Statistical Inference

Computational Mathematics and Statistics Camp
University of Chicago
September 2018

1. The distribution of the number of eggs laid by a certain species of hen during their breeding period has a mean of 35 eggs with a standard deviation of 18.2. Suppose a group of researchers randomly samples 45 hens of this species, counts the number of eggs laid during their breeding period, and records the sample mean. They repeat this 1,000 times, and build a distribution of sample means.

   a. Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

      Because we are dealing with the distribution of sample means, we need to check to see if the Central Limit Theorem applies. Our sample size is greater than 30, and we are told that random sampling is employed. With these conditions met, we expect that the distribution of the sample mean will be nearly normal and therefore symmetric.

   b. Calculate the variability of this distribution and state the appropriate term used to refer to this value.

      Because we are dealing with a sampling distribution, we measure its variability with the standard error. $SE = \frac{18.2}{\sqrt{45}} = 2.713$.

   c. Suppose the researchers' budget is reduced and they are only able to collect random samples of 10 hens. The sample mean of the number of eggs is recorded, and we repeat this 1,000 times, and build a new distribution of sample means. How will the variability of this new distribution compare to the variability of the original distribution?

      The sample means will be more variable with the smaller sample size.

2. A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter. The standard error for this estimate was 2.4%, and a normal distribution may be used to model the sample proportion.

   a. Construct a 99% confidence interval for the fraction of U.S. adult Twitter users who get some news on Twitter, and interpret the confidence interval in context.

      Recall the general formula is Point estimate $\pm\, z^* \times SE$. First, identify the three different values. The point estimate is 52%, $z^* = 2.58$ for a 99% confidence level, and $SE = 2.4\%$. Then plug the values into the formula:

      $$52\% \pm 2.58 \times 2.4\% \rightarrow (45.8\%, 58.4\%)$$

   b. Identify the follow statements as true or false. Provide an explanation to justify each of your answers.

      a. The data provide statistically significant evidence that more than half of U.S. adult Twitter users get some news through Twitter. Use a significance level of $\alpha = 0.01$.

         False. 50% is included in the 99% confidence interval, hence a null hypothesis of $p = 0.50$ would not be rejected at this level.

      b. Since the standard error is 2.4%, we can conclude that 97.6% of all U.S. adult Twitter users were included in the study.

         False. The standard error measures the variability of the sample proportion, and is unrelated to the proportion of the population included in the study.

c. If we want to reduce the standard error of the estimate, we should collect less data.

False. We need to increase the sample size to decrease the standard error.

d. If we construct a 90% confidence interval for the percentage of U.S. adults Twitter users who get some news through Twitter, this confidence interval will be wider than a corresponding 99% confidence interval.

False. As the confidence level decreases so does the margin of error, and hence the width of the confidence interval.

3. A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. She collects a simple random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. A 95% confidence interval based on this sample is (128 minutes, 147 minutes), which is based on the normal model for the mean. Determine whether the following statements are true or false, and explain your reasoning.

a. This confidence interval is not valid since we do not know if the population distribution of the ER wait times is nearly Normal.

False, even if the population distribution is not normal, with a large enough sample size we can assume that the sampling distribution is nearly Normal and calculate a confidence interval. It would however be ideal to see a histogram or Normal probability plot of the population distribution to assess the level of skewness. If the distribution is very strongly skewed, the sample size of 64 may be insufficient for the sample mean to be approximately normally distributed.

b. We are 95% confident that the average waiting time of these 64 emergency room patients is between 128 and 147 minutes.

False, inference is made on the population parameter, not the point estimate. The point estimate is always in the confidence interval.

c. We are 95% confident that the average waiting time of all patients at this hospital's emergency room is between 128 and 147 minutes.

True, this is the correct interpretation of the confidence interval.

d. 95% of random samples have a sample mean between 128 and 147 minutes.

False, the confidence interval is not about a sample mean. The true interpretation of the confidence level would be that 95% of random samples produce confidence intervals that include the true population mean.

e. A 99% confidence interval would be narrower than the 95% confidence interval since we need to be more sure of our estimate.

False. To be more confident that we capture the parameter, we need a wider interval.

f. The margin of error is 9.5 and the sample mean is 137.5.

