# Lesson 6

Descriptive Statistics

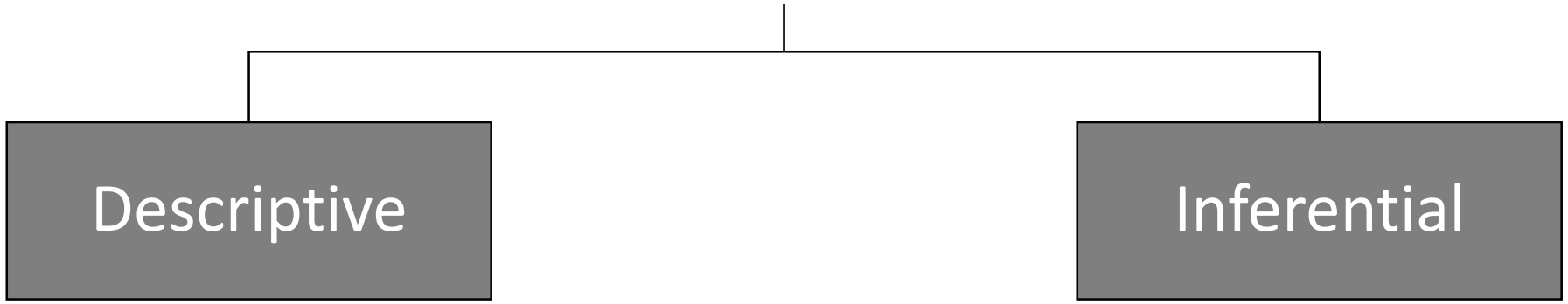
Kush Kulshrestha

# Statistics

/stəˈtɪstɪks/

*noun*

the practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.




Describing, presenting, summarizing or organizing your data via numerical calculations or graphs or tables

Complex mathematics, allow us to infer trends, make assumptions & predictions about population based on samples

# Why Descriptive statistics is important?

**Data Science** and **Machine Learning**

*"are all about understanding data and being able to take meaning out of data, which otherwise is not possible through classical techniques, and make reasonably accurate predictions."*

**Descriptive Statistics**

Is about understanding the data by describing it in some other aggregate form.
It is going to be your first line of defense against unknown data.

# Descriptive statistics - Topics

- Measure of Central Tendency
  - Mean
  - Median
  - Mode

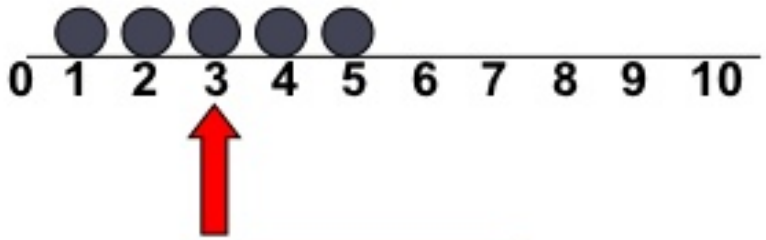
- Measure of Variability
  - Range
  - IQR
  - Variance
  - Standard Deviation
  - Skewness and Kurtosis

- Probability and Distributions
  - Basics
  - Probability Distributions
  - Normal Distribution
  - Z score

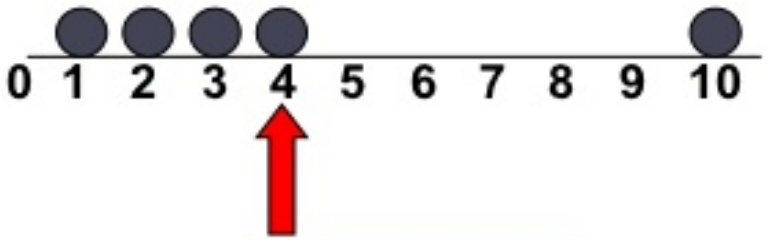
# Measure of Central Tendency - Mean

This describes the whole data with a single value that represents the center of the data distribution.

**Mean:**
Sum of observations divided by the sample size. It is a common metric used in day to day life.
Also called average.
Mean is sensitive to outliers:




$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

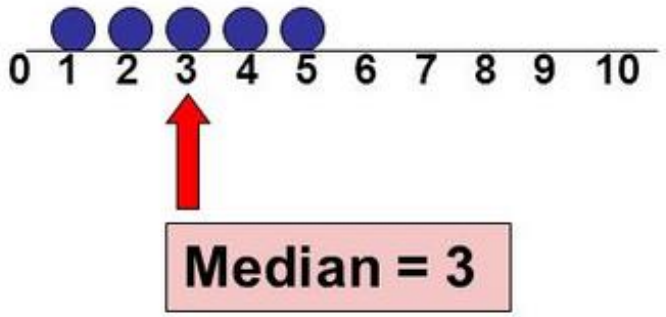


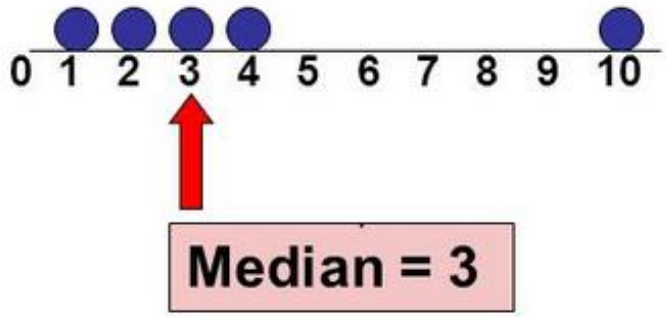

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

# Measure of Central Tendency - Median

This describes the whole data with a single value that represents the center of the data distribution.

