DATA 8

Fall 2020

Tutoring Section 11

Sample Means, Center/Spread, Normal Distribution

Slides by Kevin Miao

Logistics

- **Clarification:** Confidence Intervals (Piazza)
  - *p% Confidence Intervals*
    - If we sample 100 times from the population and we take a 95% CI, **only** then we see that the **true parameter** is captured by ~ 95 of the confidence intervals.
    - If we have **one** sample and we bootstrap (resample) it 100 times, then we are not sure about it.
      - If sample is **representative**, ~95 CIs will capture true parameter
      - If sample is **bad**, fewer will capture the true parameter
- Per usual:
  - **Feedback Form:** https://tinyurl.com/feedbackD8Kevin

All resources can be found on kevin-miao.com

Today

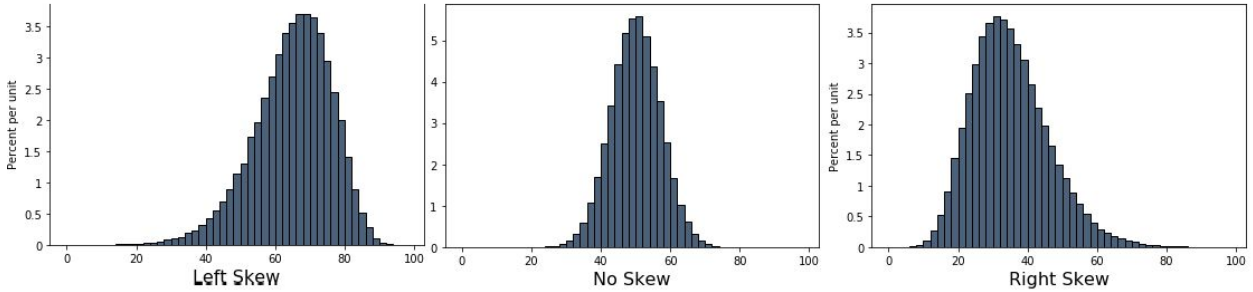
- **Busy day**
- Mean/Median
- Variability
  - *Standard Deviation and Variance*
- Standard Deviation and Normal Curves
- Central Limit Theorem
- Variability of Sample Mean

Worksheet

Link: **https://tinyurl.com/d8tutweek11**

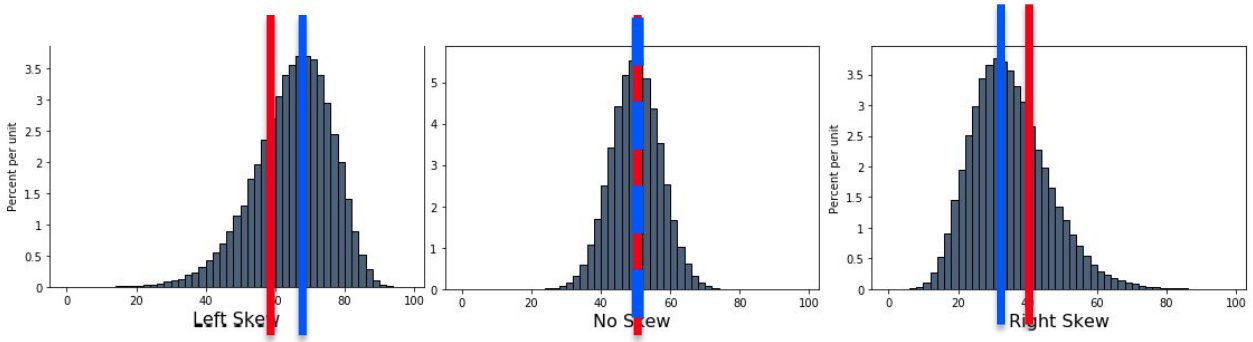
Mean/Median

- **Mean:** The sum of all elements divided by the total number of elements in the collection.
  - Analogy: Seesaw and the balance point.
- **Median**: 50th percentile of the graph



