Worksheet 7

Author:

Discussants:

==========

The following worksheet will give you practice understanding sampling distributions, bootstrap distributions and computing confidence intervals using the bootstrap. Please answers the following questions and submit a compiled pdf with your answers to Moodle by 11:59pm on Sunday October 28th.

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, μ_m, $\pm$ $\approx$

Part 1a: Using an web app to explore sampling distributions

In the following exercises you will use a web application to explore sampling and bootstrap distributions. You will also explore standard errors and confidence intervals computing from these distributions. The application to use is located at: https://asterius.hampshire.edu:3939/intro_stats/sampling_and_bootstrap_distributions/. Note, the application is a little buggy, so if it acts strangely please reload the web page.

To complete these exericses, open up the web application in another window. When the application is open you will see the following three plots:

a) The plot on the *upper left* shows the distribution of data in a population. The shape of the population distribution can be changed using the population distribution dropdown box, where the choice are: right skewed, bimodal and normal.

b) The plot on the *lower left* shows a histogram of the data from *one single sample* of size n. The size of the sample can be changed using the sample size input box.

c) The plot on the *upper right* shows an (approximate) sampling distribution (histogram) of statistics based on 5,000 samples (i.e., each statistic in the histogram was computed from one of the 5,000 samples). The type of statistic that goes into this distribution can be changed using the 'statistic' drop down box. Options for the statistics you can use are: the mean, the median and the standard deviation.

Exercise 1.1: Using the *mean* statistic and a sample size of $n = 100$, describe the shape of the *sampling distribution* for the:

a) the right skewed population distribution
b) the bimodal population distribution
c) the normal population distribution

Answer in the conclusion section below: does the shape of the *sampling distribution* change a lot depending on the shape of the *population distribution*?

Answers:

a) the shape of the sampling distribution is:

b) the shape of the sampling distribution is:

c) the shape of the sampling distribution is:

Conclusion:

Exercise 1.2: Using a sample size of n = 100 and a population distribution that is right skewed, describe the shape of the sampling distribution for the statistics of:

a) mean ($\bar{x}$)
b) standard deviation (s)
c) median

Answer in the conclusion section below: does the shape of the distribution change a lot depending on the statistic chosen?

Answers:

a) the shape of the sampling distribution is:

b) the shape of the sampling distribution is:

c) the shape of the sampling distribution is:

Conclusion:

Exercise 1.3: Keeping population distribution right skewed and the statistic being the sample mean $\bar{x}$. Compare the sample sizes of:

a) n = 20
b) n = 80
c) n = 320

In the answer section below describe the shape of the sampling distribution for these different sample sizes and what the standard errors (SE) are. Is this what you would expect? Why? Also, do you notice any relationship between the different standard errors?

Answers:

a) the shape of the sampling distribution is: SE =

b) the shape of the sampling distribution is: SE =

c) the shape of the sampling distribution is: SE =

Conclusion:

Part 1b: Using an web app to explore bootstrap distributions

For the next two questions check the box that says 'Display Bootstrap Distribution'. You should notice a new plot in the lower right that shows the bootstrap distribution that is created from the one sample in the lower left plot.

Exercise 1.4: The centers of these distributions are given by red lines on the plots, and the center values noted at the top of the 4 plots. To get a sense of the relationship between the center values between the plots, set the statistic to the mean, and explore relationships by changing the sample size to values around

100 which will generate: a) new a sample, b) a new sampling distribution and c) the resulting bootstrap distribution from the sample. Then answer the following questions

a) How does the population mean (μ) relate to the center of the sampling distribution ($E[\bar{x}]$)? Is there bias here?

b) How does the center of the sample ($\bar{x}$) related to the the population mean (μ)? Is this expected?
c) How does the center of the bootstrap distribution ($E[\bar{x}*]$) related to the center of the sample ($\bar{x}$)?

d) How does the center of the bootstrap distribution relate to the center of the population (μ)?