**Median:**
It is the middle value of data. It splits the data in half and also called 50th percentile. *It is much less affected by the outliers and skewed data than mean*.
If the no. of elements in the dataset is odd, the middle most element is the median.
If the no. of elements in the dataset is even, the median would be the average of two central elements.

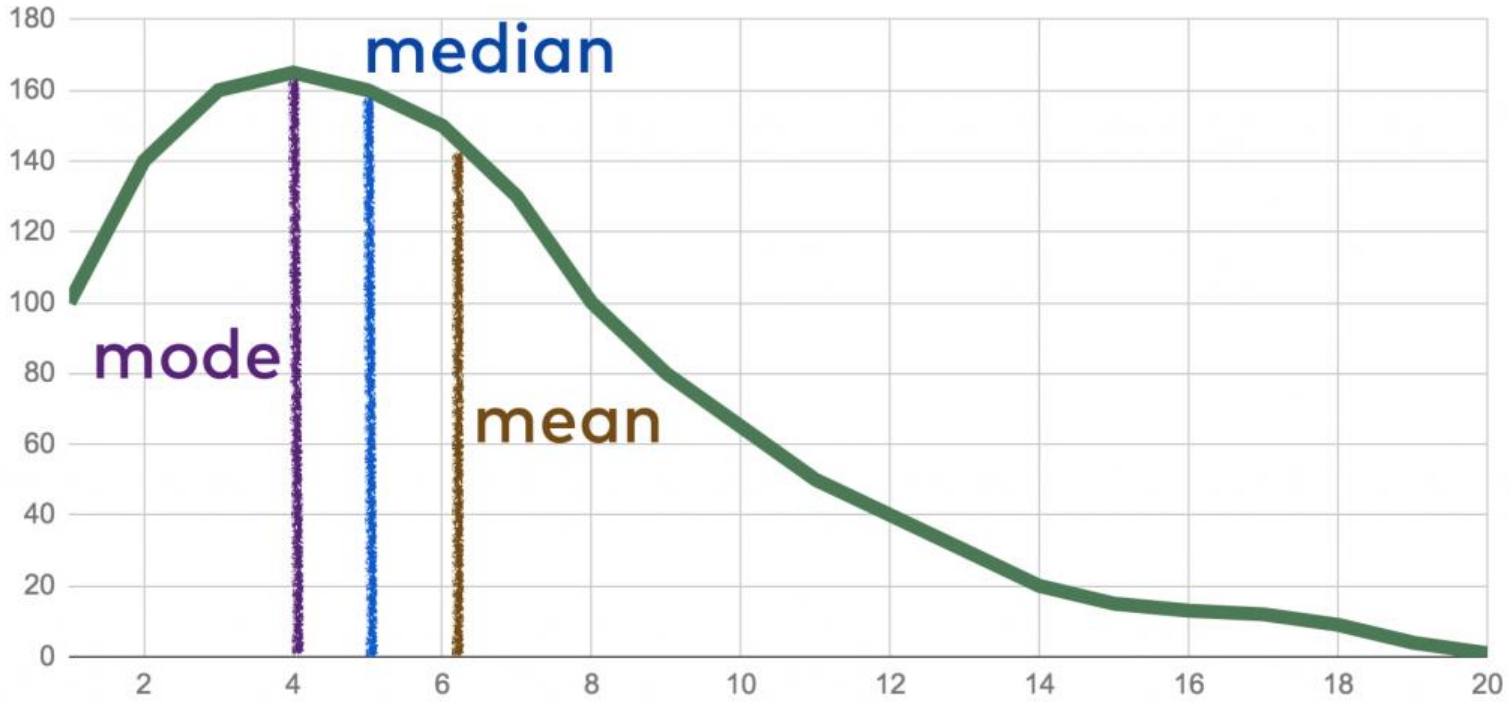


# Measure of Central Tendency - Mode

This describes the whole data with a single value that represents the center of the data distribution.

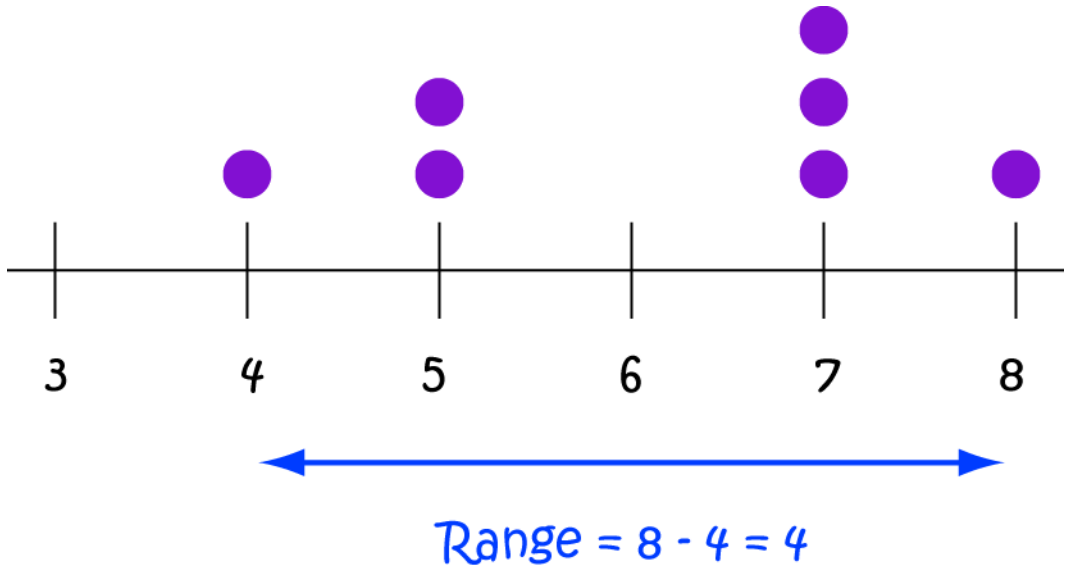
**Mode:**
It is the value that occurs more frequently in a dataset. Therefore a dataset has no mode, if no category is the same and also possible that a dataset has more than one mode. It is the only measure of central tendency that can be used for categorical variables.



