DATA 8
Spring 2020

Lecture 27

The Normal Curve and Sample Means

Weekly Goals

- Monday
  - Describing distributions: “center” and “spread”
  - How big are most of the values?
- **Today**
  - The bell shaped curve and its relation to large random samples
- Friday
  - The variability in a random sample average
  - Choosing the size of a random sample

Announcements

DATA 9

Happy April 1!

Review: Standard Units

- How many SDs above average?
- ***z = (value - average)/SD***
    - Negative z: value below average
    - Positive z: value above average
    - z = 0: value equal to average
- When values are in standard units: average = 0, SD = 1
- Gives us a way to compare/understand data no matter what the original units

The SD and the Histogram

- Usually, it's not easy to estimate the SD by looking at a histogram.
- But if the histogram has a bell shape, then you can.

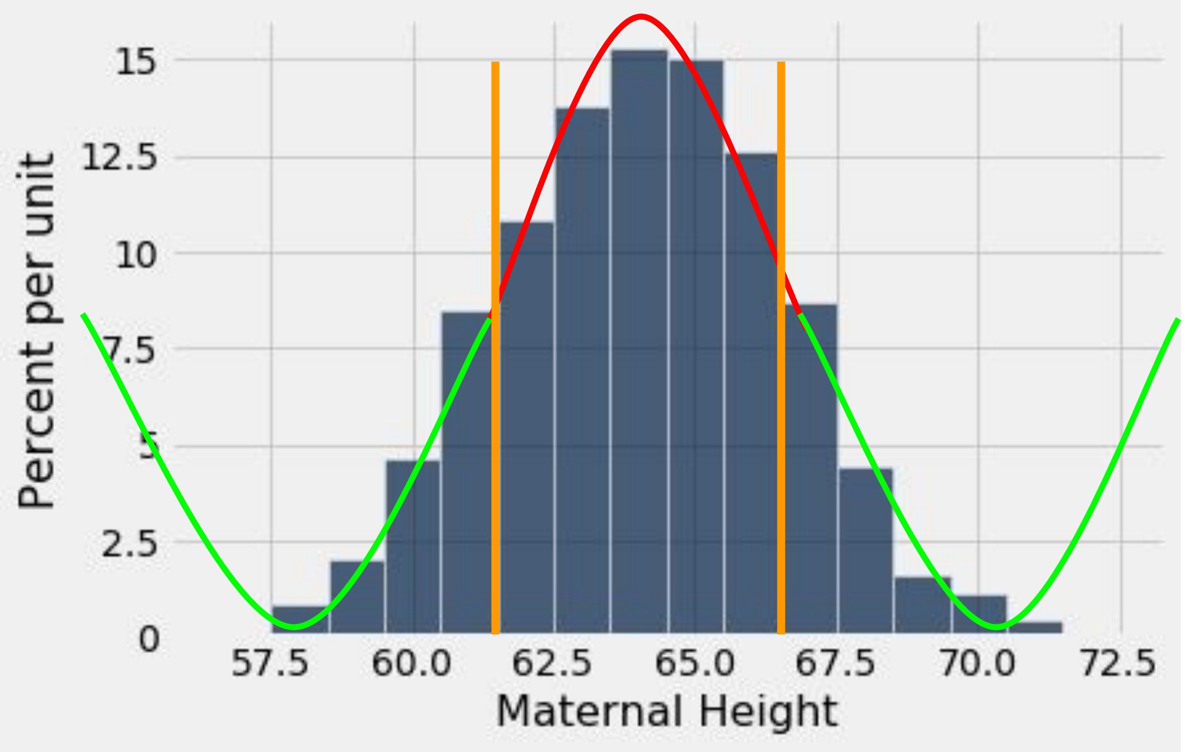
The SD and Bell-Shaped Curves

If a histogram is bell-shaped, then

- the average is at the center
- the SD is the distance between the average and the points of inflection on either side

(Demo)

