JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Lecture 3c: Practice Problem Solutions

John McGready
Johns Hopkins University

Estimating a 95% Confidence Interval

- Suppose an independent environmental group computes the gas mileage for a random sample of 100 new models of the same car make and model in order to make a statement about the gas mileage of this make and model. The results on these 100 cars include the following summary statistics:
  - Sample mean mileage, 31.4 mpg
  - Sample standard deviation: 1.2 mpg
  - Sample median: 31.2 mpg

Estimating a 95% Confidence Interval

- Assuming the gas mileage data is normally distributed, estimate a range of gas mileage for most (95%) of the cars of this make and model based on the sample results
  - If the sample data comes from a distribution of normally distributed values, then we can estimate this interval from the sample results by using $\bar{x} \pm 2 \times s$
  - In this example, applying this logic gives an interval of $31.4 \pm 2 \times 1.2 \rightarrow 31.4 \pm 2.4 \rightarrow (29.0\,mpg, 33.8\,mpg)$.

Estimating a 95% Confidence Interval

2. Without assuming normality, estimate a range of gas mileage for most (95%) of the cars of this make and model based on the sample results.
   - This is a trick question: given only the sample mean, median, and standard deviation, it is not possible to answer this question. One may note that the mean and median are similar in value and hence the sample distribution is likely symmetric, but this cannot be verified with a histogram or some other approach. If we had the appropriate sample percentile values (2.5th and 97.5th), then this question could be answered without assuming normality.

Estimating a 95% Confidence Interval

3. Assuming the gas mileage data is normally distributed, estimate a 95% confidence interval for the mean gas mileage for all cars that are this make and model.

   - As the sample is large, we can go ahead and estimate a 95% CI for the true average gas mileage among all cars of this make and model using the CLT based formula, $\bar{x} \pm 2 \times \hat{SE}(\bar{x})$
   - i.e., $\bar{x} \pm 2 \times {}^{s}\!/\!_{\sqrt{n}}$
   - For these sample results, this gives:
   - $31.4 \pm 2 \times \frac{1.2}{\sqrt{100}} \rightarrow 31.4 \pm 0.24 \rightarrow$

     $(31.16\ mpg, 31.64\ mpg) \approx (31.2\ mpg, 31.6\ mpg)$

Estimating a 95% Confidence Interval

4. Without assuming normality, estimate a 95% confidence interval for the mean gas mileage for all cars that are this make and model
   - Actually, the normality assumption about the original individual data is NOT needed to use the CLT based approach to create a 95% CI for the true underlying population mean of the population from which the sample was taken
   - Hence the answer to this question is EXACTLY the same as the answer to the previous question (3)
   - This highlights one of the powers of the CLT result: that we can estimate a valid confidence interval for a population mean using only a single random sample of data, regardless of what the population distribution (and random sample distribution) of the individual level data is

Estimating a 95% Confidence Interval

5. What is the difference in the interpretation of the intervals created in questions 1/2 and questions 3/4?
   - This is a very important distinction
   - The first two questions deal with estimating an interval that contains most (95%) of the individual (car) level values (gas mileage) in the population of all cars of this make and model: it is giving a range for individual observations
   - This interval is NOT a confidence interval, and requires specific characteristics of the population/sample data if it is to be estimated using only the sample mean and standard deviation

Estimating a 95% Confidence Interval

5. What is the difference in the interpretation of the intervals created in questions 1/2 and questions 3/4?
   - The second set of questions asks for a 95% confidence interval for a single summary measure about the population from which the sample was taken
   - A 95% CI gives a range of plausible values for a single population summary quantity (like a mean) using the information in a single sample to estimate the sampling error in the estimated sample mean
   - This 95% CI is NOT a range of possibilities for individual level values in the population from which the sample was taken, but a range of possibilities for a single number describing this entire population—in this case a mean
   - This can be estimated without making assumptions about the distribution of individual level values in the sample/population