# Measure of Variability - Range

Measures of variability are also called spreads. They describe how varied are the set of observations.

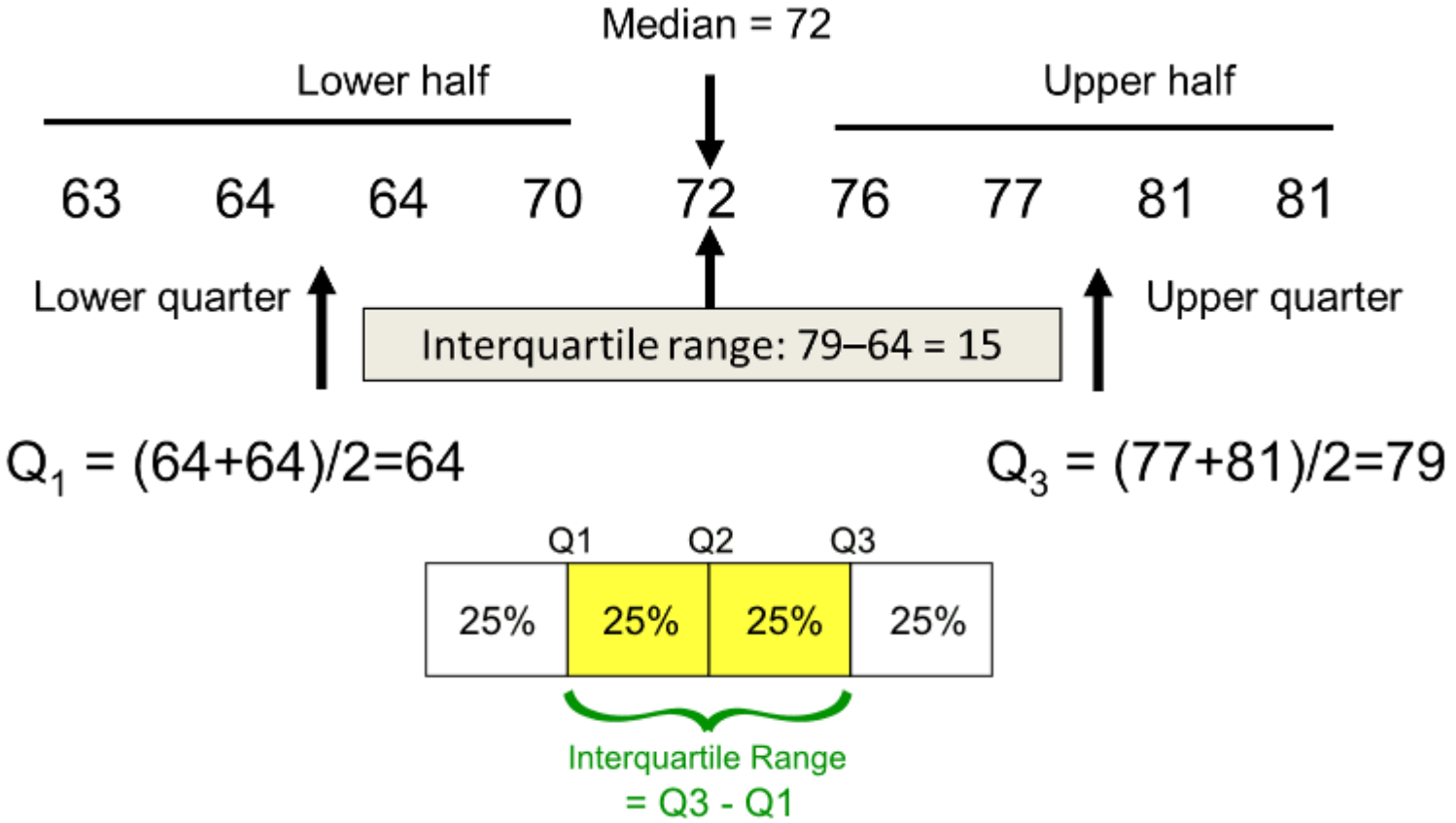
**Range:**
It is the difference between the largest and the smallest data points in the dataset.



# Measure of Variability - IQR

Measures of variability are also called spreads. They describe how varied are the set of observations.

**IQR or Inter Quartile Range:**
It is a measure of statistical dispersion between upper (75th) quartiles i.e Q3 and lower (25th) quartiles i.e Q1. While the range measures where the beginning and end of your data are, the interquartile range is a measure of where the majority of the values lie.