True, since the normal model was used to model the sample mean. The margin of error can be calculated as half the width of the interval, and the sample mean is the midpoint of the interval.

$$ME = \frac{146 - 128}{2} = \frac{19}{2} = 9.5 \quad \bar{x} = 128 + 9.5 = 137.5$$

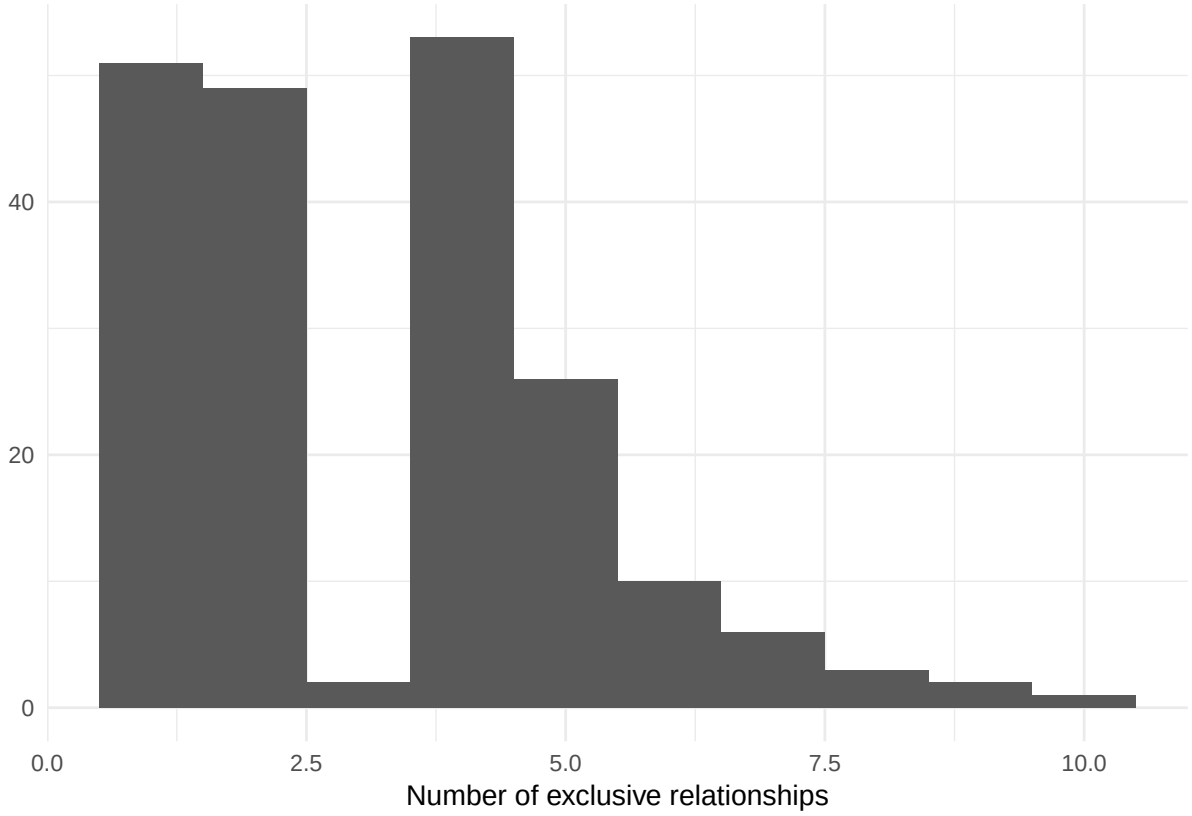
g. In order to decrease the margin of error of a 95% confidence interval to half of what it is now, we would need to double the sample size.

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 in half (and hence the margin of error) we would need to sample $2^2 = 4$ times the number of people in the initial sample.

$$ME_{\text{half as big}} = Z^* \frac{s}{\sqrt{n}} \rightarrow \frac{1}{2} ME_{\text{original}} = Z^* \frac{s}{\sqrt{4n}} = Z^* \frac{s}{2\sqrt{n}} = \frac{1}{2} Z^* \frac{s}{\sqrt{n}} = \frac{1}{2} ME_{\text{original}}$$

4. A survey conducted on a reasonably random sample of 203 undergraduates asked, among many other questions, about the number of exclusive relationships these students have been in. The histogram below shows the distribution of the data from this sample.

   The sample average is 3.2 with a standard deviation of 1.97.



