Lab 02C: Week 7

Contents

1 Quick quiz

1.1 Confidence intervals

1.2 Bootstrap

2 Group work

3 Questions

3.1 Speed of light

3.2 Cotinine

3.3 Cereal

4 For practice after the computer lab

4.1 Further reading

4.2 Three blind mice

The **specific aims** of this lab are:

- practice using the bootstrap to construct confidence intervals
- find p-values using permutation tests
- generate discussion and provide an opportunity to practise *statistical thinking* and *communicating statistical concepts*

The unit **learning outcomes** addressed are:

- LO1 Formulate domain/context specific questions and identify appropriate statistical analysis.
- LO3 Construct, interpret and compare numerical and graphical summaries of different data types including large and/or complex data sets.
- LO5 Identify, justify and implement appropriate parametric or non-parametric statistical tests.
- LO8 Create a reproducible report to communicate outcomes using a programming language.

1 Quick quiz

1.1 Confidence intervals

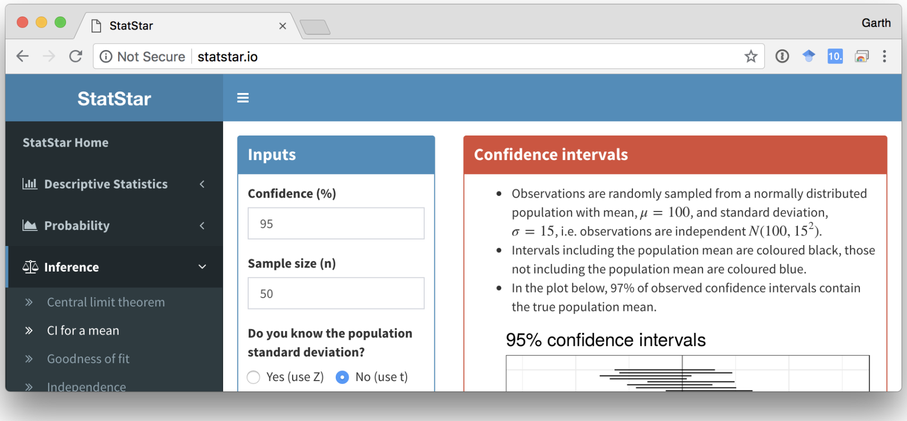
You can explore these ideas at http://statstar.io selecting `Inference` on the left and then

`CI for a mean.`

1. If I increase the sample size the confidence interval will get **wider/narrower** (holding all else constant).

2. If I increase the level of confidence (or equivalently decrease α) the confidence interval will get **wider/narrower**.

3. If I do 100 experiments under identical conditions and calculate 95% confidence intervals for each experiment then exactly 95 of those confidence intervals will contain the true population mean. **True/False**



Inputs

Confidence (%)

95

Sample size (n)

50

Do you know the population standard deviation?

○ Yes (use Z) ● No (use t)

