YaleNUSCollege

YSC2239 Lecture 10

Recap

- A/B testing
- Confidence Intervals
- Significant level (also called alpha level)

Python command

- Percentile

Today’s class

- Central and Spread
- Central limit theorem
- Correlation

- Reading: Chapter 14, 15

Confidence Intervals For Testing

Using a CI for Testing

What if we want to do a hypothesis test, but we can't simulate under the null?

- Null hypothesis: **Population average = *x***
- Alternative hypothesis: **Population average ≠ *x***
- Cutoff for P-value: *p*%
- Method:
  - Construct a (100-*p*)% confidence interval for the population average
  - If *x* is not in the interval, reject the null
  - If *x* is in the interval, can't reject the null

Center and Spread

Questions

- 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?

Average

The Average (or Mean)

Data: 2, 3, 3, 9 **Average = (2+3+3+9)/4 = 4.25**

- Need not be a value in the collection
- Need not be an integer even if the data are integers
- Somewhere between min and max, but not necessarily halfway in between
- Same units as the data
- Smoothing operator: collect all the contributions in one big pot, then split evenly

(Demo)

Comparing Mean and Median

- **Mean:** Balance point of the histogram
- **Median:** Half-way point of data; half the area of histogram is on either side of median
- If the distribution is symmetric about a value, then that value is both the average and the median.
- If the histogram is skewed, then the mean is pulled away from the median in the direction of the tail.

Standard Deviation

Defining Variability

Plan A: "biggest value - smallest value"

- Doesn't tell us much about the shape of the distribution

Plan B:

- Measure variability around the mean
- Need to figure out a way to quantify this

(Demo)

How Far from the Average?

- Standard deviation (SD) measures roughly how far the data are from their average

- SD = root mean square of deviations from average

- SD has the same units as the data

Why Use the SD?

There are two main reasons.

- **The first reason:**

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

- **The second reason:**

Coming up in the next lecture.

Standard Units

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
- Chebyshev: At least 96% of the values of *z* are between -5 and 5 (i.e.: average – 5*SD , average + 5*SD)

(Demo)

Discussion Question

Find whole numbers that are close to:

(a) the average age

(b) the SD of the ages

(Demo)

Age in Years	Age in Standard Units
27	-0.0392546
33	0.992496
28	0.132704
23	-0.727088
25	-0.383171
33	0.992496
23	-0.727088
25	-0.383171
30	0.476621
27	-0.0392546

... (1164 rows omitted)

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.

