken, we are 99% confident that the sample mean of the new sample would lie between 3.02 and 3.42.

(f) The margin of error of the 99% confidence interval (3.02, 3.42) is 0.2.

(g) In order to decrease the margin of error of a 99% confidence interval by half, we would need to use a sample twice as large.

(h) A 95% confidence interval would be narrower than the 99% confidence interval since we don't need to be as sure about our estimate.

Answer: *[14 points in total]*

(a) *[2pts]* True. The reason can be any of the following:

- If normal, the mean and median would be about equal. For this sample, the mean 3.22 was substantially higher than the median 2.
- The ideal number of children must be ≥ 0. There won't be observations more than $(3.22 - 0)/1.99 \approx 1.62$ SDs below the mean. If the sample is approximately normal, we expect to have about 5% of the data with z-score < -1.62.
- People's answer about their ideal number of children is usually a small integer: 0, 1, 2, ... Very few will give a number more than 20. So people's answer about their ideal number of children is a discrete variable. The normal distributions are continuous distributions. So the sample won't be normal.

(b) *[2pts]* False. Even though the sample is not normally distributed, by CLT, with a large enough sample size ($n = 689$), the sampling distribution of the sample mean is nearly normal. Based on that, we can construct a confidence interval.

(c) *[2pts]* False. Inference is made on the population parameter (population mean here), not the sample mean. The sample mean is always in the confidence interval.

(d) [2pts] False. Apparently, people's ideal number of children would be a whole number. There is no whole number between 3.02 and 3.42. A 99% confidence interval is to cover the population mean with 99% confidence, not to cover 99% of the population.

(e) [2pts] False, the confidence interval is not about the sample mean of another sample. The correct interpretation of the confidence level would be that 99% of random samples produce 99% confidence intervals that include the true population mean.

(f) [1pt] True. The margin of error is half the width of the interval: $(3.42 - 3.02)/2 = 0.2$.

(g) [2pts] False, since in calculation of the standard error we divide the standard deviation by square root of the sample size, in order to cut the standard error by half (and hence the margin of error) we would need to sample $2^2 = 4$ times the number of people in the initial sample.

(h) [1pt] True. A 95% CI is (point estimate)$\pm 1.96\times$ SE, narrower than a 99% CI: (point estimate)$\pm 2.58\times$ SE.
