DSFA

Spring 2021

Lecture 25

Sample Averages

Announcements

- Project 2, Part 1 due today at 5:59PM
- Part 2 due Friday 4/16 at 5:59PM
- Prelim 2, Tuesday 4/20, 8:30-10PM here for all Ithaca-resident students
    - Coverage: Through Lecture 26, Monday 4/12
    - Review session Saturday 4/17, 3:30-5:30PM
    - More on Monday

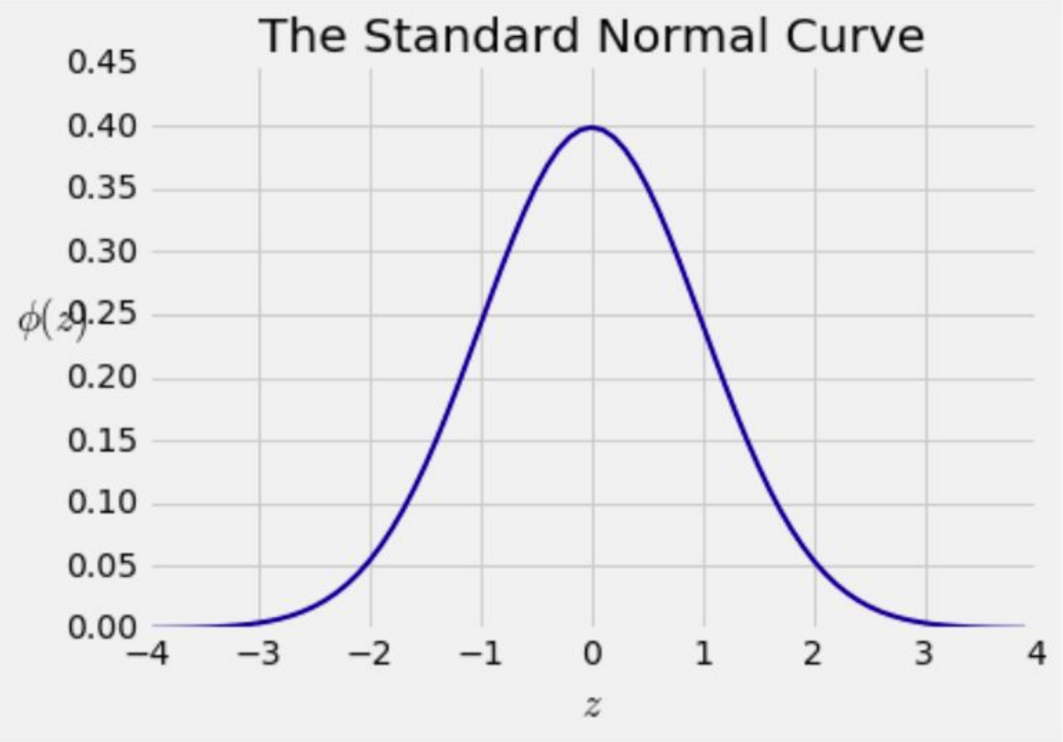
Questions for This Week

- How can we quantify natural concepts like “center” and “variability”?

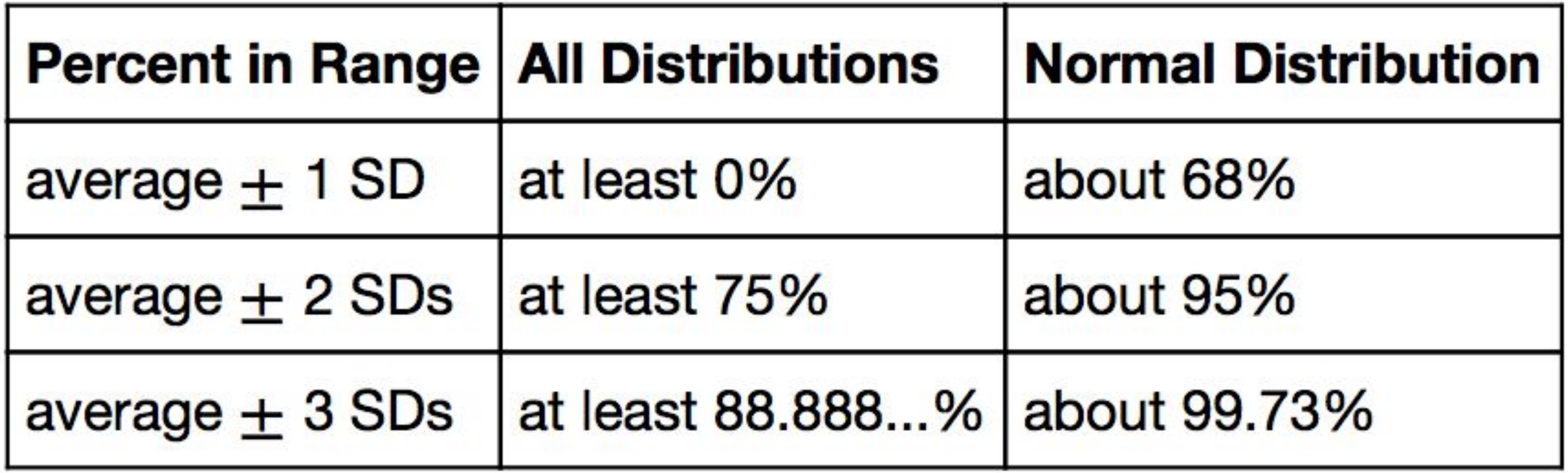
- Why do many of the empirical distributions that we generate come out bell shaped?