Answers:

a)

b)

c)

d)

Exercise 1.5: How does the standard error and confidence interval of the sampling distribution compare to the standard error and confidence interval created from the bootstrap distribution? To explore this, set the distribution to normal and the statistic to the mean, and run the following steps twice (i.e., my answers from running it once are in answer 1) fill in answers 2) and 3) by running the steps below twice):

a) Set the sample size to n = 99 and then reset it to n = 100. This will rerun the app to create a new sample and a new bootstrap distribution.

b) Write down the SE from the sampling distribution and the SE* from the bootstrap distribution.

c) Report the SE, $SE*$, and the value of $\bar{x}$. Also, using the statistic value from the one sample ($\bar{x}$), compute 95% confidence intervals from the sampling distribution (using SE) and using the bootstrap distribution ($SE*$) using the formulas: $CI = \bar{x} \pm 2 \cdot SE$ and $CI_b = \bar{x} \pm 2 \cdot SE*$.

Do all these intervals capture the population parameter?

Answers:

While the answer will diff, for the three different runs I got are:

1) $\bar{x}$ = 5.07 SE = 0.10 CI = [4.87 5.27] $SE*$ = 0.13 CI_b = [4.81 5.33]
2) $\bar{x}$ = SE = CI = [ ] $SE*$ = CI_b = []
3) $\bar{x}$ = SE = CI = [ ] $SE*$ = CI_b = [ ]

Do all three intervals capture the population parameter?

Answer:

Exercise 1.6: Repeat exercise 1.5 but using the right skewed population. Do all three intervals capture the population parameter?

a) Set the sample size to n = 99 and then reset it to n = 100. This will rerun the app to create a new sample and a new bootstrap distribution.

b) Write down the SE from the sampling distribution and the SE* from the bootstrap distribution.

c) Report the SE, $SE*$, and the value of $\bar{x}$. Also, using the statistic value from the one sample ($\bar{x}$), compute 95% confidence intervals from the sampling distribution (using SE) and using the bootstrap distribution ($SE*$) using the formulas: $CI = \bar{x} \pm 2 \cdot SE$ and $CI_b = \bar{x} \pm 2 \cdot SE*$.

Do all these intervals capture the population parameter?

Answers:

While the answer will diff, for the three different runs I got are:

1) $\bar{x}$ = 1.91 SE = 0.14 CI = [1.63 2.19] $SE*$ = 0.13 CI_b = [1.65 2.17]
2) $\bar{x}$ = SE = CI = [ ] $SE*$ = CI_b = [ ]
3) $\bar{x}$ = SE = CI = [ ] $SE*$ = CI_b = [ ]

Do all three intervals capture the population parameter?

Answer:

Part 2: calculating confidence intervals using the bootstrap in R

In worksheets 2 and 3 we examined the sale prices of cars for a few different brands that were sold on Edmunds.com. In the exercises below you will calculate confidence intervals for the mean price that Mazda's and Toyota's are sold for.

Exercise 2.1: Let's start by looking at the price of Mazda's. The code below loads the data on the prices of cars and extracts a sample of prices that different Mazda's were sold for. Calculate the size of the sample using the length() function and store the answer in an object called n_sample_size. Also calculate the mean price that Mazda's were sold for and store the result in an object called the_stat. Report the value for the sample size and the mean price Mazda's were sold for and use the appropriate symbol for the mean Mazda price.

```
# load the car price data
load("/home/shared/intro_stats/cs206_data/cars_small.Rda")

# extract the sample of Mazda prices
mazda_price_sample <- subset(car_data$price, car_data$brand ==
    "Mazda")


# get the samples size and store it in an object called
# n_sample_size



# get the mean price and store it in an object called
# the_stat
```

Answer:

Exercise 2.2: Now use a for loop to create a bootstrap distribution by: 1) sampling with replacement n_sample_size sites from mazda_price_sample 2) calculating the mean of the bootstrap sample and storing it in a vector named bootstrap_dist 3) repeating this processing 10,000 times.