   Estimate the average number of exclusive relationships Duke students have been in using a 90% confidence interval and interpret this interval in context. Check any conditions required for inference, and note any assumptions you must make as you proceed with your calculations and conclusions.

   The assumptions that need to be satisfied in order to proceed with the confidence interval are as follows:

   1. **Independence**: The sample is from less than 10% of all undergraduates, and it is a random sample. We can assume that the students in this sample are independent of each other with respect to number of exclusive relationships they have been in. Notice that there are no students who have had no exclusive relationships in the sample, which suggests some student responses are likely missing (perhaps only positive values were reported)
   2. **Normality**: The data are strongly skewed, but the sample is relatively large.

   If the assumption about the sample being reasonably equivalent to a simple random sample is acceptable, then the confidence interval can be calculated as follows:

$$\begin{aligned} \bar{x} \pm Z^* SE_{\bar{x}} &= 3.2 \pm 1.65 \times \frac{1.97}{\sqrt{203}} \\ &= 3.2 \pm 0.23 \\ &= (2.97, 3.43) \end{aligned}$$

5. The nutrition label on a bag of potato chips says that a one ounce (28~gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a sample mean of 134 calories with a standard deviation of 17 calories. Is there evidence that the nutrition label does not provide an accurate measure of calories in the bags of potato chips? We have verified the independence, sample size, and skew conditions are satisfied.

   The hypotheses are as follows:

$$H_0 : \mu = 130$$
$$H_A : \mu \neq 130$$

   The test statistic and the p-value can be calculated as follows:

$$Z = \frac{134 - 130}{\frac{17}{\sqrt{35}}} = 1.39$$
$$\begin{aligned} \text{p-value} &= P(Z < -1.39) + P(Z > 1.39) \\ &= 2 \times 0.0823 \\ &= 0.1646 \end{aligned}$$

   Since the p-value $> \alpha$ (since not given, use 0.05), fail to reject H_0. The data do not provide convincing evidence that the true average calorie content in bags of potato chips is different than 130 calories.

6. A patient named Diana was diagnosed with Fibromyalgia, a long-term syndrome of body pain, and was prescribed anti-depressants. Being the skeptic that she is, Diana didn't initially believe that anti-depressants would help her symptoms. However after a couple months of being on the medication she decides that the anti-depressants are working, because she feels like her symptoms are in fact getting better.

   a. Write the hypotheses in words for Diana's skeptical position when she started taking the anti-depressants.

      - H_0: Anti-depressants do not work for the treatment of Fibromyalgia.
      - H_A: Anti-depressants work for the treatment of Fibromyalgia. Diana might also have taken special note if her symptoms got much worse, so a more scientific approach would have been to use a two-sided test. If you proposed a two-sided approach, your answers in (b) and (c) will be different.

   b. What is a Type 1 Error in this context?

      Concluding that anti-depressants work for the treatment of Fibromyalgia symptoms when they actually do not (false positive).

   c. What is a Type~2 Error in this context?

      Concluding that anti-depressants do not work for the treatment of Fibromyalgia symptoms when they actually do.

7. A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

   a. Is the distribution of housing prices in Topanga symmetric, right skewed, or left skewed? **Hint:** Sketch the distribution.

      Right skewed. There is a long tail on the higher end of the distribution but a much shorter tail on the lower end.

b. Would you expect most houses in Topanga to cost more or less than \$1.3 million?

Less than, as the median would be less than the mean in a right skewed distribution.

c. Can we estimate the probability that a randomly chosen house in Topanga costs more than \$1.4 million using the normal distribution?

We should not since the population distribution is not normal.

d. What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

Even though the population distribution is not normal, if the conditions for inference are satisfied we may be able to use the Central Limit Theorem to estimate this probability.

1. **Independence**: Houses are sampled randomly in 60 houses are (presumably) less than 10% of all houses in Topanga, so the observations are independent.
2. **Normality**: There are a few large outliers, which suggests that the data are very strongly skewed. In practice, we should ask to see the data and check whether this assessment is accurate, and if so, we may need to perform a more advanced analysis. However, the data are not readily available, so we caution that the calculated probability may only be a rough estimate.