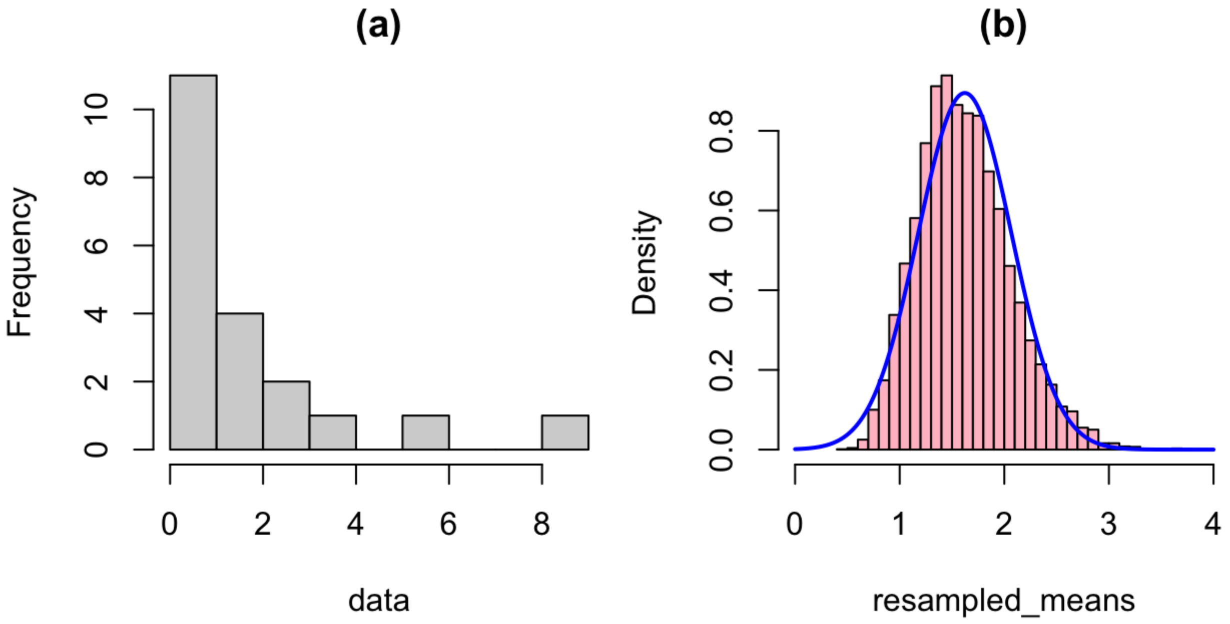
Confidence intervals

- Observations are randomly sampled from a normally distributed population with mean, $\mu = 100$, and standard deviation, $\sigma = 15$, i.e. observations are independent $N(100, 15^2)$.
- Intervals including the population mean are coloured black, those not including the population mean are coloured blue.
- In the plot below, 97% of observed confidence intervals contain the true population mean.

95% confidence intervals

1.2 Bootstrap

A researcher wanted to empirically estimate the distribution of the sample mean in the data she had collected. To do this she decided to use bootstrap resampling. Below is a histogram of (a) the raw data and (b) the resampled means with the normal approximation curve.



Which of the following statements is the least correct

a. The distribution of the resampled means is right skewed.

b. The upper bound of the confidence interval derived from the bootstrapped resamples should be greater than the upper bound derived from a normal approximation.

c. The mean of the resampled means should be greater than the median of the resampled means.

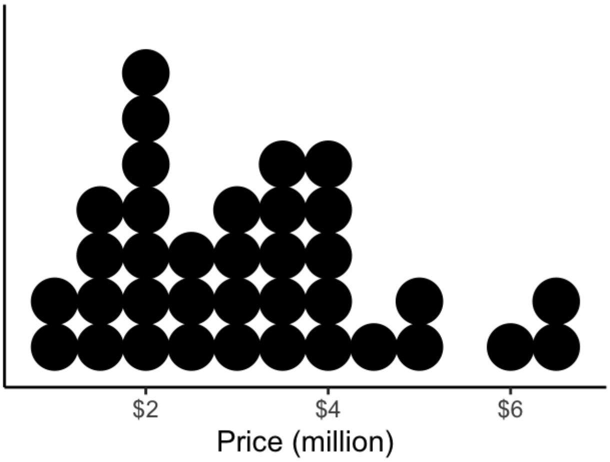
d. The skew of the resampled means is in the same direction as the skew of the observed data.

e. The lower bound of the confidence interval derived from the bootstrapped resamples should be less than the lower bound derived from a normal approximation.

2 Group work

We would like to report the average house price for a street of 36 houses. The prices of these house are give in the table below (in millions of dollars).

Prices	1	2	3	4	5	6
1	2.1	2.4	5.9	2.8	2.9	6.4
2	3.4	1.4	1.9	2.2	5.0	3.3
3	3.3	2.9	2.1	6.6	3.5	1.0
4	3.9	2.1	1.6	2.4	1.6	1.9
5	2.0	1.2	4.1	2.9	1.5	5.1
6	3.4	2.3	4.3	4.2	4.1	3.8



The true mean of the houses in the street is $3.1 million. Unfortunately, you only have money to buy price estimates on 6 houses.

1. Using two rolls of a dice, buy 6 unique housing reports (i.e. without replacement). If you don’t carry dice with you you can use `sample(1:6, 2, replace = TRUE)` to simulate rolling two

dice. Use these two numbers to identify a random house price in the table above (numbers 1-6 across the top and down the side). Repeat 6 times to get a sample of six prices. (If you hit the same house, roll again.) Report the mean of your sample to the rest of the class. Your tutor will provide a link to an online spreadsheet to aggregate the results.

- What does the distribution of the class' sample means look like?

2. Using your individual sample, estimate the distribution of the sample mean using bootstrap resampling. Use 200 bootstrap resamples. Calculate a bootstrap confidence interval for the population mean average house price of the street.

- Does the true population mean lie in this confidence interval?

- Does your bootstrap distribution of the sample mean look like the distribution of the means collected from your classmates?

3 Questions

3.1 Speed of light

In the lecture, we discussed a famous dataset where Simon Newcomb measured the time required for light to travel from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a total distance of about 7400 meters. You can download the data file that contains 66 sets of measurements used to estimate the speed of light.

Getting started: Read in the data and check the size of your data.

```
library(tidyverse)
speed_file =
          read_csv("https://raw.githubusercontent.com/DATA2002/data/master/speed_of_light.tx

speed = speed_file$Speed_of_Light
ggplot(speed_file, aes(x="", y = Speed_of_Light)) +
  geom_boxplot(colour = "red", outlier.size = 4) +
  theme_classic(base_size = 16) +
  labs(x = "", y = "Speed") + coord_flip()
```

a. Generate one bootstrap sample and compare this sampled data with the original data.

b. Compute the mean and median of the bootstrap sample and compare with the corresponding values in the original data.

c. Draw another bootstrap sample and repeat the comparison. Repeat this 20 times and see if your conclusion changes. Inspect first ten bootstrap estimates of the mean. Visualise the result. *Hint:* Write a for loop (see example in lecture notes).

d. Typically one draws a large number of bootstrap samples, say 1000 or more. Try different numbers of bootstrap samples and see how the shape of the histogram changes.

e. Find a 95% bootstrap confidence interval for the mean using the 2.5 and 97.5 percentiles as the confidence limits. Compare this with a "traditional" confidence interval that uses the t-distribution.

f. Generate 95% bootstrap confidence intervals calculation for the median and the MAD.

