Worksheet 6

Author:

Discussants:

==========

The following worksheet will give you practice understanding sampling distributions and standard errors. Please answers the following questions and submit a compiled pdf with your answers to Moodle by 11:59pm on Sunday October 21th.

A list of functions we have used in class can be found on a shared class google document here. Some useful symbols are $\bar{x}$, $\hat{p}$, μ, π

Part 1: A few Lock 5 exercises examining sampling distributions and confidence intervals

Lock5 exercise 3.24 (1st edition): **Average Household Size** The latest US Census lists the average household size for all households in the US as 2.61. (A household is all people occupying a housing unit as their primary place of residence). Figure 3.6 shows possible distributions of means for 1000 samples of household sizes. The scale on the horizontal axis is the same in all four cases.

a) Assume that two of the distributions show results from 1000 random samples, while two others show distributions from a sampling method that is biased. Which two dotplots appear to show samples produced using a biased sampling method. Explain your reasoning. Pick one of the distributions that you listed as biased and describe a sampling method that might produce this bias.

b) For the two distributions that appear to show results from random samples, suppose that one comes from 1000 samples of size n = 100 and one comes from 1000 samples of size n = 500. Which distribution goes with which sample size? Explain.

Answers:

a)

b)

Lock5 exercise 3.26 (1st edition): **Mix It Up for Better Learning** In preparing for a test on a set of material, is it better to study one topic at a time or to study topics mixed together? In one study, a sample of fourth graders were taught for equations. Half of the children learned by studying repeated examples of one equation at a time, while the other half studied mixed problem sets that included examples of all four types of calculations grouped together. A day later, all the students were given a test on the material. The students in the mixed practice group had an average grade of 77, while the students in the one-at-a-0time group had an average grade of 30. What is the best estimate for the difference in the average grade between fourth-grade students who study mixed problems and those who study each equation independently? Give notation (as a difference with a minus sign) for the quantity we are trying to estimate, notation for the quantity that gives the best estimate, and the value of the best estimate. Be sure to clearly define any parameters in the context of this situation.

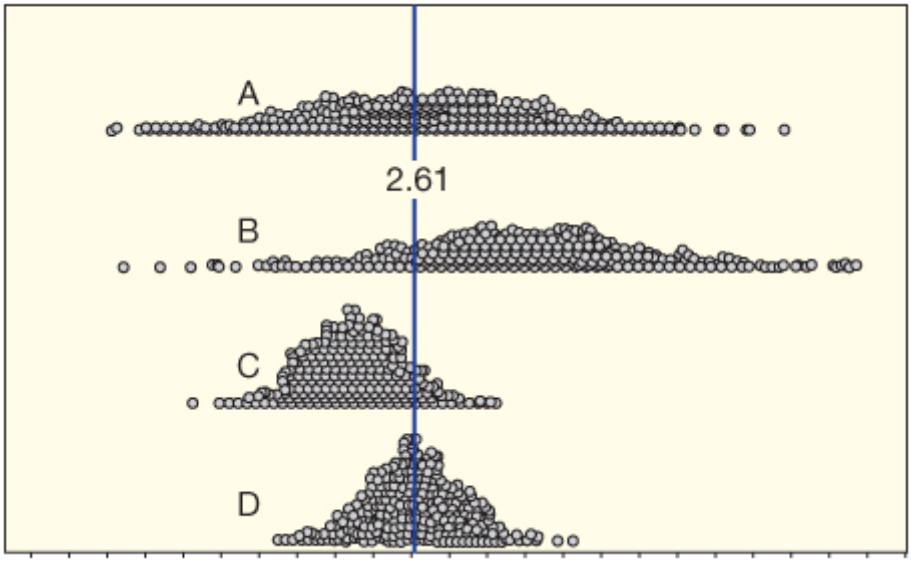

Figure 1: Figure 3.6

Answers:

Lock5 exercise 3.28 (1st edition): *3.28 Hollywood Movies Data* The data set HollywoodMovies2011 contains information on 136 movies that came out in 2011. One of the variables in this data set the budget (in millions of dollars) to make the movie. Figure 3.8 shows two boxplots. One represents the budget data for one random sample of size n = 30. The other represents the value in a sampling distribution of 1000 means of budget data for samples of size 30.

a) Which is which? Explain.

b) From the boxplot showing the data from one random sample, what does one value in the sample represent? How many values are included in the data to make the boxplot? Estimate the minimum and maximum values. Give a rough estimate of the mean of the values and use appropriate notation for your answer.

c) From the boxplot showing the data from a sampling distribution, what does one value in the sampling distribution represent? How many values are included in the data to make the boxplot? Estimate the minimum and maximum values. Give a rough estimate of the value of the population parameter and use appropriate notation for your answer.

Answers:

a)

b)

c)

Lock5 exercise 3.30 (1st edition): **Number of Screws in a Box** A company that sells boxes of screws claims that a box of its screws contains on average 50 screws ($\mu = 50$). Figure 3.10 shows a distribution of sample means collected from many simulated random samples of size 10 boxes.

