DATA 8

Fall 2020

Tutoring Section 12

Sample Means, Designing Experiments

Slides by Kevin Miao

Today

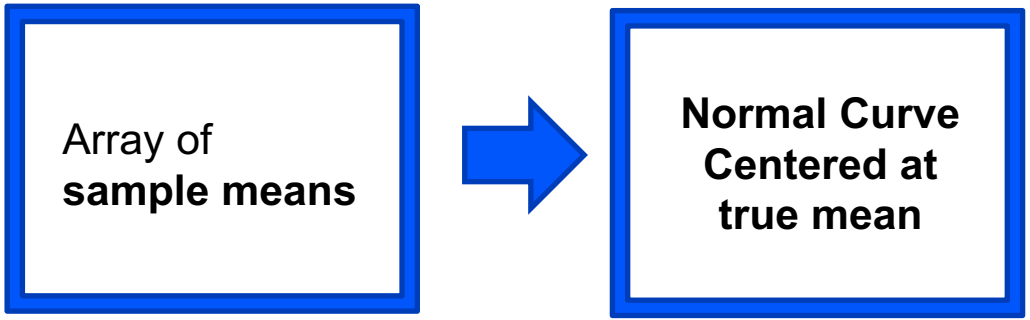
- The Variability of the Sample Mean (continued)
- Designing Experiments and Choosing Sample Size
- Bootstrap/CI Review
  - **Tutoring Worksheet**
  - **Final Fall ‘19**

Worksheet

Link: **https://tinyurl.com/d8tutweek12**

Variability of Sample Mean

The **standard deviation** of these **sample means** depends on the **sample size!**



- SD of Sample Means = $\frac{Population\ SD}{\sqrt{Sample\ Size}}$
- Population size **does not** influence SD of sample means

Remember from last week?

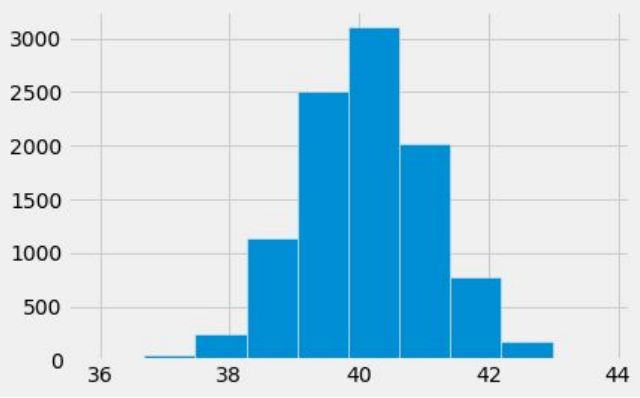
Practice Problems

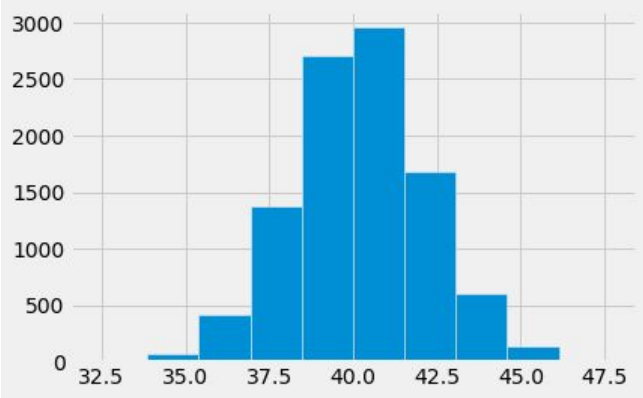
4.1 Suppose you simulate the proportion of purple-flowered plants in a sample of 200 plants (from Mendel's 75% purple- and 25% white-flower plant population) using `sample_proportions` 1000 times. Then, you plotted distribution of the proportion of purple-flowered plants from each of the 1000 trials. What would this distribution look like? Where would the distribution be centered?

- CLT only applies to **means** or **sums**
- Here we worked with **proportions** and CLT still applied?
- **Proportions** are **means**
- purple_flowers = [True, False, True, False … False, True]
- What is **np.mean(purple_flowers)**?

Q1.1

1.1 Suppose we have a population with a mean of 40 and an SD of 10. One of the histograms below is an empirical distribution of the means of 10000 bootstrap resamples each of size 100 from the population. Which histogram?





Q1.2-1.3

1.2 Fill in the blank: Based on the population from the previous question, there is a _____% chance that a random resample has a mean that lies within the range [38, 42].

1.3 Suppose a redwood forest has trees whose average height are 200 feet with an SD of 30 feet. A random sample of 400 trees is taken. Fill in the blank: There is a 68% chance that the average height of the sample lies within the range 200 plus or minus _____.