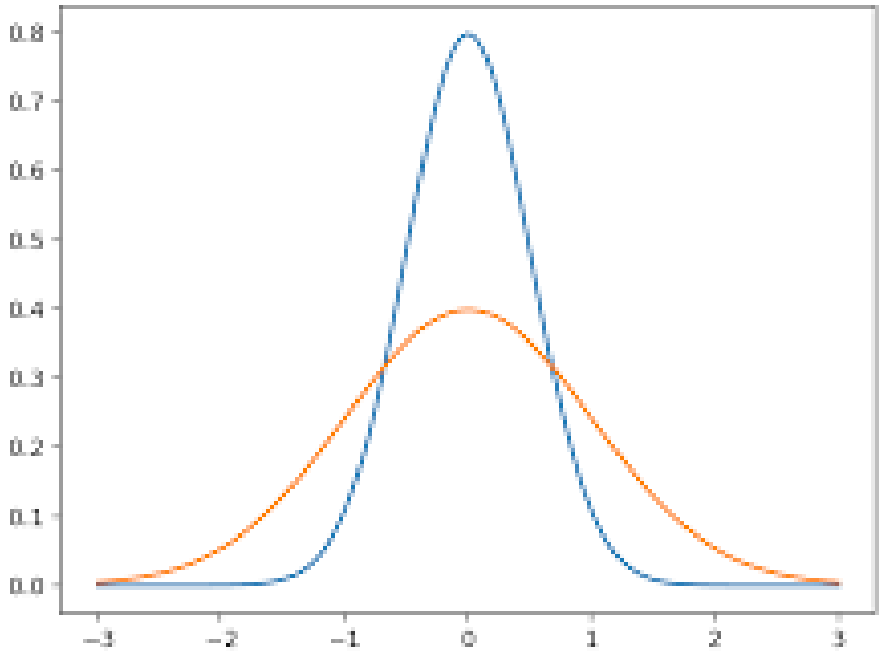


# Measure of Variability - Variance

Measures of variability are also called spreads. They describe how varied are the set of observations.

**Variance:**
It is the average squared deviation from mean. The variance is computed by finding the difference between every data point and the mean, squaring them, summing them up and then taking the average of those numbers. It is a measure of how far values in the dataset lie from the mean.

***The problem with Variance is that because of the squaring, it is not in the same unit of measurement as the original data.***



# Measure of Variability – Standard Deviation

Measures of variability are also called spreads. They describe how varied are the set of observations.

**Standard Deviation:**
Standard Deviation is used more often because it is in the original unit. It is simply the square root of the variance and because of that, it is returned to the original unit of measurement.
When you have a low standard deviation, your data points tend to be close to the mean. A high standard deviation means that your data points are spread out over a wide range.

The squares are used during the calculation because they weight outliers more heavily than points that are near to the mean. This prevents that differences above the mean neutralize those below the mean.

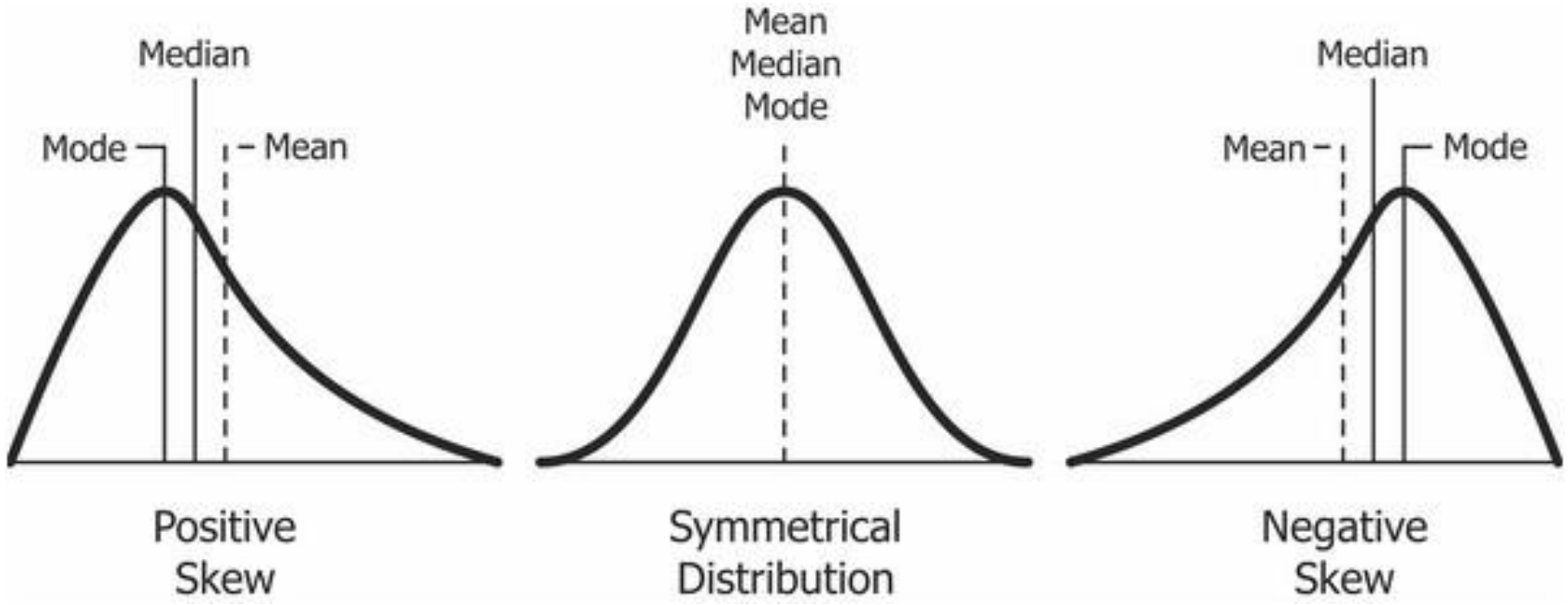
**Units of Variance and Std dev.:**
Imagine a data set that contains centimeter values between 1 and 15, which results in a mean of 8. Squaring the difference between each data point and the mean and averaging the squares renders a variance of 18.67 (squared centimeters), while the standard deviation is 4.3 centimeters.