Mean/Median

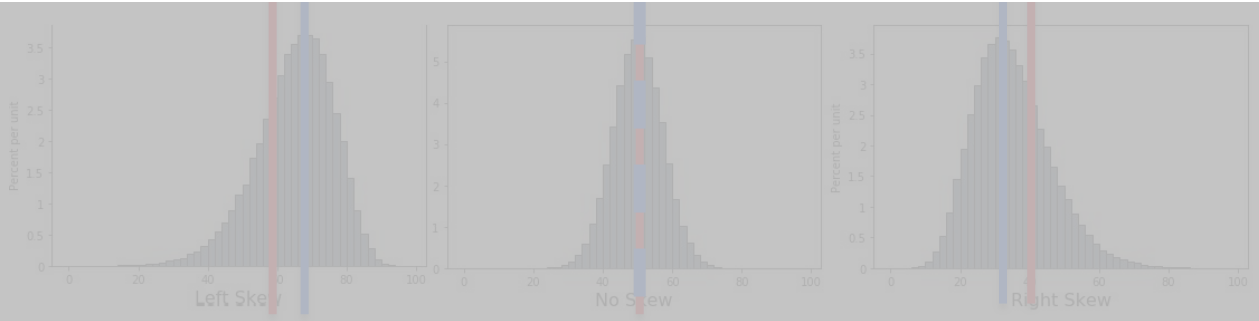
- **Mean:** The sum of all elements divided by the total number of elements in the collection.
  - Analogy: Seesaw and the balance point.
- **Median**: 50th percentile of the graph



Properties - Mean/Median

- **Mean:** The sum of all elements divided by the total number of elements in the collection

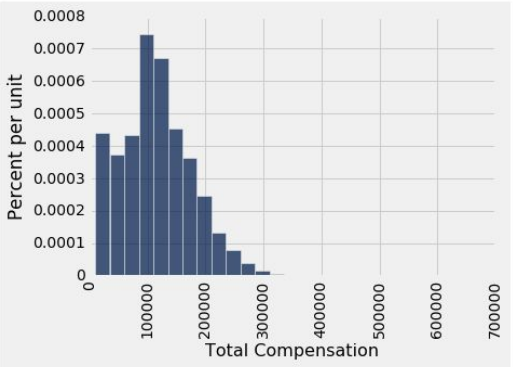
- **The mean/median might not be true values**
- **The mean/median can become decimals**
- **Same units as the values you measured**



Q1

1.1 Suppose a set of numbers has mean value 15 and median value 20. Is the distribution of the values in the data skewed *left* or skewed *right*?

1.2 In the graph to the right, is the mean or the median larger?



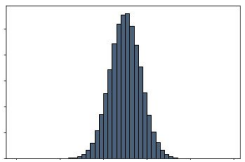
1.3 Suppose you have an array containing three 18s, seven 11s, and a 74.

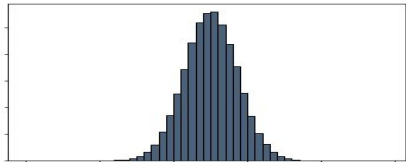
a. Write an arithmetic expression to calculate the mean of the array. How does the 74 affect the histogram?

b. Now suppose we replace the 74 with 350. How does this affect the mean? How about the median?

Variability

- These graphs have the same mean, but their spread spread is different.

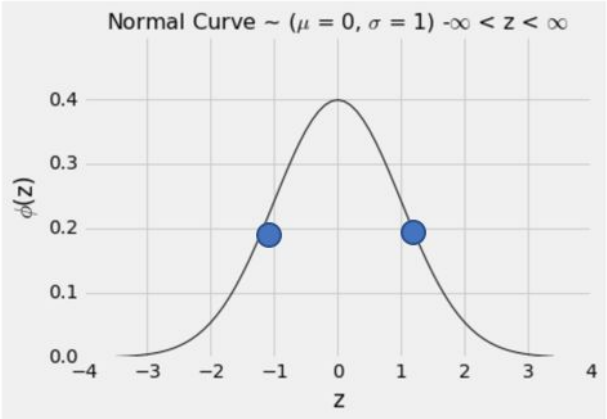




- **SD =** *Root (Mean (Squared (Difference from the average))*
- **Variance = Standard Deviation ^{2}**
- **Converting to SU**
  - Sometimes units are on different scales, i.e. you are predicting ($) vs gallons.
  - SU = $\frac{value - average}{SD}$
  - Just think of it as converting from Celsius to Fahrenheit.

Standard Normal Curve

- **Standard Normal Curve:**
  - Symmetric
  - Bell-shaped
  - Standard Deviation of **1**
  - Mean of **0**



- When we convert to standard units, we force a graph to look like this graph above!

