Week 11 : Sample Means, Center/Spread, Normal Distribution

Data 8 Tutoring

# 1 Mean and Median

## Key Concepts

**Mean: Definition**

The average, or mean, of a collection of numbers is the sum of all the elements divided by the total number of elements in the collection.

**Median: Definition**

The median is the 50th percentile of a collection of numbers. It is the “middle” element.

**Properties of the Mean and Median**

* They mean and median aren’t necessary elements of the set of numbers.
* They might not be an integer even if all the elements of the collection are integers.
* If the collection consists of values measured in specified units, then it has the same units too.

**Mean vs. Median**

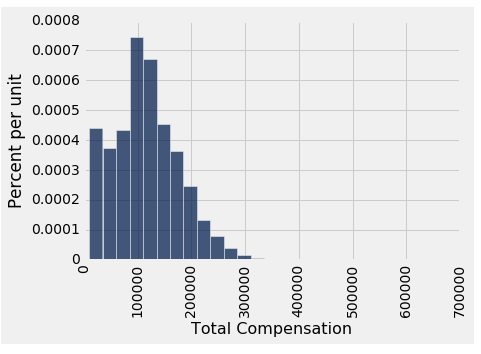
The median is always the midpoint of the data, while the mean is affected by the magnitude of the data points. For example, if the data is concentrated to the right with fewer values on the left, the mean is dragged to the left by those tail values.



## 

## Practice Problems

**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.

1. Write an arithmetic expression to calculate the mean of the array. How does the 74 affect the histogram?
2. Now suppose we replace the 74 with 350. How does this affect the mean? How about the median?

# 

# 2 Variability

## Key Concepts

**Calculating Variance and SD**

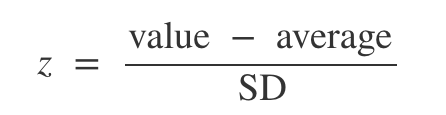
SD: “Root mean squared deviation from average”

5 4 3 2 1

Assume dist = [2, 4, 6, 8, 10]

* First, find the average of the distribution.
  + average = 6
* Next, find the difference between each number in the distribution and the average.
  + differences = [-4, -2, 0, 2, 4]
* Square each difference (so there are no negatives).
  + squared\_differences = [16, 4, 0, 4, 16]
* Now take the mean of all the squared differences. (Variance)
  + mean\_squared\_differences = 8
* Take the square root of that mean. (Standard Deviation)
  + root\_mean\_squared\_differences =

Alternatively, you can also use np.std(array) to calculate the standard deviation!

**Standard Units**

To convert a value to standard units (a unitless measure), first find how far it is from the average of the distribution, and then compare that deviation with the standard deviation of the distribution.