# Measure of Variability – Skewness

Measures of variability are also called spreads. They describe how varied are the set of observations.

**Skewness:**
It is a measure of symmetry of a distribution.

Therefore it describes how much a distribution differs from a normal distribution, either to the left or to the right. The skewness value can be either positive, negative or zero.
A perfect normal distribution would have a skewness of zero because the mean equals the median.



# Measure of Variability – Skewness

**We speak of a positives skew if the data is piled up to the left**, which leaves the tail pointing to the right.

**A negative skew occurs if the data is piled up to the right**, which leaves the tail pointing to the left. Note that positive skews are more frequent than negative ones.

A good measurement for the skewness of a distribution is **Pearson's skewness coefficient** that provides a quick estimation of a distributions symmetry.

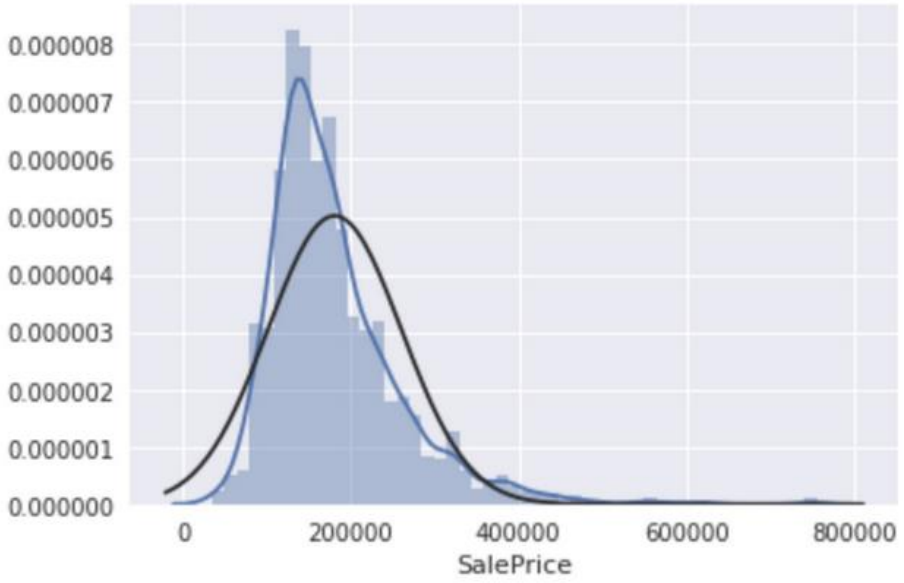
To compute the skewness in pandas you can just use the skew() function which we will apply in the notebooks.

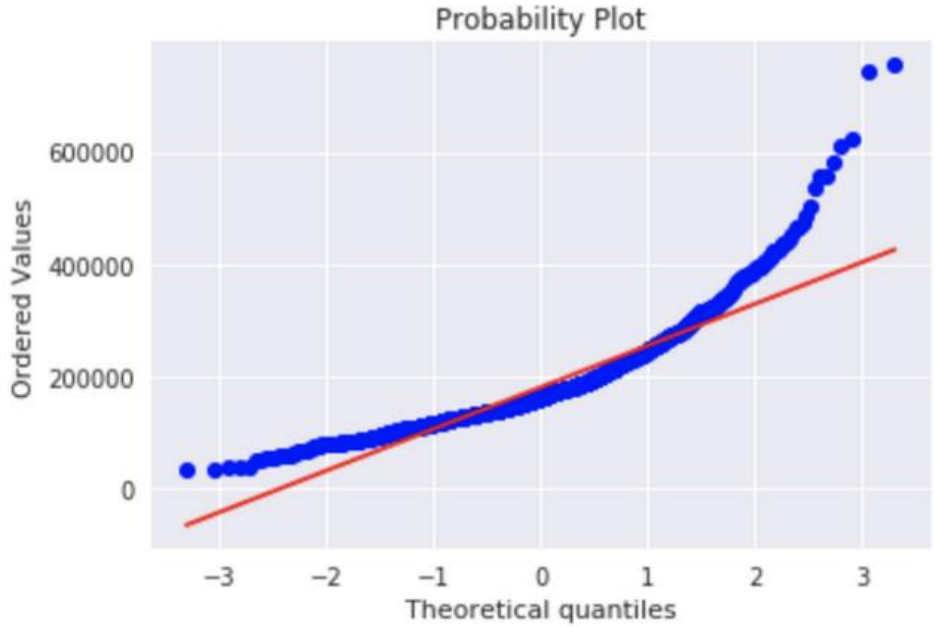
# Measure of Variability – Kurtosis

Measures of variability are also called spreads. They describe how varied are the set of observations.