- How is sample size related to the accuracy of an estimate?

Bell Curve


The Standard Normal Curve
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
$\phi(z)$
−4
−3
−2
−1
0
1
2
3
4
z

Bounds and Normal Approximations



Percent in Range	All Distributions	Normal Distribution
average ± 1 SD	at least 0%	about 68%
average ± 2 SDs	at least 75%	about 95%
average ± 3 SDs	at least 88.888...%	about 99.73%

Central Limit Theorem

If the sample is

- large, and
- drawn at random with replacement,

Then, *regardless of the distribution of the population,*

the distribution of the sample sum
(or of the sample average) is roughly bell-shaped

(Demo)

Sample Averages

- The Central Limit Theorem describes how the normal distribution (a bell-shaped curve) arises in the context of random sampling.
- Most distributions we observed were not bell-shaped, but empirical distributions of sample averages were.
- We care about sample averages because they estimate population averages.

Distribution of the Sample Average

Why is There a Distribution?

- You have only one random sample, and it has only one average.
- But **the sample could have come out differently**.
- And then the sample average might have been different.
- So there are many possible sample averages.

Distribution of the Sample Average

- Imagine all possible random samples of the same size as yours. There are lots of them.

- Each of these samples has a mean.

- The **distribution of the sample average** is the distribution of the means of all the possible samples.

Shape of the Distribution

Central Limit Theorem

If the sample is

- large, and
- drawn at random with replacement,

Then, *regardless of the distribution of the population,*

the distribution of the sample sum (or of the sample average) is roughly bell-shaped

(Demo)

Specifying the Distribution

Suppose the random sample is large.

- We have seen that the distribution of the sample average is roughly bell shaped.

- Important questions remain:
  - Where is the center of that bell curve?
  - How wide is that bell curve?

Center of the Distribution

The Population Average

The distribution of the sample average is roughly a bell curve centered at the population average.

Variability of the Sample Average

Why Is This Important?

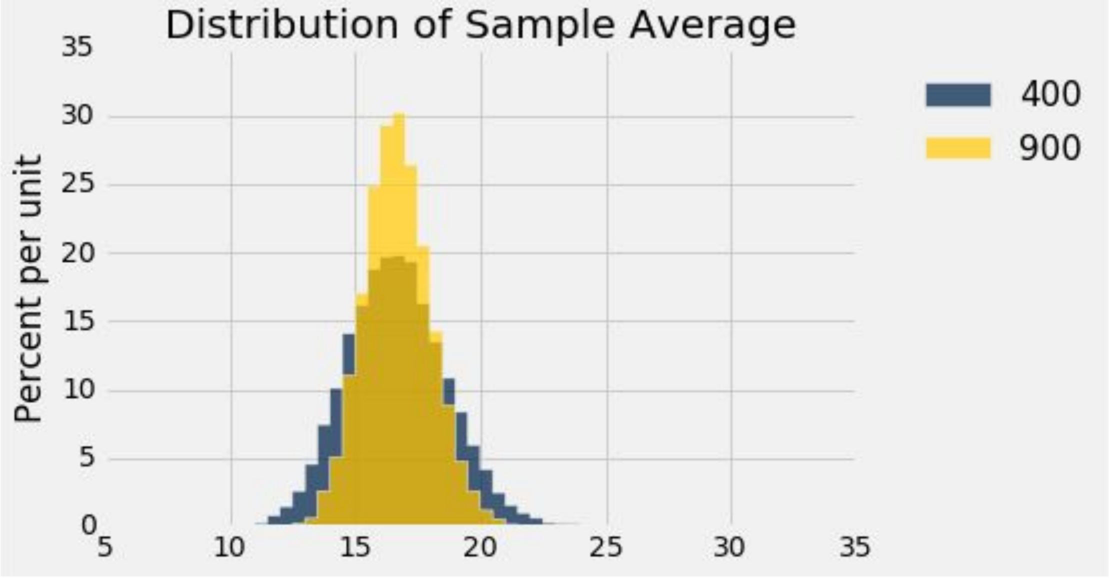
- Along with the center, the spread helps identify exactly which normal curve is the distribution of the sample average.
- The variability of the sample average helps us measure how accurate the sample average is as an estimate of the population average.
- If we want a specified level of accuracy, understanding the variability of the sample mean helps us work out how large our sample has to be.