## Screen Shot 2017-10-26 at 8.29.18 PM.png

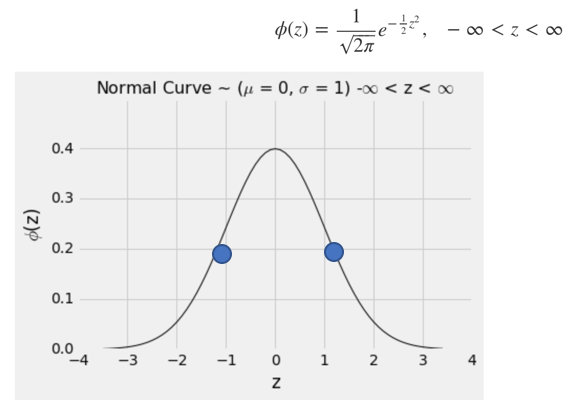
## Practice Problems

**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.

# 

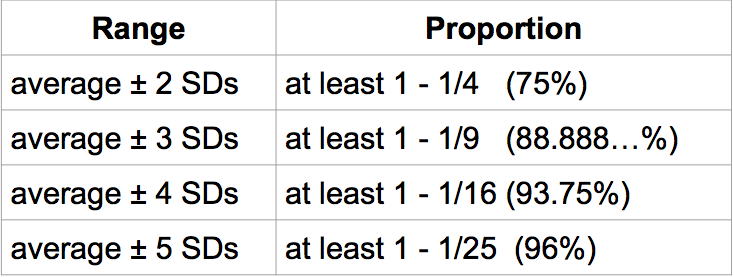
# 3 SD and Normal Curve

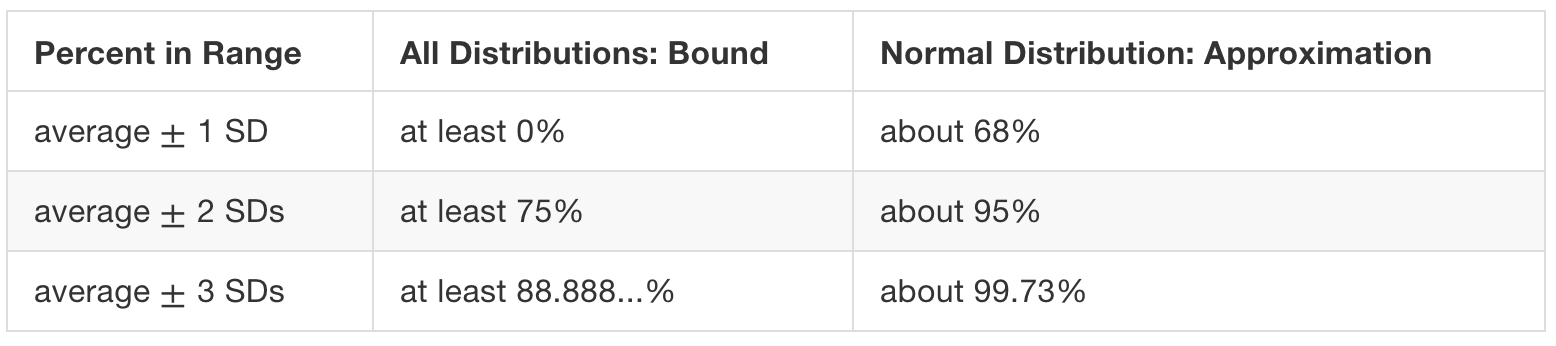
## Key Concepts

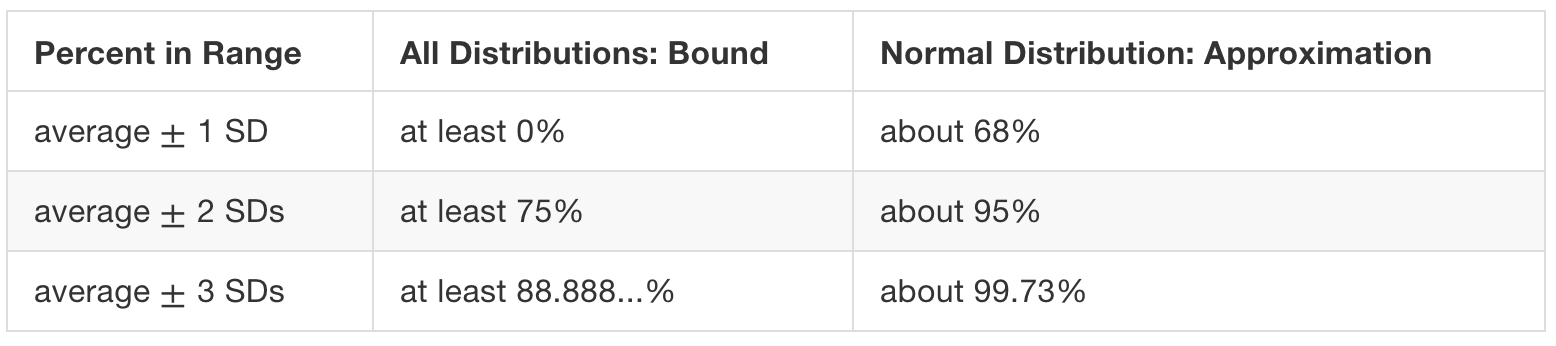
**Overview**

Here is the standard normal curve (mean = 0, SD = 1) and some of its properties:

* The total area under the curve is 100%.
* The curve is symmetric about 0, with its mean and median both equal to 0.
* If a variable has this distribution, its SD is 1. The standard normal curve is one of the few distributions that has a SD so clearly identifiable on the histogram.



**Chebyshev’s Bounds**: The table on the right uses Chebyshev’s inequality to calculate the following proportion of values that fall within *k* SDs of the mean. Remember that Chebyshev’s bound works for **ALL** distributions, which is why it is a weaker bound.

The table below shows the Chebyshev bounds for the normal distribution.

## 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.

# 4 Central Limit Theorem

## Key Concepts

**Overview**

The Central Limit Theorem says that the probability distribution of the **sum or average of a large random sample drawn with replacement will be roughly normal**, regardless of the distribution of the population from which the sample is drawn.

## Practice Problems

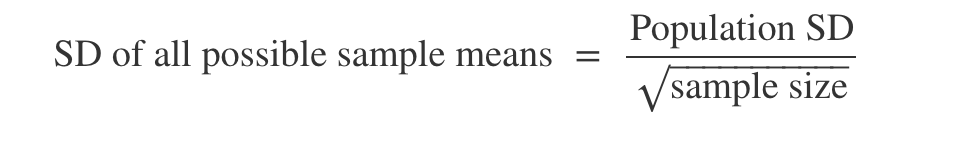
**4.1** Suppose you simulate the proportion of purple-flowered plants in a sample of 200 plants (from Mendel’s 75% purple- and 25% white-flower plant population) using sample\_proportions 1000 times. Then, you plotted distribution of the proportion of purple-flowered plants from each of the 1000 trials. 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?

# 5 Variability of the Sample Mean

## Key Concepts

**The SD of the Sample Mean**

****

This is the standard deviation of the averages of all the possible samples that could be drawn. **It measures roughly how far off the sample means are from the population mean.** The smaller the SD, the more accurate the estimate.

## Practice Problems

**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?