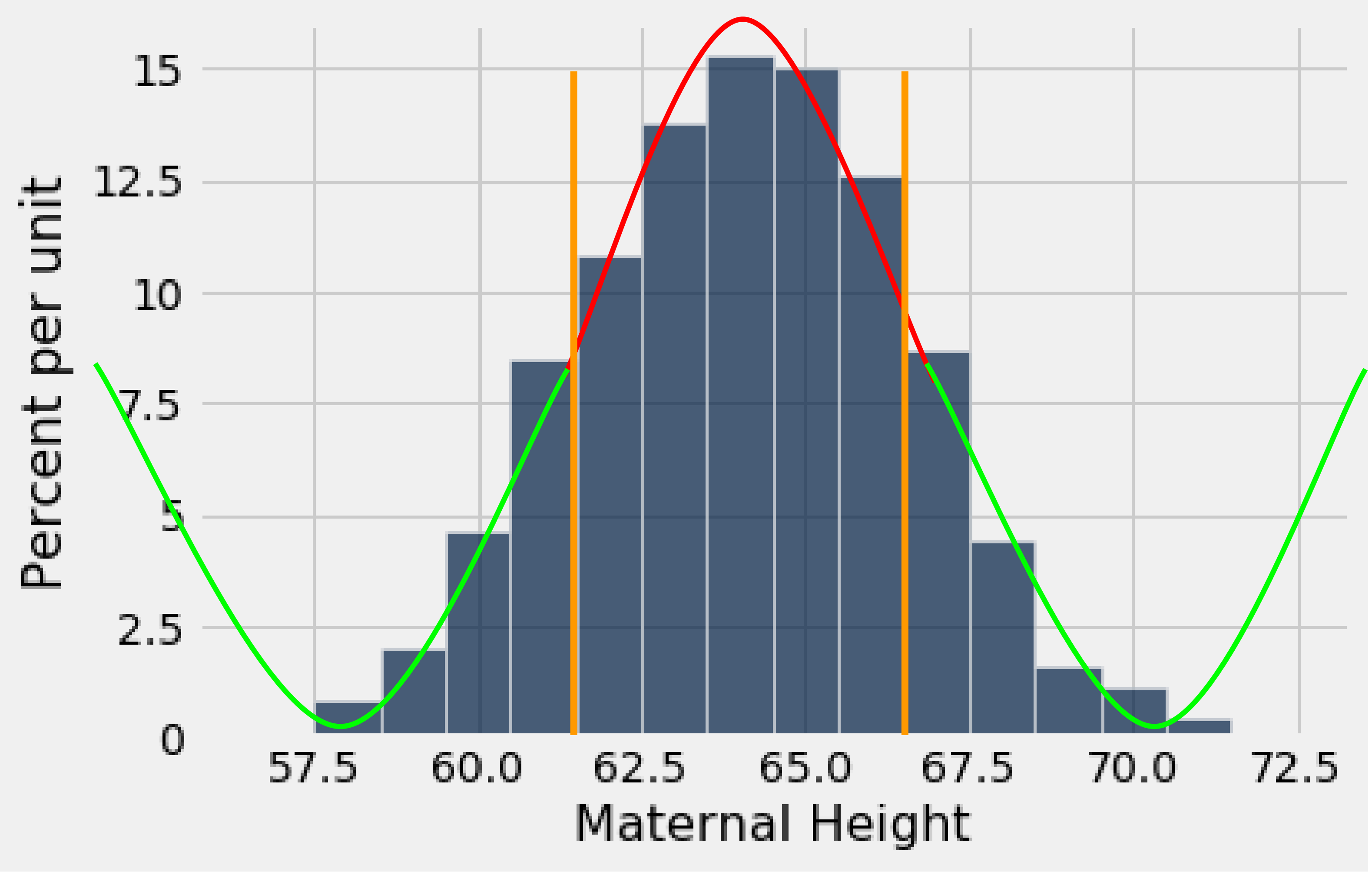
The SD and Bell-Shaped Curves

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



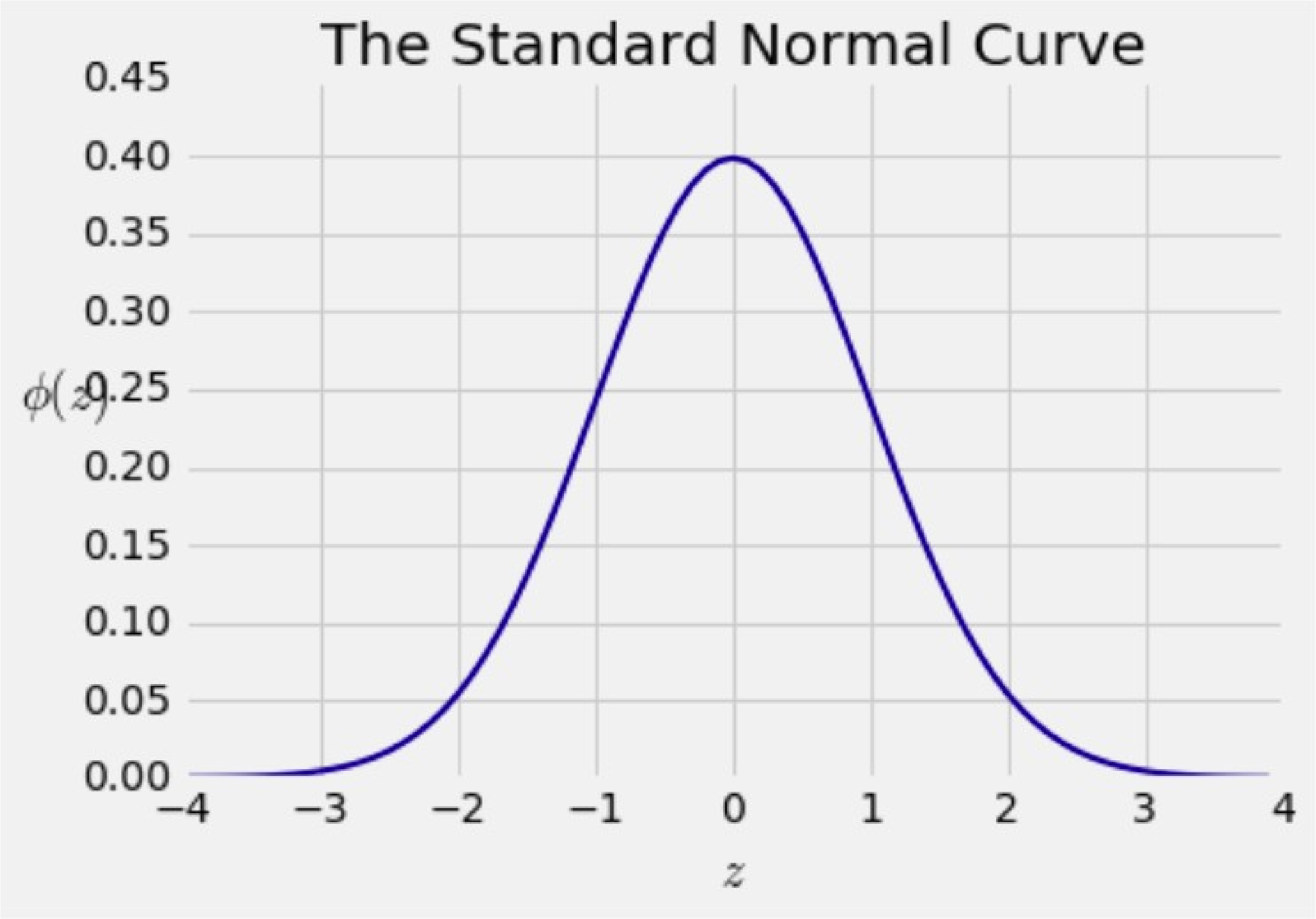
The Normal Distribution

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"