(Demo)

Discussion Question

The gold histogram shows the distribution of ___________ values, each of which is ___________________________.

(a) 900
(b) 10,000
(c) a randomly sampled flight delay
(d) an average of flight delays



The Two Histograms

- The gold histogram shows the distribution of 10,000 values, each of which is an average of 900 randomly sampled flight delays.
- The blue histogram shows the distribution of 10,000 values, each of which is an average of 400 randomly sampled flight delays.
- Both are roughly bell shaped.
- The larger the sample size, the narrower the bell.

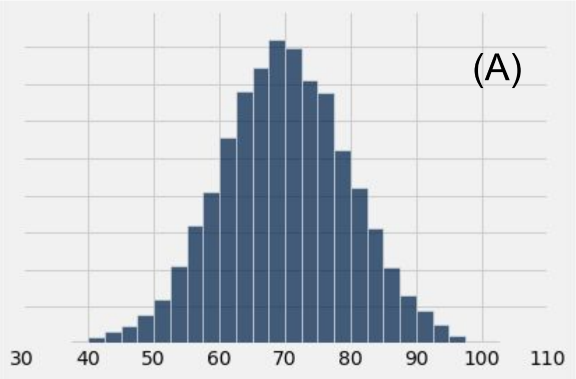
(Demo)

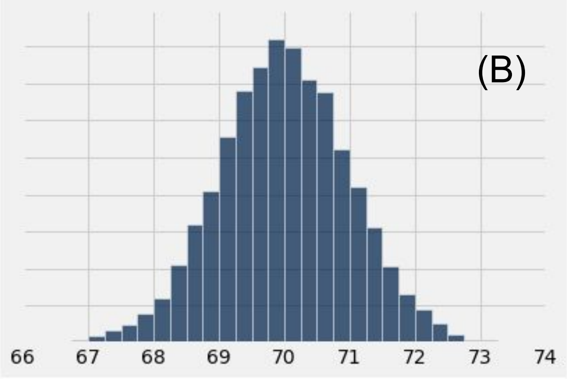
Variability of the Sample Average

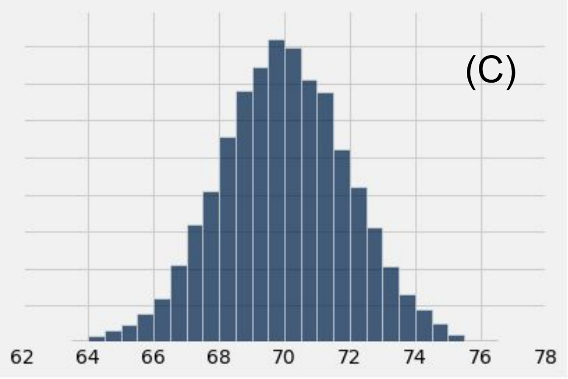
- Fix a large sample size.
- Draw all possible random samples of that size.
- Compute the average of each sample.
- You’ll end up with a lot of averages.
- The distribution of those is called the *distribution of the sample average.*
- It’s roughly normal, centered at the population average.
- SD = (population SD) / $\sqrt{\text{sample size}}$

Discussion Question

A population has average 70 and SD 10. One of the histograms below is the empirical distribution of the averages of 10,000 random samples of size 100 drawn from the population. Which one?







When poll is active, respond at **pollev.com/dsfa**

Text **DSFA** to **22333** once to join

Which distribution?