Point of Inflection



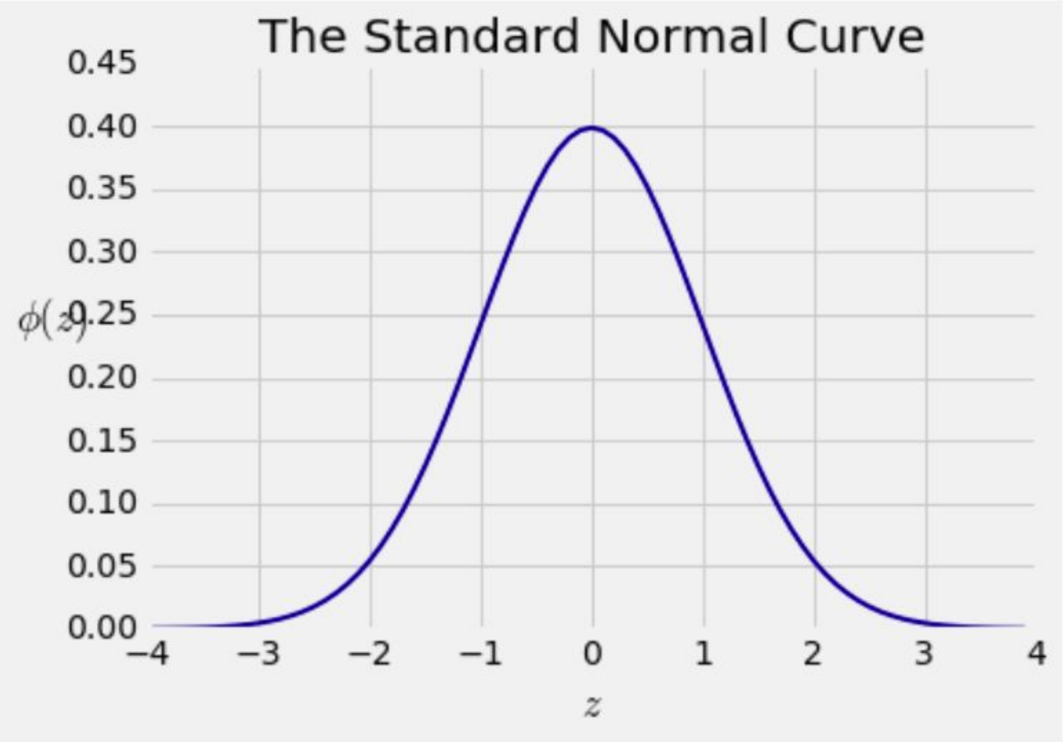
The Normal Distribution

The Standard Normal Curve

A beautiful formula that we won’t use at all:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \qquad -\infty < z < \infty$$

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

Normal Proportions

How Big are Most of the Values?

No matter what the shape of the distribution,
the bulk of the data are in the range “average ± a few SDs”