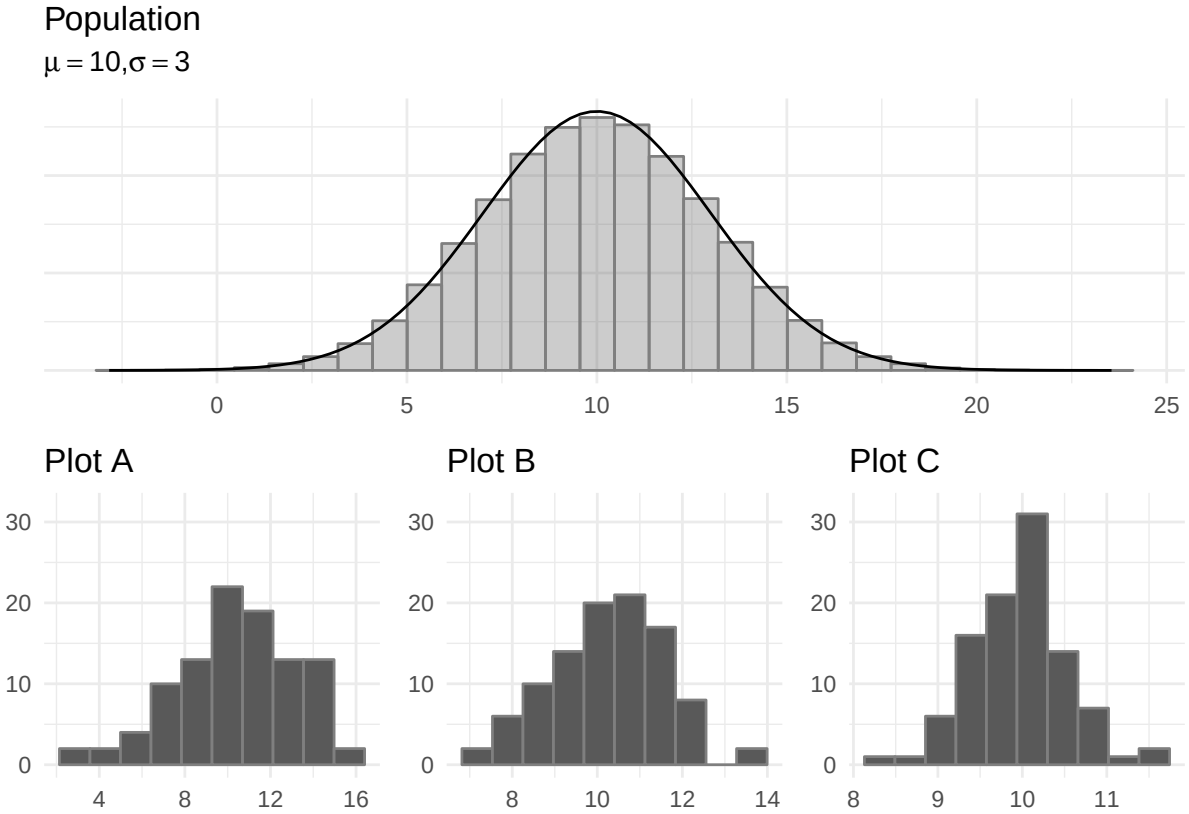
$$\begin{aligned} \bar{X} &\sim N\left(\mu_{\bar{x}} = 1.3, SE_{\bar{x}} = \frac{0.3}{\sqrt{60}}\right) \\ P(\bar{X} > 1.4) &= P\left(Z > \frac{1.4 - 1.3}{\frac{0.3}{\sqrt{60}}}\right) \\ &= P(Z > 2.58) \\ &= 1 - 0.9951 \\ &= 0.0049 \end{aligned}$$

e. How would doubling the sample size affect the standard deviation of the mean?

Doubling the sample size would decrease the standard error by a factor of $\sqrt{2}$ – that is, the new standard error will be $\sqrt{2}$ times the old standard error:

$$SE_n = \frac{s}{\sqrt{n}} \rightarrow SE_{2n} = \frac{s}{\sqrt{2n}} = \frac{s}{\sqrt{2}\sqrt{n}} = \frac{1}{\sqrt{2}}\frac{s}{\sqrt{n}} = \frac{1}{\sqrt{2}}SE_n$$

8. Four plots are presented below. The plot at the top is a distribution for a population. The mean is 10 and the standard deviation is 3. Also shown below is a distribution of (1) a single random sample of 100 values from this population, (2) a distribution of 100 sample means from random samples with size 5, and (3) a distribution of 100 sample means from random samples with size 25. Determine which plot (A, B, or C) is which and explain your reasoning.



