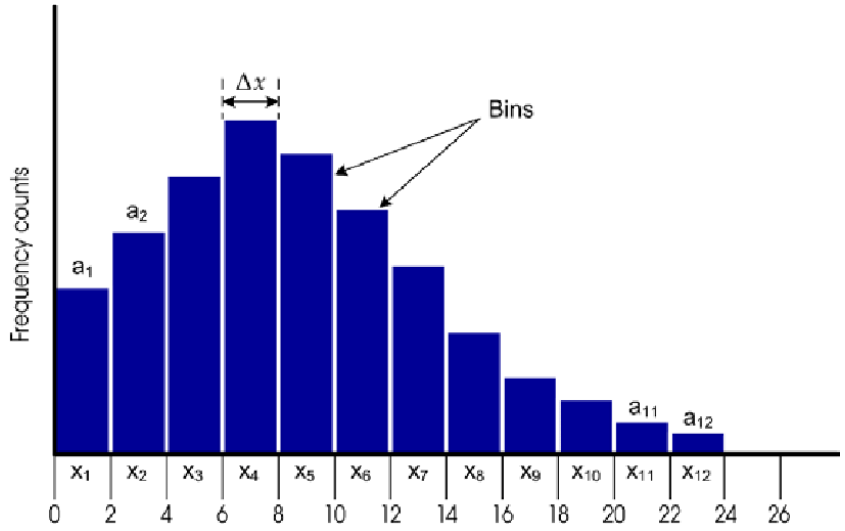
SESSIONS 4

Descriptive Statistics:
Numerical and Table Summary

Data Science Program

Descriptive Statistic

- Descriptive statistic includes the construction of graphs, charts, tables, and calculation of various descriptive measures such as averages, variation, and percentile.




```
tips = sns.load_dataset('tips')
```

```
tips.describe()
```

	total_bill	tip	size
count	244.000000	244.000000	244.000000
mean	19.785943	2.998279	2.569672
std	8.902412	1.383638	0.951100
min	3.070000	1.000000	1.000000
25%	13.347500	2.000000	2.000000
50%	17.795000	2.900000	2.000000
75%	24.127500	3.562500	3.000000
max	50.810000	10.000000	6.000000

Purwadhika
Startup and Coding School

Numerical Summary

There are generally 2 types of numerical summary:

- **Measures of Central Tendency.** It is the way of describing the central position of a frequency distribution for a group of data. We can describe it by using, for example Mean, Median, Mode

- **Measures of Spread.** It is the way to summarize the group of data by describing how spread the data are. We can describe it by using, for example Range, Quartile, Variance, and Standard Deviation

Measure of Central Tendency

There are three most common way to describe the central measurement of the frequency distribution:

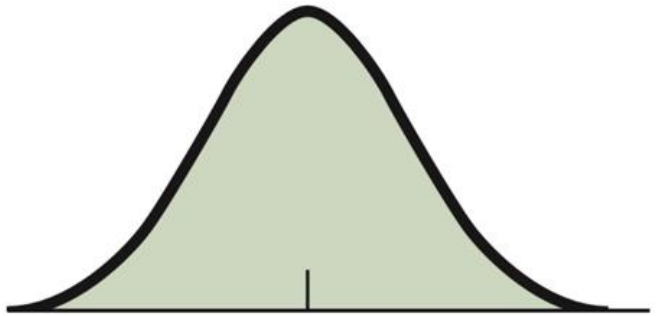
- **Mode**: Value of a qualitative or a countable quantitative variable where the frequency is occurring the most.
- **Median**: The middle value in the ordered list. If the number of observation is odd, then the sample median is the observed value exactly in the middle. If the number of observation is even, then the sample median is the number halfway between the two middle observed values in the ordered list. In either case, the sample median position is at n+1/2 when n is the number of observation
- **Mean**: The sum of observed values in a data divided by the number of observations. The most commonly used measure of center for quantitative variable.

Which measurement to choose?