Graph Fact

- *By Chebyshev's bound (if you are interested, hit me up for the proof):*
  - ***For all distributions,*** we know this is true:

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%)

Graph Fact: Normal Distribution

- ***For the normal distribution (symmetric bell-shaped curve),*** we know more:

Percent in Range	Normal Distribution: Approximation
average $\pm$ 1 SD	about 68%
average $\pm$ 2 SDs	about 95%
average $\pm$ 3 SDs	about 99.73%

Q2

Practice Problems

2.1 Write code to convert the delay times in column "Delay" from the `united` table at right to standard units. Name the array of converted times `delay_standard`.

Date	Flight Number	Destination	Delay
6/21/15	1964	SEA	580
6/22/15	300	HNL	537
6/21/15	1149	IAD	508
6/20/15	353	ORD	505
8/23/15	1589	ORD	458

Q3

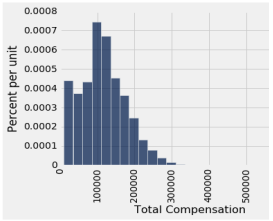
Practice Problem

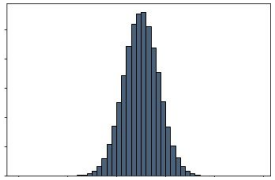
3.1 Vehicle speeds on a highway are normally distributed with mean 90 mph and SD 10 mph. Using the table above, what is the approximate probability that a randomly chosen car is going more than 100 mph?

Hint: Remember that the total area under the normal curve is 1, and that the area under a region of the curve represents the proportion of total data that falls in that region.

Central Limit Theorem

- There is something cool about the **mean:**
  - If we collect a **large**, **random** sample **with replacement,** regardless of the distribution of the population, **the distribution of all your sample means** (or the sum of the samples) will be **approximately normal**.





- Sample many times (**large samples, with replacement**)
- Take the **mean**

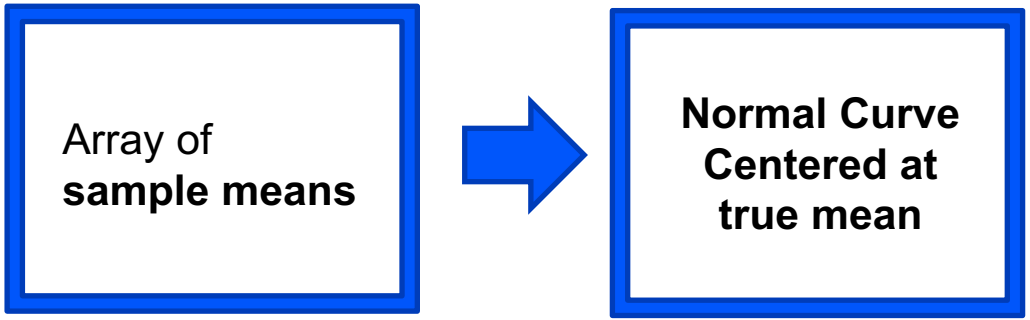
→ Array of **sample means** →

Normal Curve Centered at true mean

Where you read mean, you can also sub in **sum**

Variability of Sample Mean

The **standard deviation** of these **sample means** depends on the **sample size!**



- SD of Sample Means = $\frac{Population\ SD}{\sqrt{Sample\ Size}}$
- So the smaller the SD, the more accurate my estimate.

Q4

Practice Problems

4.1 Suppose you simulate the proportion of purple-flowered plants in a sample of 200 plants (from Mendel's 75% purple- and 25% white-flower plant population) using `sample_proportions` 1000 times. Then, you plotted distribution of the proportion of purple-flowered plants from each of the 1000 trials. What would this distribution look like? Where would the distribution be centered?

4.2 What would it look like if we used a sample size of 800 instead?

Q5

5.1 As sample size increases, what happens to the distribution of the sample mean? Does it become narrower or wider? Where is it centered?

5.2 Does population size affect the variability of the sample mean?

5.3 If you had a sample size of 100, but wanted to increase accuracy by a factor of 4, what should the new sample size be?

End of Section

- Please complete the anonymous Feedback form so I can improve my teaching:
  - **https://tinyurl.com/feedbackD8Kevin**
- Solutions and notes will be posted after Wednesday.
- Email me if you have any questions: kevinmiao@berkeley.edu