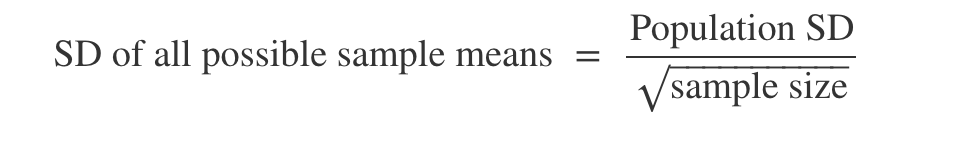
Week 12 : Designing Experiments, More CLT

Data 8 Tutoring

# 1 The Variability of the Sample Mean

## Key Concepts

**The SD of the Sample Mean**

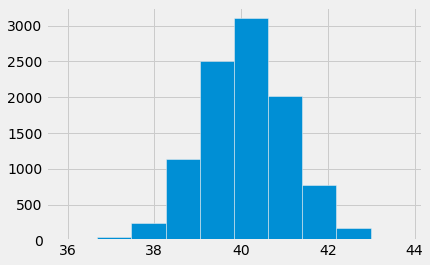
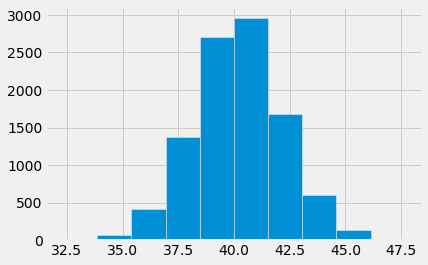
****

This is the standard deviation of the averages of all the possible samples that could be drawn. **It measures roughly how far off the sample means are from the population mean.** The smaller the SD, the more accurate the estimate.

**Proportions are Means**

* Proportions are the means of a list of 1s and 0s (they tell you the fraction of 1s in the list)--this means that the CLT applies to them!
* Oftentimes, you’ll work with data where the population is binary. For example, you might have polling data where the answer to a question is “Yes” or “No”. You might want to estimate the sample proportion of “Yes” votes.
* The SD of all possible sample proportions, calculated using the formula above, is **at most 0.5 regardless of the proportion of 1s in the population**.

## Practice Problems

**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?

The left one. We know that these distributions are approximately normal because they’re distributions of sample means. The SD, based on the information given about the population, should be 10 / sqrt(100) = 1. The left one seems to have an SD of 1.

**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].

95%. We know from properties of the normal distribution that there is a 95% chance that the data lie within 2 SDs of the mean.

**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 \_\_\_\_\_\_\_\_\_.

This is just the SD of the sample means, 30/sqrt(400) = 30/20.

# 2 Designing Experiments and Choosing Sample Size

## Key Concepts

**Choosing the Sample Size of an Experiment**

* Sometimes, you’ll need to conduct an experiment and estimate a population parameter (like a mean) up to a certain accuracy--this accuracy is usually measured by the width of the confidence interval.
* Ultimately this means that we want to limit the variability of our estimate. We know from the CLT that the variability of the sample mean is affected by the sample size!
* For a normal distribution, the “middle 95%” is within 2 SDs of the mean. We can use this information to create a confidence interval: the center ± 2 SDs.
* Then, the width of this confidence interval is:
* The term being multiplied is the SD of all the sample means--if the bounds of our 95% confidence interval are 2 SDs to the left of the mean and 2 SDs to the right, then the width is 4 SDs total!
* We can use this equation above to calculate the sample size we need for an experiment.

## Practice Problems

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?

4 \* 0.1 / sqrt(sample size) <= 0.05

0.4 <= 0.05 \* sqrt(sample size)

0.4 / 0.05 <= sqrt(sample size)

(0.4 / 0.05)^2 <= sample size → sample size >= 64

**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?

We need the population SD to calculate what sample size we need, but most of the time, we don’t have it! However, we can replace it with the maximum population SD of a list of 0s and 1s, which is 0.5.

**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?

Since you don’t have the population SD, you can replace it with 0.5.

4 \* 0.5 / sqrt(sample size) <= 0.05

2 <= 0.05 \* sqrt(sample size)

2 / 0.05 <= sqrt(sample size)

(2 / 0.05)^2 <= sample size → sample size >= 1600

# 3 (Optional) Confidence Interval Review

## Practice Problems

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 = make\_array()

for i in np.arange(replications):

resample\_median = spooky\_sample.sample()

median = percentile(50, resample\_median.column(0))

result\_medians = np.append(result\_medians, median)

left\_end = percentile(2.5, result\_medians)

right\_end = percentile(97.5, result\_medians)

return make\_array(left\_end, right\_end)

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

No, it will not always return the same interval! The confidence interval is calculated based on values from the empirical distributions, which are randomly generated each iteration.

**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?

We would expect to see the true median scariness level in 95 out of 100 of the confidence intervals.

**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?

We can’t say what the probability is that this *particular* interval contains the true median. There is no chance involved--the true median is either in the interval, or it isn’t!