**Kurtosis:**
Kurtosis measures whether your dataset is heavy-tailed or light-tailed compared to a normal distribution.
**Data sets with high kurtosis have heavy tails and more outliers and data sets with low kurtosis tend to have light tails and fewer outliers.**

Note that a histogram is an effective way to show both the skewness and kurtosis of a data set because you can easily spot if something is wrong with your data.



# Probability and Distributions

## **Probability basic terminology**:

**Experiment**: An experiment could be something like — whether it rains in Delhi on a daily basis or not.

**Outcome**: Outcome is the result of a single trial. If it rains today, the outcome of today's trial is "it rained".

**Event**: An event is one or more outcomes of an experiment. For the experiment of whether it rains in Delhi every day the event could be "it rained" or it didn't rain.

**Probability**: This simply the likelihood of an event. So it there's a 60% chance of it raining today, the probability of raining is 0.6.
Mathematically, the probability that an event will occur is expressed as a number between 0 and 1.
Notationally, the probability of event A is represented by P(A).

- If P(A) equals zero, event A will almost definitely not occur.
- If P(A) is close to zero, there is only a small chance that event A will occur.
- If P(A) equals 0.5, there is a 50-50 chance that event A will occur.
- If P(A) is close to one, there is a strong chance that event A will occur.
- If P(A) equals one, event A will almost definitely occur.

# Probability and Distributions

## **Probability basics**:

The sum of probabilities for all possible outcomes is equal to one. This means, for example, that if an experiment can have three possible outcomes (A, B, and C), then P(A) + P(B) + P(C) = 1.

The probability that the experiment results in a successful outcome (S) is:

P(S) = ( Number of successful outcomes ) / ( Total number of equally likely outcomes )

**Example:**
Consider the following experiment. An urn has 10 marbles. Two marbles are red, three are green, and five are blue. If an experimenter randomly selects 1 marble from the urn, what is the probability that it will be green?

In this experiment, there are 10 equally likely outcomes, three of which are green marbles. Therefore, the probability of choosing a green marble is 3/10 or 0.30.

# Probability and Distributions

## **Rules of Probability**:

- Two events are **mutually exclusive** or **disjoint** if they cannot occur at the same time.
- The probability that Event A occurs, given that Event B has occurred, is called a **conditional probability**. The conditional probability of Event A, given Event B, is denoted by the symbol P(A|B).
- The **complement** of an event is the event not occurring. The probability that Event A will not occur is denoted by P(A').
- The probability that Events A and B *both* occur is the probability of the **intersection** of A and B. The probability of the intersection of Events A and B is denoted by P(A ∩ B). If Events A and B are mutually exclusive, P(A ∩ B) = 0.
- The probability that Events A or B occur is the probability of the **union** of A and B. The probability of the union of Events A and B is denoted by P(A ∪ B) .
- If the occurrence of Event A changes the probability of Event B, then Events A and B are **dependent**. On the other hand, if the occurrence of Event A does not change the probability of Event B, then Events A and B are **independent**.

# Probability and Distributions

**Rule of Multiplication:**
The probability that Events A and B both occur is equal to the probability that Event A occurs times the probability that Event B occurs, given that A has occurred.

$$P(A \cap B) = P(A)\, P(B|A)$$

**Rule of Addition:**
The probability that Event A or Event B occurs is equal to the probability that Event A occurs plus the probability that Event B occurs minus the probability that both Events A and B occur.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

**Rule of Subtraction:**
The probability that event A will occur is equal to 1 minus the probability that event A will not occur.

$$P(A) = 1 - P(A')$$

# Probability and Distributions

## Example 1:

An urn contains 6 red marbles and 4 black marbles. Two marbles are drawn *without replacement* from the urn. What is the probability that both of the marbles are black?

**Sol.**

Let A = the event that the first marble is black; and let B = the event that the second marble is black. We know the following:

In the beginning, there are 10 marbles in the urn, 4 of which are black. Therefore, $P(A) = 4/10$.

After the first selection, there are 9 marbles in the urn, 3 of which are black. Therefore, $P(B|A) = 3/9$.

Therefore, based on the rule of multiplication:

$P(A \cap B) = P(A)\, P(B|A)$

$P(A \cap B) = (4/10) * (3/9) = 12/90 = 2/15 = 0.133$

**What if we do it with replacement?**

# Probability and Distributions

## Example 2:

A student goes to the library. The probability that she checks out (a) a work of fiction is 0.40, (b) a work of non-fiction is 0.30, and (c) both fiction and non-fiction is 0.20. What is the probability that the student checks out a work of fiction, non-fiction, or both?

**Sol.**

Let F = the event that the student checks out fiction; and let N = the event that the student checks out non-fiction.

Then, based on the rule of addition:

$P(F \cup N) = P(F) + P(N) - P(F \cap N)$

$P(F \cup N) = 0.40 + 0.30 - 0.20 = 0.50$

# Probability and Distributions

## Random Variable

When the value of a variable is determined by a chance event, that variable is called a **random variable**. Random variables can be discrete or continuous.

**Discrete**. Within a range of numbers, discrete variables can take on only certain values. Suppose, for example, that we flip a coin and count the number of heads. The number of heads will be a value between zero and plus infinity. Within that range, though, the number of heads can be only certain values.

**Continuous**. Continuous variables, in contrast, can take on any value within a range of values. For example, suppose we randomly select an individual from a population. Then, we measure the age of that person. In theory, his/her age can take on any value between zero and plus infinity, so age is a continuous variable.