- **Mode** should be used when calculating the measure of center for the qualitative variable
- **Mean** is the proper measure of center if dealing with the quantitative variable with symmetric distribution (often bell shaped)
- **Median** is the good choice if the quantitative variable have a skewed distribution. We do not used mean in this case, because mean could be highly influenced by an observation that falls far from the rest of the data (outlier)
- It should be noted that this measurement assume that the sample measurement is corresponding to the population measures of center, which are unknown. The sample measurement can be used to estimate this unknown parameter.

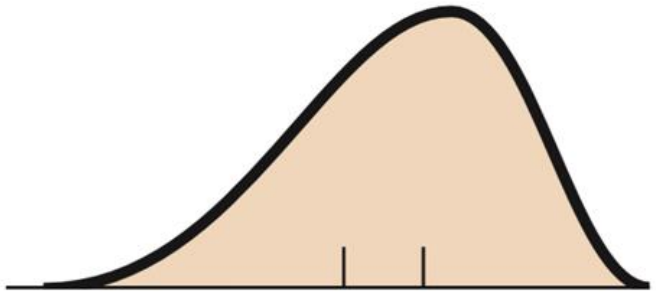
Median vs Mean

Symmetric Distribution



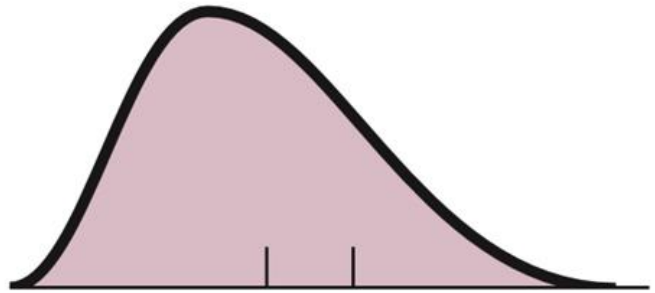
Mean = Median

Left-Skewed Distribution



Mean Median

Right-Skewed Distribution



Median Mean

Purwadhika
Startup and Coding School

Measure of Central Tendency Example

Patient	Gender	Age
Andrew	Male	22
Jacob	Male	22
Ros	Female	23
Andersen	Male	23
Lina	Female	29
Robert	Male	24
Jack	Male	27
Annie	Female	28

- **Mode**
  - Gender: Male
  - Age: 22 and 23
- **Mean**
  - Age
    (22 + 22 + … + 28) / 8 = 24.75
- **Median**
  Age: 22, 22, 23, **23, 24**, 27, 28, 29,
  Median = 23.5

Measure of Spread

- Another important aspect of descriptive study is numerically measuring the extent of variation around the center.
- Two dataset of the same variable may possess similar position of center but remarkably different with respect to variability.
- Most frequently used measures of variation; the sample range, the sample interquartile range, and the sample standard deviation

Range

- The sample range is obtained by computing the difference between the largest observed value of the variable and the smallest one

Range = Max – Min

- Range is overly sensitive to extreme value

Standard Deviation

- The sample standard deviation is the most frequently used measure of variability. It can be considered as a kind of average of the absolute deviations of observed values from the mean of the variable in question.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

S = Sample Standard Deviation
X_i = Each value of dataset
$\bar{x}$ = Mean of the dataset
N = Sample size