Designing Experiments

- **CLT**
  - **SD of Sample Means** = $\frac{Population\ SD}{\sqrt{Sample\ Size}}$
  - Sample Means Normally Distributed
- **95% Confidence Interval / Bootstrappng**
  - Resample sample many times
  - Width of the middle 95%
- **Chebyshev for normal distributions**
  - Average ± 2SD corresponds to **95%**

$$\textbf{Width of 95\% CI} = 4 \times \frac{Population\ SD}{\sqrt{Sample\ Size}}$$

Designing Experiments

$$\textbf{Width of 95\% CI} = 4 \times \frac{Population\ SD}{\sqrt{Sample\ Size}}$$

- **Cool**, what's the utility of this function?
- **Imagine this:**
  - Goal: Need to know **sample size** for 95% CI
  - Infer Population SD from Sample SD
  - Solve formula for **sample size**
  - E.g. you want to know how many people to survey for a certain experiment!

Q2.1

Let's say you want to poll the population of UC Berkeley students to ask whether they like vanilla ice cream or chocolate ice cream. You can only take a sample, but you want to estimate the population proportion of students who like vanilla ice cream. Let's say you need your estimate to have a confidence interval width of at most 0.05.

2.1 Suppose the population SD of the proportion of students who like vanilla ice cream is 0.1. What sample size do you need to achieve a 95% confidence interval width of **at most** 0.05?

Q2.2-2.3

2.2 Is it possible to calculate what sample size you need if you don't know the population SD? If not, can we bound what the population SD could be?

2.3 Suppose you **do not** know the population SD of students who like vanilla ice cream. What sample size do you need to achieve a 95% confidence interval of width **at most** 0.05?

Q3 on worksheet (optional)

Tonight is the Monster Mash. We're trying to determine the median scariness level of ghosts. We are given a sample of ghosts in the form of a one column table, `spooky_sample`, that contains 200 numbers each of which describe how scary a ghost is on a scale from 0 to 10. You can assume that the sample is a simple random sample from the population of all ghosts.

3.1 Fill in the code below to create a function that computes a 95% confidence interval for the median scariness level in the population of ghosts. Assume `spooky_sample` is a one column table with scariness levels of the ghosts in our sample.

```
def candy_cornfidence_interval(spooky_sample, replications):
   result_medians = ______________________
      for i in ______________________________:
         resample_median = ___________________________________
         median = ____________________________________________
         result_medians = ____________________________________
   left_end = ____________________________
   right_end = ___________________________
   return ____________________________
```

3.2 If we run the function you wrote above multiple times, will it always return the same interval? Why or why not?

Q3 on worksheet (optional)

3.3 If we consider the population of all ghosts, there exists a median scariness level. We call this the “true median scariness level” of the population. Recall that since we don’t have access to the population, we don’t have access to the true median scariness level either.
If we were to compute 100 confidence intervals with the function from 3.1, how many of those confidence intervals would we expect to capture the true median scariness level?

3.4 If we picked out one of the 100 confidence intervals from the previous question and found that it was [5.6, 6.8], what is the probability that this interval contains the true median scariness level?

Exam Prep (Fall 19) – 9a

9. (30 points) Foreign Aid

Enji, an International Relations major, is writing his Masters thesis on aid given to foreign governments by the World Bank. He finds a sample of donations given to various countries over the last decade and collects these findings into a table called `aid`. Here are the first few rows.

Date	Recipient	Amount	Purpose
May 24, 2016	Zambia	595,321	agriculture
Aug 13, 2012	India	2,571,991	rail
Dec 4, 2018	Bangladesh	1,633,020	agriculture
Dec 16, 2019	Turkey	510,410	manufacturing

The table contains four columns:

- **Date**: a string, the date upon which the donation was made
- **Recipient**: a string, the country receiving the money
- **Amount**: an int, the amount of the donation in USD
- **Purpose**: a string, the reason listed for the aid

(a) (2 pt) To get a sense of the data, Enji first plots a histogram of the aid `'Amount'` in his sample. He finds that the empirical distribution of `'Amount'` has an average of \$3,532,423 and an SD of \$1,121,240. The distribution of `'Amount'` in his sample is:

- ○ Approximately normal
- ○ Not approximately normal
- ○ There isn't enough information to answer this question

Exam Prep (Fall 19) – 9b

9. (30 points) Foreign Aid

Enji, an International Relations major, is writing his Masters thesis on aid given to foreign governments by the World Bank. He finds a sample of donations given to various countries over the last decade and collects these findings into a table called `aid`. Here are the first few rows.