3.2 Cotinine

A variant of nicotine found in cigarettes is cotinine (which, not coincidentally, is an anagram of nicotine). It is found in the blood stream and the amount is proportional to the amount of exposure a person has to tobacco smoke. Therefore, cotinine is used as an indicator of tobacco smoke exposure. For example, cotinine levels < 10 ng/ml are considered to be consistent with no active smoking; values between 10 - 100 ng/ml are associated with light smoking, or moderate passive exposure while heavy smokers have at least 300 ng/ml. Levels in active smokers typically reach 500 ng/ml.

The following data lists the cotinine levels of 40 passive smokers who are not smokers of tobacco products.

```
x = c(0, 87, 173, 253, 1, 103, 173, 265, 1, 112, 198, 266, 3, 121, 208,
      277, 17, 123, 210, 284, 32, 130, 222, 289, 35, 131, 227, 290, 44, 149,
      234, 313, 48, 164, 245, 477, 86, 167, 250, 491)
```

1. Calculate some simple descriptive measures of the data, construct a histogram and a QQ plot. Provide a brief description of the sample data.

2. Based on your descriptive summary of the data, do you think there are any outlying, or unusually large, observations that may impact upon any inferential test that you perform? In your description, take into consideration the summary statistics and histogram of the remaining data.

3. Using R and the complete sample, perform a standard t test of the hypotheses $H_0\colon \mu = 130$ vs $H_1\colon \mu \neq 130$. At the 5% level of significance, what can you conclude about the cotinine levels of the smokers in the population?

4. Perform a sign test to test $H_0\colon \mu = 130$ vs $H_1\colon \mu \neq 130$.

4. Perform a permutation test by generating 10,000 resamples. What conclusion do you reach based on the permutation test p-value?

5. Which of the procedures above provides the more appropriate inference of the population mean. Why?

6. Generate 95% bootstrap confidence intervals for the standard deviation and the median absolute deviation from the median (MAD). Plot histograms of both bootstrap distributions. Which of these two estimators of scale is more reliable in this setting.

7. Say we were interested in the coefficient of variation, $CV = s/\bar{x}$, for this data set. Generate a 90% bootstrap confidence interval for the coefficient of variation

bootstrap confidence interval for the coefficient of variation.

3.3 Cereal

This cereal data is taken from Data Description, Inc. (2021) and can be downloaded from https://raw.githubusercontent.com/DATA2002/data/master/Cereal.csv.

One of the variable `mfr` represents the manufacturer of cereal where `A` = American Home Food Products, `G` = General Mills, `K` = Kelloggs, `N` = Nabisco, `P` = Post, `Q` = Quaker Oats, `R` = Ralston Purina.

Please note that the code provided should be used as a guide only and not complete solutions.

Getting started: Read the data from the website. Check the size of your data. Think about what the number of rows actually means.

```
library(tidyverse)
cereal = read_csv("https://raw.githubusercontent.com/DATA2002/data/master/Cereal.csv",
    na = "-1")
# if you've downloaded it to your computer cereal =
# read_delim('Cereals.txt', delim = '\t', na = '-1') Looking at the
# start of the data
dplyr::glimpse(cereal)
```

1. Produce some basic summary statistics the nutrients “sugar” and “sodium”
2. Restricting attention to `G` = General Mills and `K` = Kelloggs cereals, visualise the distribution of sodium content between the two manufacturers. Does it look like there is equal variance between the two groups? Could you safely assume normality within each group?
3. Perform a permutation test to test whether there is a significant difference in the mean sodium content between the two manufacturers.

4 For practice after the computer lab

4.1 Further reading

For permutation testing see Curley and Milewski (2020) sections 13.1 and 13.3.

4.2 Three blind mice

Three blind mice are put on a diet. Their weights before and after the diet are {2.9, 2.4, 3.1} and {2.8, 2.2, 2.8} respectively. Using a permutation test and the test statistic $\bar{d}$, where d_i are the observed weight changes, is there any evidence that the three blind mice lost weight on the diet at a significance level of $\alpha = 0.1$

References

Curley, James P., and Tyler M. Milewski. 2020. *Psy317l Guidebook*. https://bookdown.org/curleyjp0/psy317l_guides/.

Data Description, Inc. 2021. *The Data and Story Library*. https://dasl.datadescription.com/.