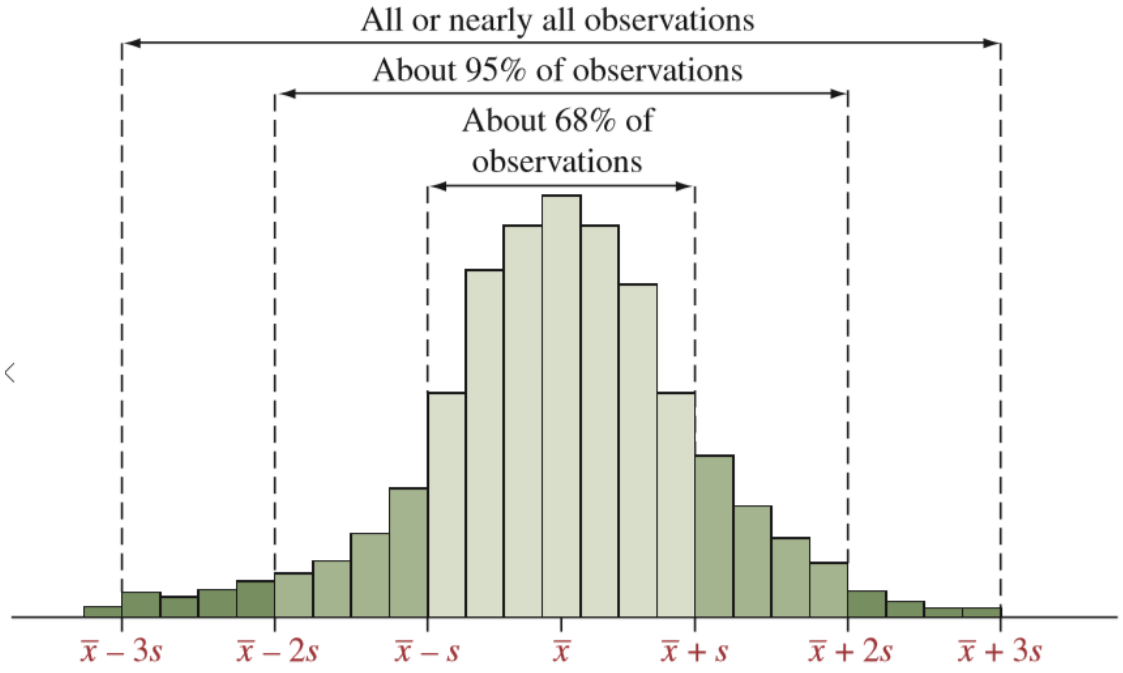
- Since Standard Deviation defined by the sample mean, it is preferred measure of variation if the mean is used as the measure of center(ex: Symmetric Distribution)

Standard Deviation Properties

- The more variation in the observed value, the larger the standard deviation for the variable observed.
- Standard Deviation is greatly affected by a few extreme observation (usually outlier).

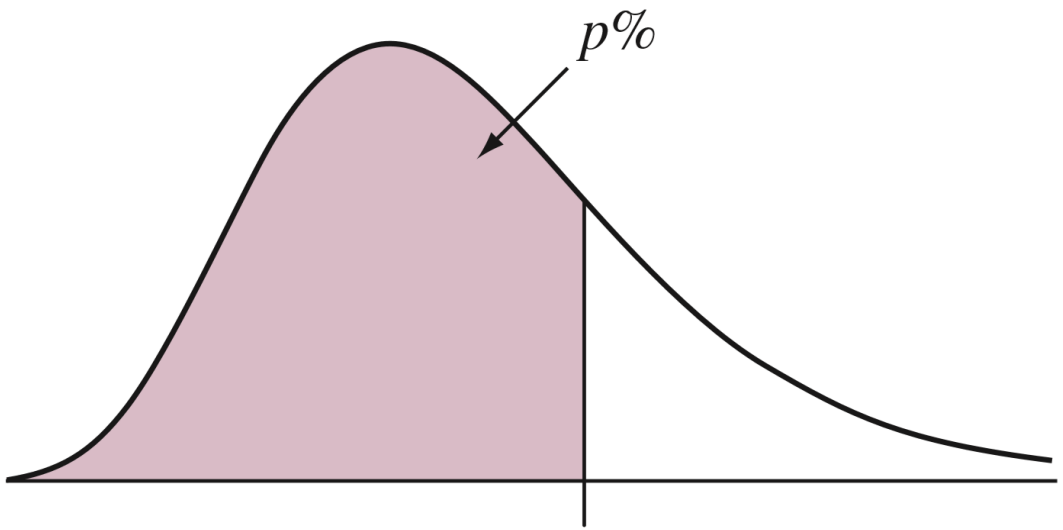
When a distribution of Data is bell shaped, then approximately:

- 68% of observation fall between $\bar{x} - s$ and $\bar{x} + s$
- 95% of observation fall between $\bar{x} - 2s$ and $\bar{x} + 2s$
- 99.7% of observation fall between $\bar{x} - 3s$ and $\bar{x} + 3s$
- $s \approx$ Range / 4 = (Max – Min) / 4