Date	Recipient	Amount	Purpose
May 24, 2016	Zambia	595,321	agriculture
Aug 13, 2012	India	2,571,991	rail
Dec 4, 2018	Bangladesh	1,633,020	agriculture
Dec 16, 2019	Turkey	510,410	manufacturing

The table contains four columns:

- **Date**: a string, the date upon which the donation was made
- **Recipient**: a string, the country receiving the money
- **Amount**: an int, the amount of the donation in USD
- **Purpose**: a string, the reason listed for the aid

(b) (4 pt) Suppose Enji wants to use his sample data to create a 95% confidence interval of the true average amount of aid of all donations. If the distribution of all World Bank donations has an SD of $1,000,000 and the `aid` table contains 10,000 rows, can Enji create a 95% confidence interval that has a width less than $25,000?

Note: an interval of [-5,5] has a width of 10.

- ⬭ He can because the sample size is large enough
- ⬭ He can't because the sample size is too small
- ⬭ There isn't enough information to answer this question

Exam Prep (Fall 19) – 9c

(c) **(3 pt)** As Enji is combing through the data set, he notices that some countries in South Asia appear to have received a disproportionate amount of aid with the purpose of `'rail'` and `'manufacturing'` compared to others in the region. He creates the following table, which displays the aid given to countries in the region with the following proportions. For example, the last column tells us that of the aid that Pakistan received from the World Bank, 20% was for agriculture, 20% was for rail, and 60% was for manufacturing. Note that each country's column adds up to 1.

Purpose	India	Bangladesh	Pakistan
agriculture	0	0.9	0.2
rail	0.4	0.1	0.2
manufacturing	0.6	0	0.6

According to the above distributions, what is the empirical total variation distance of aid `'Purpose'` between India and Bangladesh? You may leave your answer as a mathematical expression (not Python).

Exam Prep (Fall 19) – 9d

(d) (8 pt) The World Bank claims the total variation distance of aid `'Purpose'` between India and Bangladesh is 0.3. Enji is not sure if his empirical TVD (from part (a)) is different from 0.3 just due to chance, but he thinks he could bootstrap his sample to get a better idea.

Complete the code below to write a function `purpose_tvd` that takes in a table `tbl` with the same column labels as `aid`, two country names, `country_a` and `country_b`, and computes the total variation distance between the two countries' `'Purpose'` distributions.

For example, `purpose_tvd(aid, 'Bangladesh', 'India')` should return your answer from part (c).

```
def purpose_tvd(tbl, country_a, country_b):

    dist_a = tbl.where(_____).____________________

    counts_a = dist_a.sort('Purpose').____________________

    dist_b = tbl.where(_____).____________________

    counts_b = dist_b.sort('Purpose').____________________

    props_a = counts_a / np.sum(counts_a)

    props_b = counts_b / np.sum(counts_b)

    return _____* np.sum(abs(_____))
```

Exam Prep (Fall 19) – 9e

(e) **(7 pt)** Next, complete the code below to simulate 500 bootstrap samples (bootstrap a sample of India and Bangladesh independently), compute the total variation distance between the `'Purpose'` distributions of the aid received by Indian and Bangladesh in each bootstrap sample, and store all of the results in the array `boot_tvds`. You may assume that `purpose_tvd` has been defined correctly.

```
boot_tvds = make_array()

for _____:

    india = aid.where(_____)

    bangladesh = aid.where(_____)

    boot_india = india._____(_____)

    boot_bangladesh = bangladesh._____(_____)

    boot_tbl =_____.append(_____)

    new_tvd = _____

    boot_tvds = _____
```

Exam Prep (Fall 19) – 9fg

(f) **(3 pt)** Finally, complete the code below to compute an approximate 95% confidence interval for the population total variation distance between the `'Purpose'` distributions of the aid received by Indian and Bangladesh. After the code is executed, `left` should store the left endpoint of our interval and `right` should store the right endpoint. You may assume that `boot_tvds` has been computed correctly.

```
left = _____

right = _____
```

(g) **(3 pt)** Suppose it turns out that the values `left` and `right` are 0.24 and 0.78, respectively, so the confidence interval in part (f) is $[0.24, 0.78]$. Suppose we test whether or not the World Bank actually followed its claims in distributing aid to India and Bangladesh (i.e. the TVD is actually 0.3), using this confidence interval and a 5% cutoff for the P-value. Pick **ALL** the correct ways to complete the sentence:

The test will conclude that the purpose distributions of aid received by India and Bangladesh

- ◯ are the same as the World Bank's claims
- ◯ are different from the World Bank's claims
- ◯ could be the same as the World Bank's claims
- ◯ probably are different from the World Bank's claims

End of Section

- Please complete the anonymous Feedback form so I can improve my teaching:
  - **https://tinyurl.com/feedbackD8Kevin**
- Solutions and notes will be posted after Wednesday.
- Email me if you have any questions: kevinmiao@berkeley.edu