When comparing discrete and continuous variables, it is more correct to say that continuous variables can always take on an infinite number of values; whereas some discrete variables can take on an infinite number of values, but others cannot.

# Probability and Distributions

## Probability Distribution

A **probability distribution** is a table or an equation that links each possible value that a random variable can assume with its probability of occurrence.

### Discrete Probability Distributions

The probability distribution of a discrete random variable can always be represented by a table.
For example, suppose you flip a coin two times. This simple exercise can have four possible outcomes: HH, HT, TH, and TT. Now, let the variable X represent the number of heads that result from the coin flips. The variable X can take on the values 0, 1, or 2; and X is a discrete random variable.
The probability of getting 0 heads is 0.25; 1 head, 0.50; and 2 heads, 0.25. Thus, the table is an example of a probability distribution for a discrete random variable.

| Number of heads, x | Probability, P(x) |
|---|---|
| 0 | 0.25 |
| 1 | 0.50 |
| 2 | 0.25 |

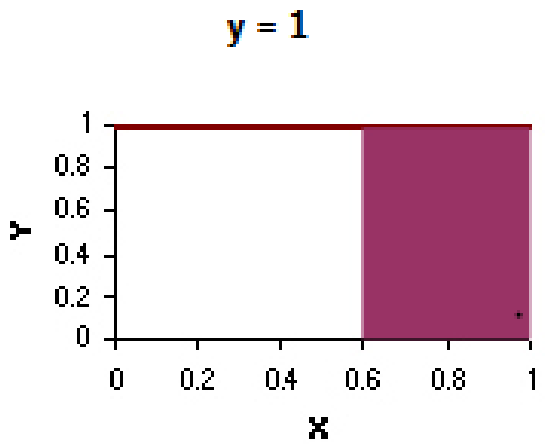
# Probability and Distributions

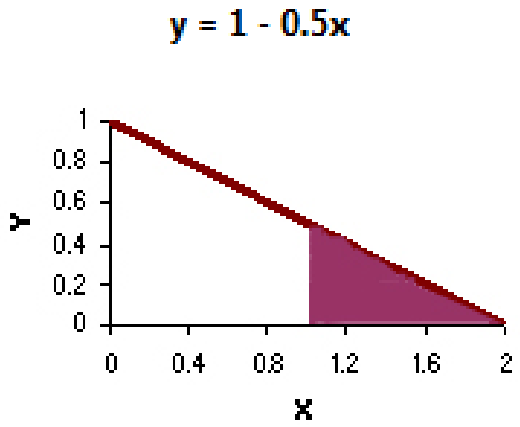
## Probability Distribution

### Continuous Probability Distributions

The probability distribution of a continuous random variable is represented by an equation, called the **probability density function** (pdf)

The charts below show two continuous probability distributions.
The first chart shows a probability density function described by the equation y = 1 over the range of 0 to 1 and y = 0 elsewhere.
The second chart shows a probability density function described by the equation y = 1 - 0.5x over the range of 0 to 2 and y = 0 elsewhere. **The area under the curve is equal to 1 for both charts**.



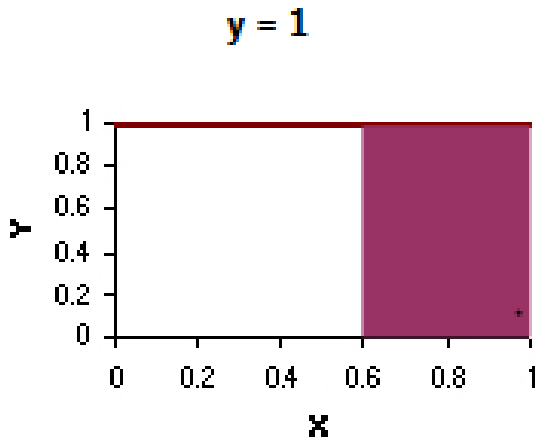
# Probability and Distributions

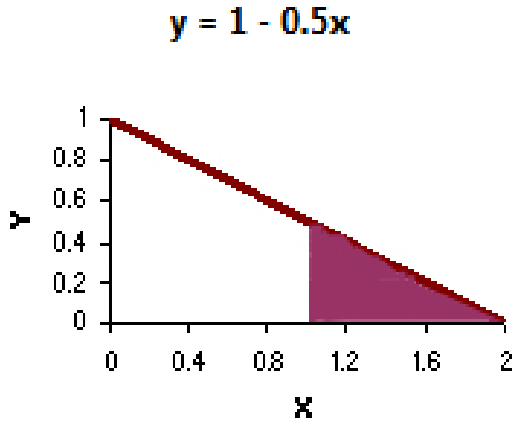
## Probability Distribution

### Continuous Probability Distributions

The probability that a continuous random variable falls in the interval between *a* and *b* is equal to the area under the pdf curve between *a* and *b*.

For example, in the first chart below, the shaded area shows the probability that the random variable X will fall between 0.6 and 1.0. That probability is 0.40. And in the second chart, the shaded area shows the probability of falling between 1.0 and 2.0. That probability is 0.25.



# Probability and Distributions

## Discrete Variables – Binomial Distribution

**Binomial Experiment:**

- The experiment consists of *n* repeated trials.
- Each trial can result in just two possible outcomes.
- The probability of success, denoted by *P*, is the same on every trial.
- The trials are independent.