Purwadhika
Startup and Coding School

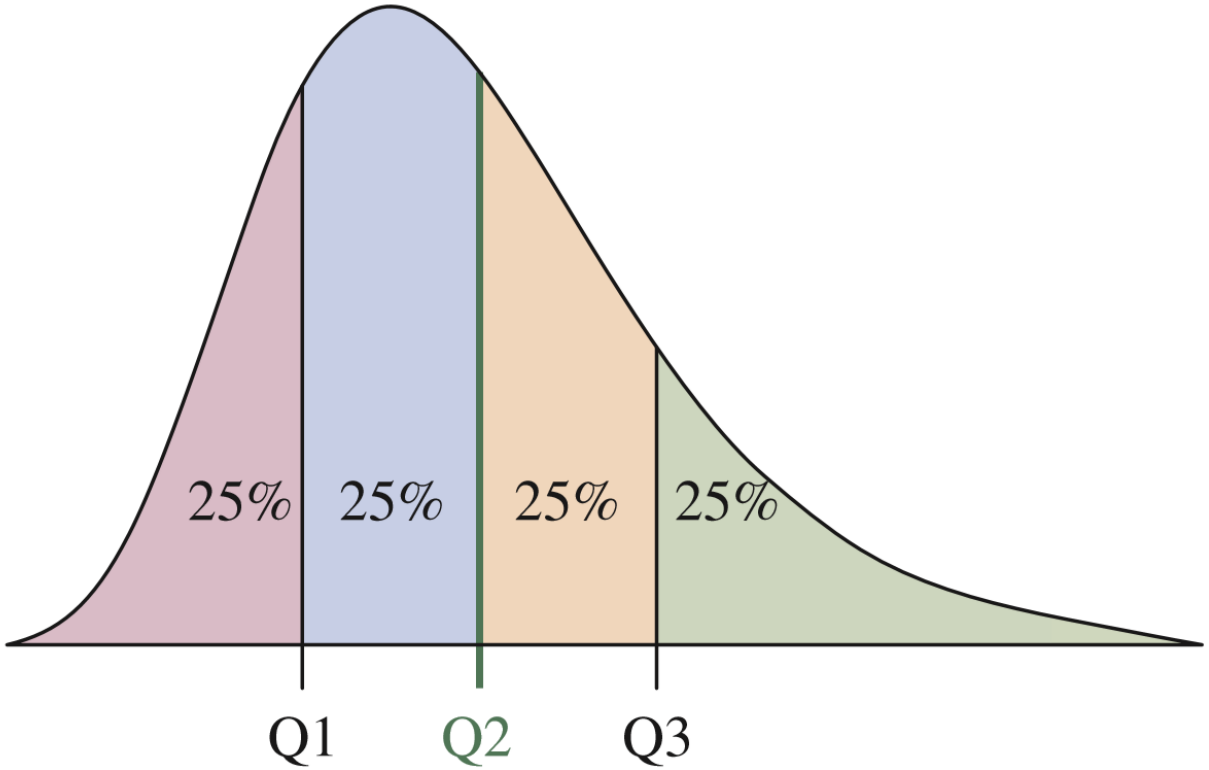
Percentile



The ***p*th Percentile** is a value such that *p* percent of the observation fall below or at that value.

Purwadhika
Startup and Coding School

Quartile



Quartile is a special case of **percentile**.

- Q1 is percentile 25
- Q2 is percentile 50 or also known as median
- Q3 is percentile 75

Purwadhika
Startup and Coding School

IQR and Outlier

Interquartile range (IQR) is the distance between Q1 and Q2:

IQR = Q3 - Q1

IQR can be used as replacement for standard deviation when data is normally distributed because standard deviation is very sensitive to outliers

