DSC 10, Spring 2018
Lecture 19

Sample Means

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Review: Variability

Deviation from the Mean

- Standard deviation measures how far the data values are spread from the mean.
    - SD = root mean square of deviations from average
- Standard units measure how many standard deviations above average.
    - z = (value - mean)/SD
- Most of the data falls within a few standard units of the mean.
    - Chebyshev's inequality gives lower bound

Chebyshev's Bounds

Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)
average ± 3 SDs	at least 1 - 1/9 (88.888...%)
average ± 4 SDs	at least 1 - 1/16 (93.75%)
average ± 5 SDs	at least 1 - 1/25 (96%)

No matter what the distribution looks like

(Demo)

Normal Proportions

The Normal Distribution

Every bell-shaped curve is called "the normal distribution"

- The average (center) could be different
- The standard deviation (spread) could be different
- These two numbers alone determine the whole shape

How Big are Most of the Values?

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

If a histogram is bell-shaped,
almost all of the data falls 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%

(Demo)

Central Limit Theorem

Sample Averages

- The Central Limit Theorem describes scenarios in which the normal distribution (a bell-shaped curve) arises
- Most data distributions we observed were not bell-shaped, *but* empirical distributions of sample averages were bell-shaped.
  - Sample averages estimate population averages
  - A proportion within a sample is a sample average

Sample Proportions are Sample Averages

Say, I keep a record of days that I wore jeans:

- "1" - indicates that I wore jeans
- "0" - indicates that I did not wear jeans

This is my data for a week: [1, 1, 1, 0, 0, 1, 1]

- On what proportion of the days did I wear jeans?
- What is the average of these 0's and 1's?

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 average (or sample sum or sample proportion) is roughly bell-shaped**

Variability of the Sample Mean

Repeated Sampling

- The purpose of repeated sampling is to understand how a statistic could have been different
- If the statistic is an average of a large random sample, then CLT says the statistic is drawn from a bell curve
- Important questions remain:
  - Where is the center of that bell curve?
  - How wide is that bell curve?

(Demo)

Variability of the Sample Mean

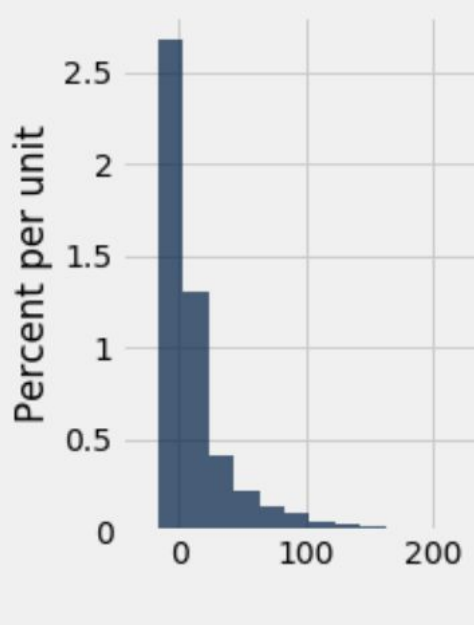
- Fix a large sample size
- Draw *all* possible random samples of that size
- Compute the mean of each sample (lots of them)
- The distribution of those is the *probability distribution of the sample mean*
- It's a normal distribution, centered at the population mean

sample mean's average = population average

sample mean's SD = (population SD) / $\sqrt{\text{sample size}}$

Discussion Question 1

Sampling from the flight delay distribution.



If you repeatedly compute the mean from a sample size of **1,** what will be the shape of the probability histogram?

A. Impossible to predict

B. Bell - shaped

C. Resembles the original histogram

(Demo)

Discussion Question 2

Population: Incomes with mean \$10,000 and SD \$20,000

Sample: 100 chosen uniformly at random with replacement

What's the chance that the sample average is above \$14,000?

A. 2.5%
B. 37%
C. 75%
D. I need a hint

sample mean's average = population average

sample mean's SD = (population SD) / $\sqrt{\text{sample size}}$

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%

Discussion Question 2: Solution

Population: Incomes with mean \$10,000 and SD \$20,000

Sample: 100 chosen uniformly at random with replacement

Question: What's the chance that the sample average is above \$14,000?

- SD of sample mean = population SD / $\sqrt{\text{sample size}}$
  = \$20,000 / 10
  = \$2,000
- \$14,000 is 2 SD above the population mean
- About 95% are within 2 SD of the population mean
- About **2.5% are above**; about 2.5% are below

Discussion Question 3

Population: A perfect bell shape. Mean 10; SD 20

Sample: 100 chosen uniformly at random with replacement

What's the chance that *all* are below 50?

A. 2.5%
B. 95%
C. 97.5%
D. None of the above
E. I need a hint

sample mean's average = population average

sample mean's SD = (population SD) / $\sqrt{\text{sample size}}$

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%

Discussion Question 3: Solution

Population: A perfect bell shape. Mean 10; SD 20
Sample: 100 chosen uniformly at random with replacement
Question: What's the chance that *all* are below 50?

- 50 is 2 population SD above the population mean
- The chance of drawing one value below 50 is 97.5%
- The chance of drawing 100 below 50 is **0.975 ** 100**

Discussion Question 4

You want to estimate the height of the tallest person on campus. You sample 100 people at random and compute a 99.9999% confidence interval using the bootstrap. Its upper bound is 6'4".

A 6'5" person walks by! What might have gone wrong?

A. Standard deviation of the population is too large to estimate
B. Sample size is too small for 99.9999% confidence interval
C. Height of tallest person is difficult to estimate with bootstrap
D. Empirical distribution of height of tallest person is not bell-shaped
E. More than one of the above