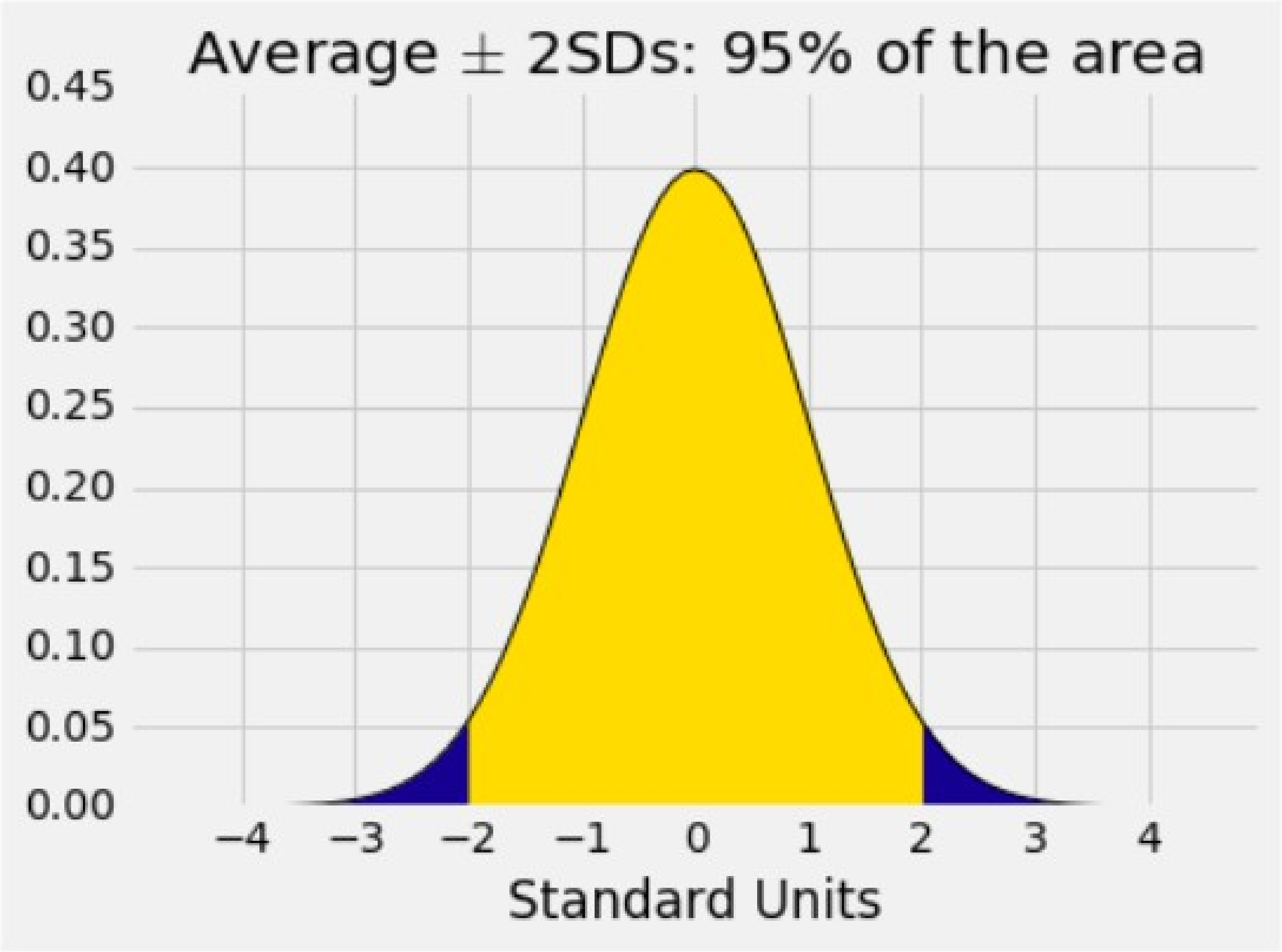
If a histogram is bell-shaped, then

- 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)

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) / √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.

Correlation Coefficient

Definition of *r*

Correlation Coefficient (*r*) =

average of	product of	x in standard units	and	y in standard units

Measures how clustered the scatter is around a straight line

The Correlation Coefficient *r*

- Measures **linear** association
- Based on standard units
- $-1 \leq r \leq 1$
  - $r = 1$: scatter is perfect straight line sloping up
  - $r = -1$: scatter is perfect straight line sloping down
- $r = 0$: No linear association; *uncorrelated*

(Demo)

Watch Out For ...

- Nonlinearity
- Outliers
- Correlation does not imply causations (https://www.tylervigen.com/spurious-correlations)

To-do

- Lab 5
- Assignment 5