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

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.
    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 $\alpha = 0.01$.
        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.
        c. If we want to reduce the standard error of the estimate, we should collect less data.
        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.

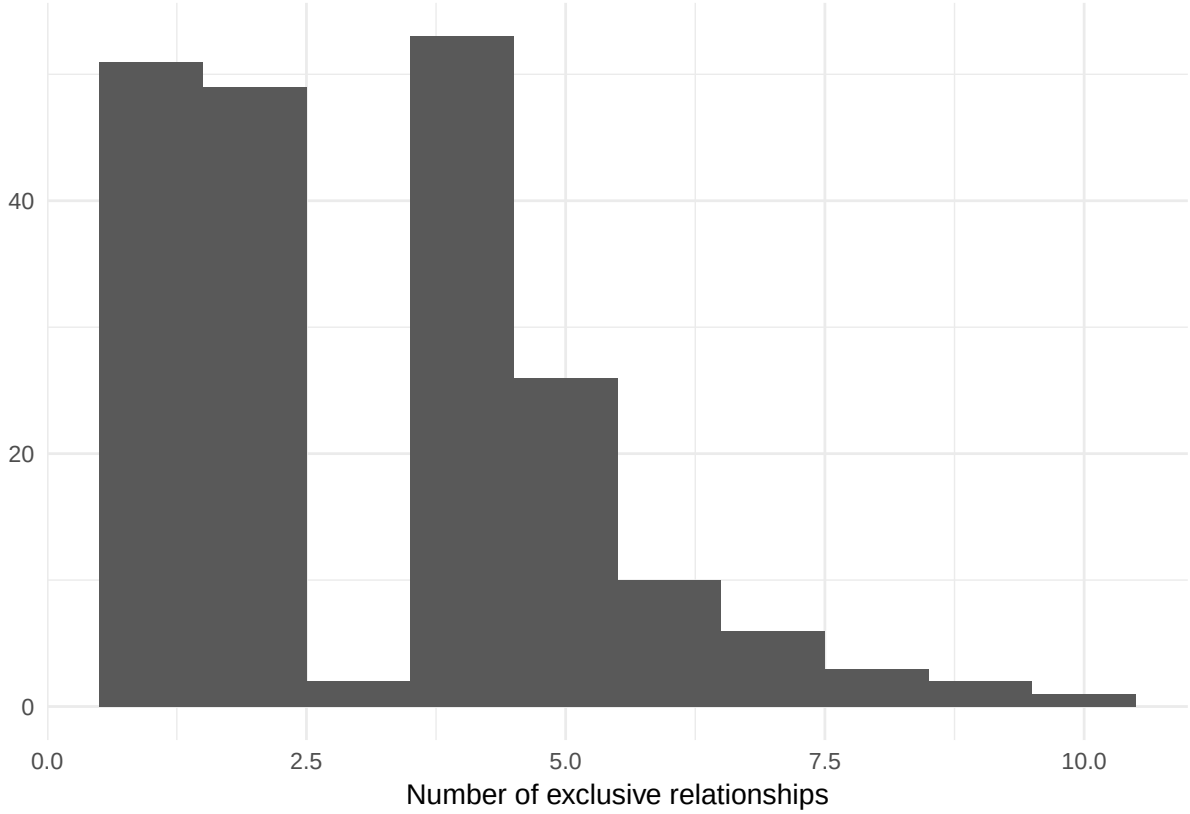
3. A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. She collects a simple random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. 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.
    b. We are 95% confident that the average waiting time of these 64 emergency room patients is between 128 and 147 minutes.
    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.
    d. 95% of random samples have a sample mean between 128 and 147 minutes.
    e. A 99% confidence interval would be narrower than the 95% confidence interval since we need to be more sure of our estimate.
    f. The margin of error is 9.5 and the sample mean is 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.

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.

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.

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.

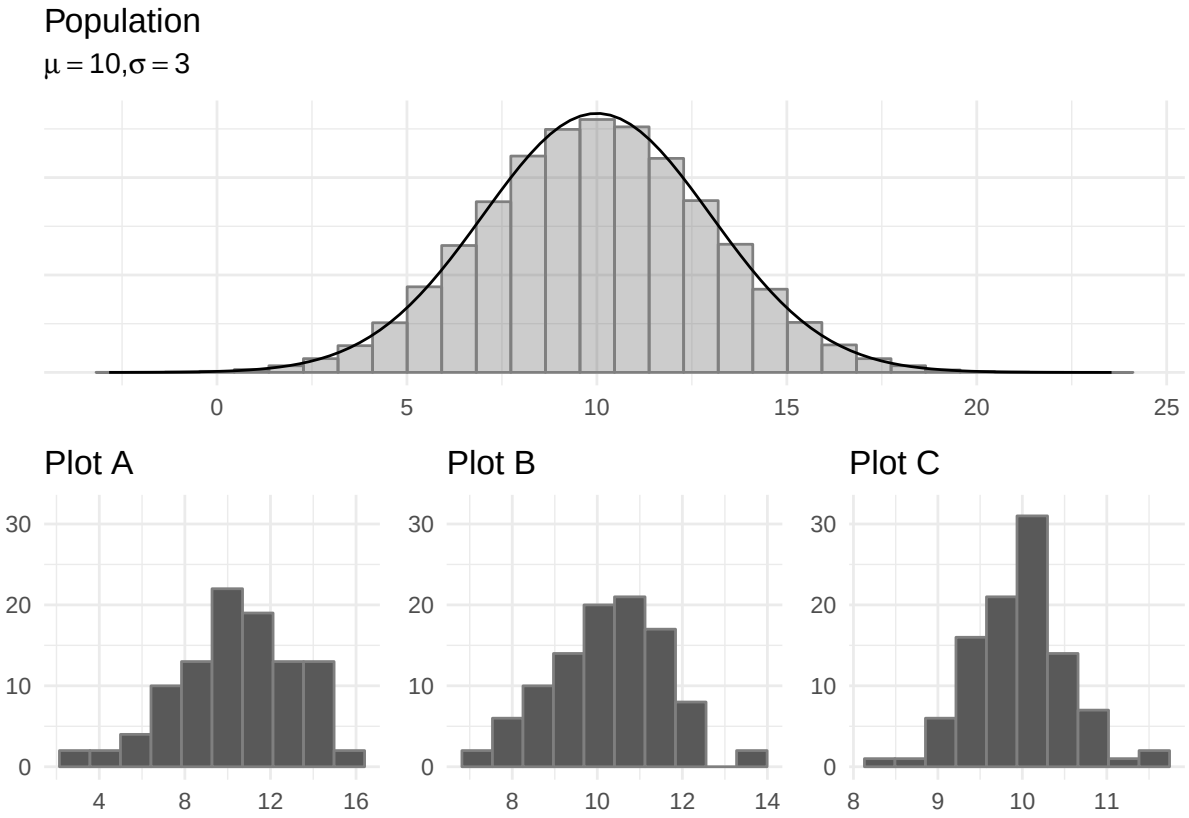
   a. Write the hypotheses in words for Diana's skeptical position when she started taking the anti-depressants.
   b. What is a Type 1 Error in this context?
   c. What is a Type~2 Error in this context?

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.
b. Would you expect most houses in Topanga to cost more or less than \$1.3 million?
c. Can we estimate the probability that a randomly chosen house in Topanga costs more than \$1.4 million using the normal distribution?
d. What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?
e. How would doubling the sample size affect the standard deviation of the mean?

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.



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

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?
b. What proportion of children in this sample are nearsighted?
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).
d. What is the p-value for this hypothesis test?
e. What is the conclusion of the hypothesis test?

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."