Then plot a histogram of the bootstrap distribution, and calculate the bootstrap standard error $SE*$. Finally, calculate a 95% confidence interval using the formula: $CI_b = \bar{x} \pm 2 \cdot SE*$ and report what the interval is below.

Answer:

Exercise 2.3 Below is code that extracts the prices for a number of Toyota's that are were also sold on Edmunds.com. Repeat steps 2.1 and 2.2 with the toyota price sample. Report the following:

1) The mean price of Toyota's based on the sample.
2) The sample size (n)
3) The bootstrap standard error $SE*$
4) The 95% confidence interval based on the statistic and the bootstrap standard error
5) Answer the question: could the population mean price for Mazda's (μ_{mazda}) be the same as the population mean price for Toyota's (μ_{toyota})? Why or why not?

```
# a sample of toyota prices
toyota_price_sample <- subset(car_data$price, car_data$brand ==
    "Toyota")
```

Answers:

1)

2)

3)

4)

5)

Reflection

How did this worksheet go? Are you feeling ok with the concepts and code related calculating confidence intervals using the bootstrap?

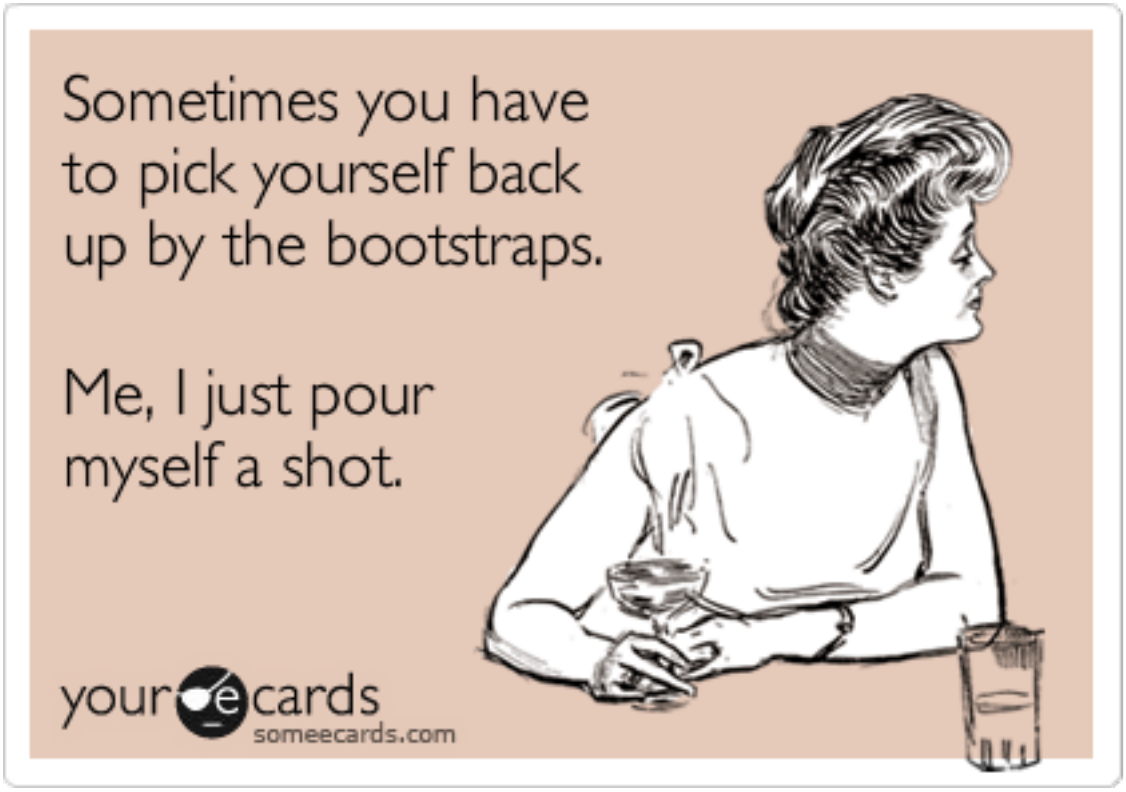

Figure 1: Bootshot