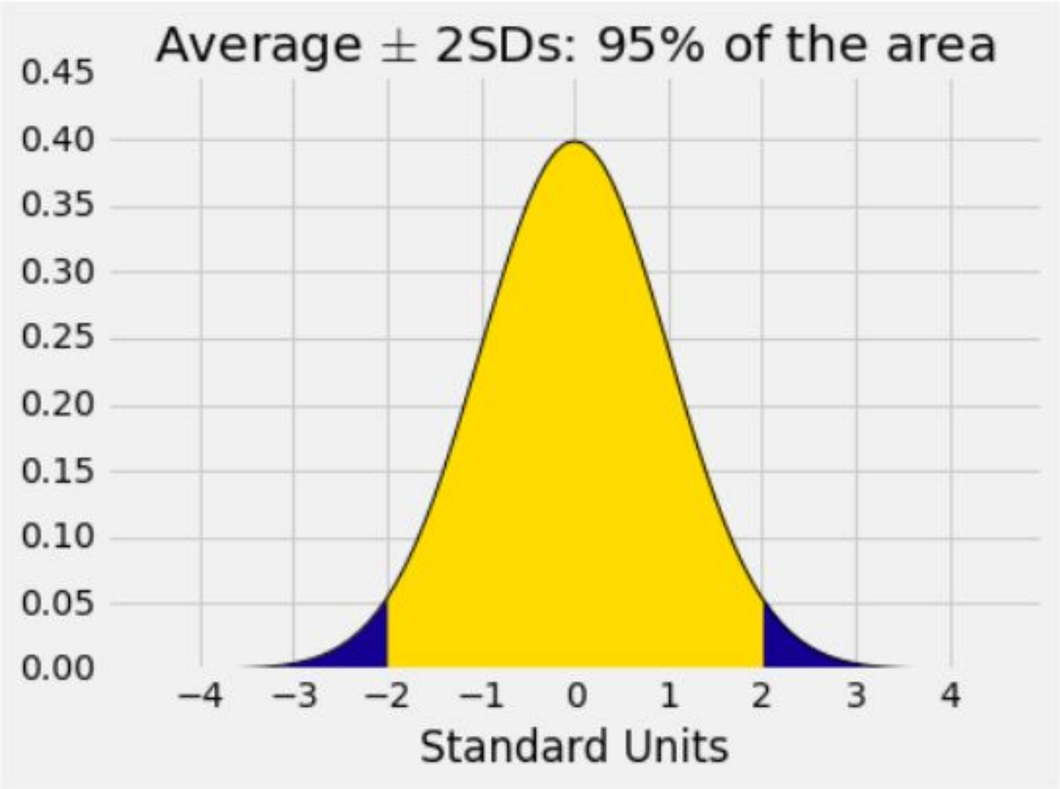
If a histogram is bell-shaped, then

- Almost all of the data are in the range “average ± 3 SDs”

Bounds and Normal Approximations

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

A "Central" Area



Central Limit Theorem

Sample Averages

- The Central Limit Theorem describes how the normal distribution (a bell-shaped curve) is connected to random sample averages.
- We care about sample averages because they estimate population averages.

Central Limit Theorem

If the sample is

- large, and
- drawn at random with replacement,

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

the probability distribution of the sample sum (or the sample average) is roughly normal

(Demo)

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 an average.

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

(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?

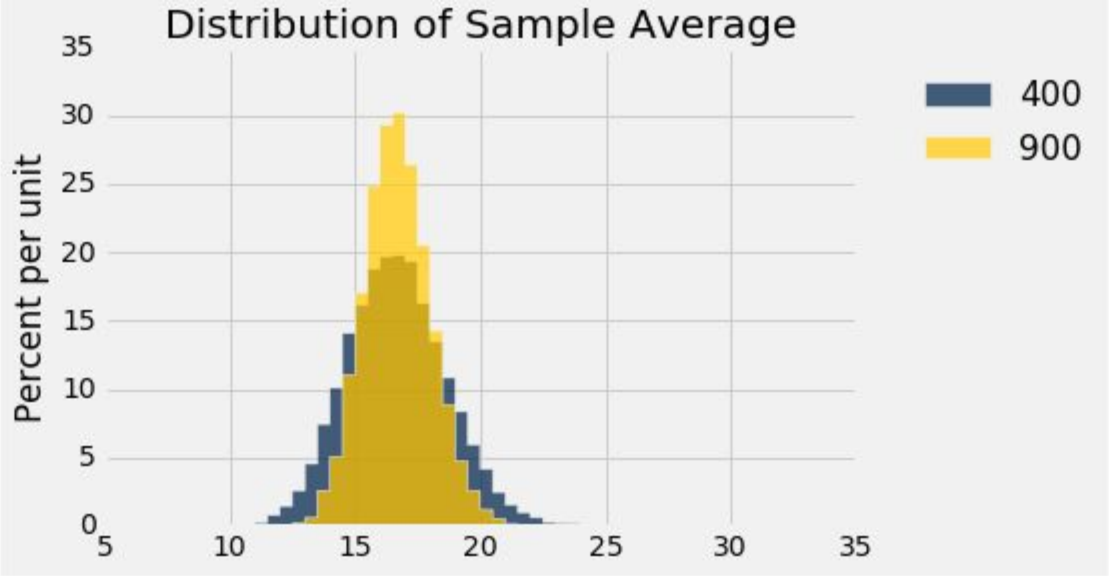
- Along with the center, the spread helps identify exactly which normal curve is the distribution of the sample average.
- The variability of the sample average helps us measure how accurate the sample average is as an estimate of the population average.
- If we want a specified level of accuracy, understanding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

- The distribution of all possible sample averages of a given size is called the *distribution of the sample average.*
- We approximate it by an empirical distribution.
- By the CLT, it's roughly normal:
  - Center = the population average
  - SD = (population SD) / $\sqrt{\text{sample size}}$

(Demo)

Discussion Question

A city has 500,000 households. The annual incomes of these households have an average of $65,000 and an SD of $45,000. The distribution of the incomes [pick one and explain]:

(a) is roughly normal because the number of households is large.
(b) is not close to normal.
(c) may be close to normal, or not; we can't tell from the information given.