S = 1.34898 x IQR

IQR is used to detect the potential outlier

An observation considered outlier if the value

- below Q1 - 1.5 x IQR
- above Q3 + 1.5 x IQR

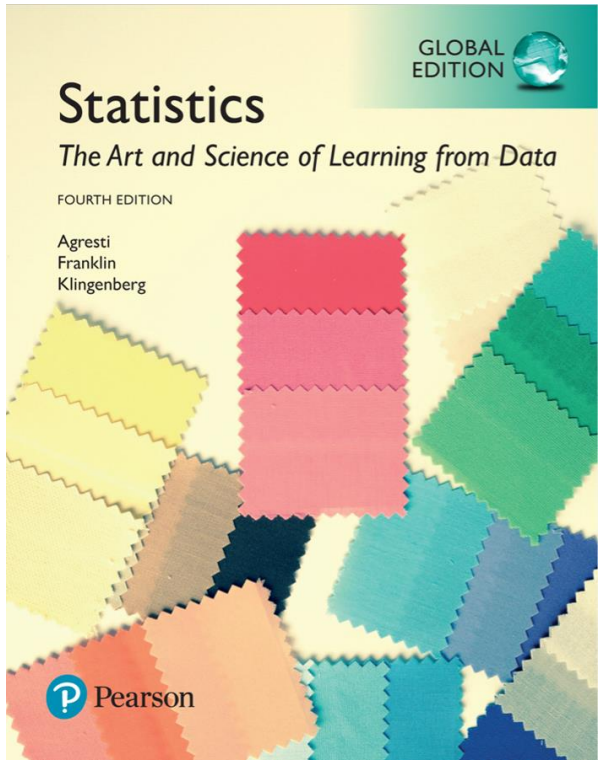
Measure of Spread Example

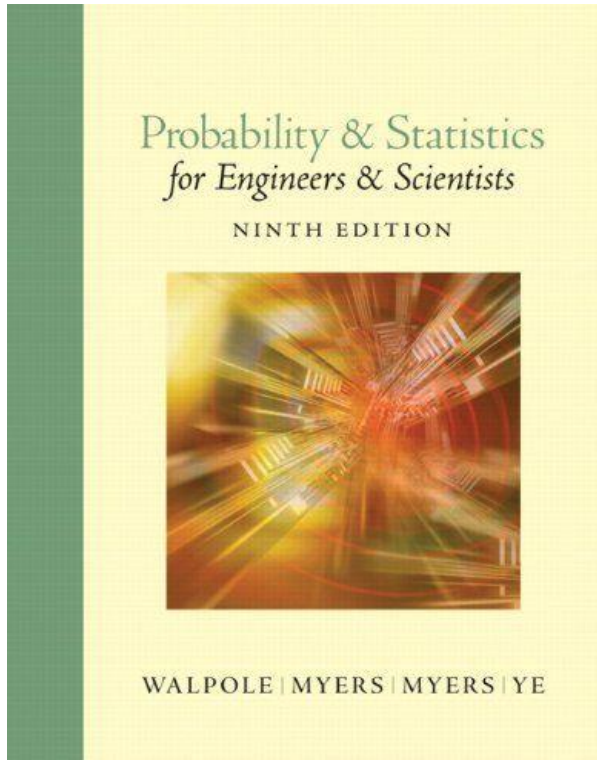
Patient	Gender	Age
Andrew	Male	22
Jacob	Male	22
Ros	Female	23
Andersen	Male	23
Lina	Female	29
Robert	Male	24
Jack	Male	27
Annie	Female	28

- Range
  - Age
    29 - 22 = 27
- Standard Deviation
  - Age
    s = 2.63391
- Q2
  22, 22, 23, **23, 24**, 27, 28, 29
  23.5
- Q1 22.75
- Q3 27.25
- IQR 4.5

Purwadhika
Startup and Coding School

Reference





Purwadhika
Startup and Coding School

Reference

https://towardsdatascience.com/data-science-you-need-to-know-a-b-testing-f2f12aff619a

https://towardsdatascience.com/data-science-fundamentals-a-b-testing-cb371ceecc27

https://www.niagahoster.co.id/blog/ab-testing-adalah/

https://vwo.com/blog/ab-testing-examples/

https://www.scribbr.com/methodology/sampling-methods/

Purwadhika
Startup and Coding School