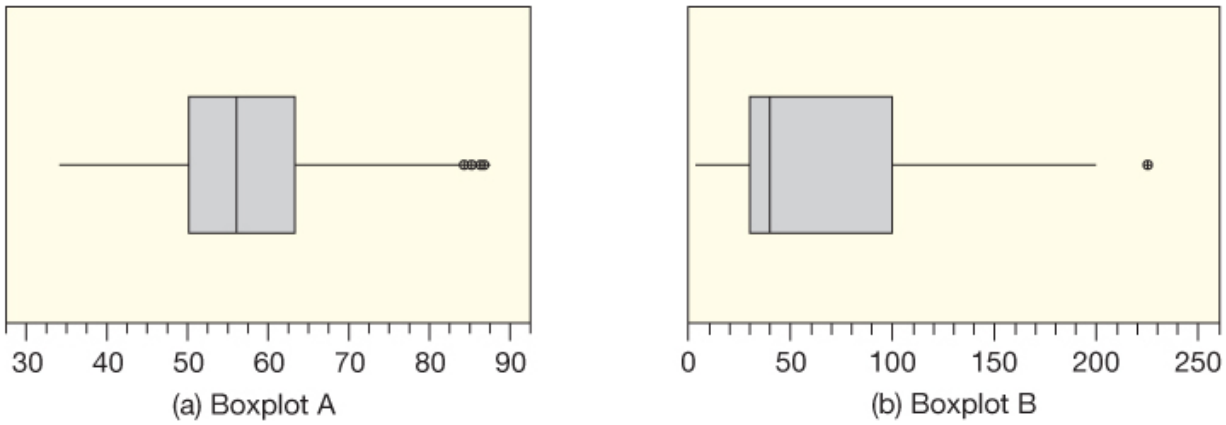

Figure 2: Figure 3.8

a) For a random sample of 10 boxes, is it unlikely that the sample mean will be more than 2 screws different from μ? What about more than 5? 10?

b) If you bought a random sample of 10 boxes at the hardware store and the mean number of screws per box was 42, would you conclude that the company's claim ($\mu = 50$) is likely to be incorrect ?

c) If you bought a random box at the hardware store and it only contained 42 screws, would you conclude that the company's clam is likely to be incorrect?

Answers:

a)

b)

c)

Lock5 exercise 3.52 (1st edition): **Employer-Based Health Insurance** A report from a Gallup poll in 20011 started by saying "Forty-five percent of American adults reported getting their health insurance from an employer. . . " Later in the article we find information on the sampling method, "a random sample of 147,291 adults, aged 18 and over, living in the US," and a sentence about the accuracy of the results, "the maximum margin of sampling error is $\pm$ 1 percentage point".

a) What is the population? What is the sample? What is the population parameter of interest? What is the relevant static?

b) Use the margin of error to give an interval estimate for the parameter of interest. Interpret it in terms of getting =health insurance from an employer.

Answers:

a)

b)

Lock5 exercise 3.60 (1st edition): **Effects of Overeating for One Month: Average Long-Term Weight Gain** Overeating for just four weeks can increase fat mass and weight over two years later, a Swedish study shows. Researchers recruited 18 healthy and normal-weight people with an average age of 26. For a

four-week period, participants increased calorie intake by 70% (mostly by eating fast food) and limited daily activity to a maximum of 5000 steps per day (considered sedentary). Nor surprisingly, width and body fa of the participants went up significantly during the study and then decreased after the study ended. Participants are believed to have returned to the diet and lifestyle they had before the experiment. However, two and aa half years after the experiment, the mean weight gain for participants was 6.8 lbs with a standard error of 1.2 lbs. A control group that did not binge had no change in weight.

a) what is the relevant parameter?

b) How could we find the actual exact value of the parameter?

c) give a 95% confidence interval for the parameter and interpret it.

d) Give the margin of error and interpret it.

Answers:

a)

b)

c)

d)

Part 2: Examining sampling distributions and calucating SEs and CIs in R

In worksheet 1, exercise 1.5 you used the function get_sprinkle_sample(n) to get a (virtual) sample of sprinkle colors and then you calculated the proportion of red sprinkles in the sample. You then repeated this process for different sizes of n. In the exercises below we will now create a sampling distribution for the proportion of red sprinkles.

Exercise 2a: For the first exercise, let's just review calculating the proportin of red sprinkles. Use the get_sprinkle_sample() function to get a sample of size n = 100 and calculate the proportion of red sprinkles. Report this statistic using the proper symbol. Note: the table() and the prop.table() functions might be useful here (if you forget how to do this, please look back at your first worksheet!).

Answer:

Exercise 2b: Now use a for loop to create a sampling distribution of the proportion of red sprinkles. The each statistic in the sampling distribution should be based on a sample size of n = 100, and there should be 10,000 points in the sampling distribution. Then create a histogram of the sampling distribution and report the standard error and what the shape of the sampling distribution looks like.

Answer:

Exercise 2c Now repeat exercise 2b for sample sizes of n = 25 and n = 400. How does the standard error change as function of the sample size n? Does this make sense to you?

Answers:

Exercise 2d: Finally, draw one additional sample of size n = 100 and calculate the confidence interval based on this sample and the standard error calculated in exercise 2b. Then repeat this process of n = 25, and n = 400 using the standard errors calculated in exercise 2c. Report all these confidence intervals and whether they change as you expected. Also, what do you think the value of the parameter is, and do you think any of these confidence intervals do not contain this parameter?

Answer:

a) The confidence interval for n = 100 is:

b) The confidence interval for n = 25 is:

c) The confidence interval for n = 400 is:

My best guess at the parameter is...

Reflection

How did this worksheet go? Are you feeling ok with the concepts of sampling distributions, standard errors and confidence intervals?