The centers are the same in each plot. Additionally, each data set is from a nearly normal distribution, though the histograms may not look very normal since each represents only 100 data points. The only way to tell which plot corresponds to which scenario is to examine the variability of each distribution. Plot A is the most variable, followed by Plot B, then Plot C. This means Plot A will correspond to the original data, Plot B to the sample means with size 5, and Plot C to the sample means with size 25.

9. Suppose the area that can be painted using a single can of spray paint is slightly variable and follows a nearly normal distribution with a mean of 25 square feet and a standard deviation of 3 square feet.

   a. What is the probability that the area covered by a can of spray paint is more than 27 square feet?

   Let X be the amount of area a can of spray paint covers, then $X \sim N(\mu = 25, \sigma = 3)$.

$$\begin{aligned} P(X > 27) &= P\left(Z > \frac{27 - 25}{3}\right) \\ &= P(Z > 0.67) \\ &= 1 - 0.7486 \\ &= 0.2514 \end{aligned}$$

   b. Suppose you want to spray paint an area of 540 square feet using 20 cans of spray paint. On average, how many square feet must each can be able to cover to spray paint all 540 square feet?

   $\frac{540 \text{ square feet}}{20 \text{ cans}} = 27$ square feet per can.

   c. What is the probability that you can cover a 540 square feet area using 20 cans of spray paint?

   We are looking for the probability that the total of 20 cans cover more than 540 square feet. This means that one can on average should cover $\frac{540}{20} = 27$ square feet. To find this using CLT, but first we must check that the conditions hold.

1. Independence: We are not told that this is a random sample but it may be a safe assumption that the amount of paint in a can of spray paint is determined by a random process. 20 cans certainly make up less than 10% of all cans of spray paint. Therefore, the area the sampled cans of spray paint cover are independent.
2. Normality: The population distribution is nearly normal.

Let $\bar{X}$ be the average area covered by one can in a sample of 20 cans, then

$$\bar{X} \sim N\left(\mu_{\bar{x}} = 25, SE_{\bar{x}} = \frac{3}{\sqrt{20}}\right)$$

$$\begin{aligned} P(\bar{X} > 27) &= P\left(Z > \frac{27 - 25}{\frac{3}{\sqrt{20}}}\right) \\ &= P(Z > 2.98) \\ &= 1 - 0.9986 \\ &= 0.0014 \end{aligned}$$

d. If the area covered by a can of spray paint had a slightly skewed distribution, could you still calculate the probabilities in parts (a) and (c) using the normal distribution?

We could not calculate (a) without a nearly normal population distribution. Without this and a small sample size of $n = 20$, the sampling distribution will not be nearly normal. Therefore we could not calculate (c) either.

10. It is believed that nearsightedness affects about 8% of all children. In a random sample of 194 children, 21 are nearsighted.

a. Construct hypotheses appropriate for the following question: do these data provide evidence that the 8% value is inaccurate?

- $H_0 : p = 0.08$: Proportion of children who are near sighted is 8%
- $H_A : p \neq 0.08$: Proportion of children who are near sighted is not 8%

b. What proportion of children in this sample are nearsighted?

$\frac{21}{194} = 0.108 \rightarrow 10.8\%$ of children in this sample are near sighted.

c. Given that the standard error of the sample proportion is 0.0195 and the point estimate follows a nearly normal distribution, calculate the test statistic (the Z-statistic).

$$Z = \frac{\hat{p} - p}{SE} = \frac{0.108 - 0.08}{0.0195} = 1.44$$

d. What is the p-value for this hypothesis test?

$$\text{p-value} = 2 \times P(Z > 1.44) = 2 \times (1 - 0.9251) = 0.1498$$

e. What is the conclusion of the hypothesis test?

Since the p-value is large we fail to reject H_0. The data do not provide convincing evidence that the proportion of children who are nearsighted is different than 8%.

11. Determine whether the following statement is true or false, and explain your reasoning: "With large sample sizes, even small differences between the null value and the point estimate can be statistically significant."

True. If the sample size is large, then the standard error will be small, meaning even relatively small differences between the null value and point estimate can be statistically significant. **This is kind of a big deal when analyzing big data.**