**Notations**:

***x***: The number of successes that result from the binomial experiment.
***n***: The number of trials in the binomial experiment.
***P***: The probability of success on an individual trial.
***Q***: The probability of failure on an individual trial. (This is equal to 1 - *P*.)
**b(*x*; *n, P*): Binomial probability** - the probability that an *n*-trial binomial experiment results in exactly *x* successes, when the probability of success on an individual trial is *P.*

# Probability and Distributions

## Discrete Variables – Binomial Distribution

A **binomial random variable** is the number of successes *x* in *n* repeated trials of a binomial experiment. The probability distribution of a binomial random variable is called a **binomial distribution**.

**Example:**
Suppose we flip a coin two times and count the number of heads (successes). The binomial random variable is the number of heads, which can take on values of 0, 1, or 2. The binomial distribution is presented below.

| Number of heads | Probability |
|---|---|
| 0 | 0.25 |
| 1 | 0.50 |
| 2 | 0.25 |

# Probability and Distributions

## Discrete Variables – Binomial Distribution

**Properties of binomial distribution:**
The mean of the distribution ($\mu_x$) is equal to $n * P$ .
The variance ($\sigma^2_x$) is $n * P * ( 1 - P )$.
The standard deviation ($\sigma_x$) is sqrt[ $n * P * ( 1 - P )$ ].

**Binomial Formula:**
Suppose a binomial experiment consists of *n* trials and results in *x* successes. If the probability of success on an individual trial is *P*, then the binomial probability is:

$$b(x; n, P) = {}_nC_x * P^x * (1 - P)^{n-x}$$

# Probability and Distributions

## Refresher – n C r

**Combinations** are a way to calculate the total outcomes of an event where order of the outcomes does not matter. To calculate combinations, we will use the formula $_nC_r = n! / r! * (n - r)!$, where $n$ represents the total number of items, and $r$ represents the number of items being chosen at a time.

A **factorial** is the product of all the positive integers equal to and less than your number. 4! = 4 * 3 * 2 * 1

**Example:**
There are ten new movies out to rent this week on DVD. John wants to select three movies to watch this weekend. How many combinations of movies can he select?

In this problem, John is choosing three movies from the ten new releases. 10 would represent the $n$ variable, and 3 would represent the $r$ variable. So, our equation would look like $_{10}C_3 = 10! / 3! * (10 - 3)!$.

# Probability and Distributions

## Discrete Variables – Binomial Distribution – Examples

**Example1.**
Suppose a die is tossed 5 times. What is the probability of getting exactly 2 fours?

This is a binomial experiment in which the number of trials is equal to 5, the number of successes is equal to 2, and the probability of success on a single trial is 1/6 or about 0.167.

Therefore, the binomial probability is:

$$b(x; n, P) = {}_nC_x * P^x * (1 - P)^{n-x}$$
$$b(2; 5, 0.167) = {}_5C_2 * (0.167)^2 * (0.833)^3$$
$$b(2; 5, 0.167) = 0.161$$

# Probability and Distributions

## Discrete Variables – Cumulative Binomial Distribution

A **cumulative binomial probability** refers to the probability that the binomial random variable falls within a specified range.

For example, we might be interested in the cumulative binomial probability of obtaining 45 or fewer heads in 100 tosses of a coin (see Example 1 below). This would be the sum of all these individual binomial probabilities.

$b(x \leq 45; 100, 0.5) = b(x = 0; 100, 0.5) + b(x = 1; 100, 0.5) + ... + b(x = 44; 100, 0.5) + b(x = 45; 100, 0.5)$

# Probability and Distributions

## Discrete Variables – Binomial Distribution – Examples

**Example2.**
The probability that a student is accepted to a prestigious college is 0.3. If 5 students from the same school apply, what is the probability that at most 2 are accepted?

To solve this problem, we compute 3 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer we seek.

$$b(x \leq 2; 5, 0.3) = b(x = 0; 5, 0.3) + b(x = 1; 5, 0.3) + b(x = 2; 5, 0.3)$$
$$b(x \leq 2; 5, 0.3) = 0.1681 + 0.3601 + 0.3087$$
$$b(x \leq 2; 5, 0.3) = 0.8369 \text{ `}$$

# Probability and Distributions

## Discrete Variables – Binomial Distribution – Examples

**Example3.**
What is the probability that the world series will last 4 games and 5 games? Assume that the teams are evenly matched. In the world series, there are two baseball teams. The series ends when the winning team wins 4 games.
Teams are evenly matched: The probability that a particular team wins a particular game is 0.5.

Series lasts just 4 games:
This can occur if one team wins the first 4 games. The probability of the National League team winning 4 games in a row is:

$$b(4; 4, 0.5) = {}_4C_4 * (0.5)^4 * (0.5)^0 = 0.0625$$

Similarly, when we compute the probability of the other League team winning 4 games in a row, we find that it is also 0.0625. Therefore, probability that the series ends in four games would be 0.0625 + 0.0625 = 0.125

For finding the probability of series to last 5 games: This is possible if a team has won 3 games out of first 4 games: $b(3; 4, 0.5) = {}_4C_3 * (0.5)^3 * (0.5)^1 = 0.25$.
Now we have 50/50 chance of winning the $5^{th}$ games: 0.25*0.5 = 0.125. Also take into account $2^{nd}$ team.

# Probability and Distributions

## Discrete Variables – Negative Binomial Distribution

**Properties of Negative Binomial Distribution:**

- The experiment consists of *x* repeated trials.
- Each trial can result in two outcomes. We call one of these outcomes a success and other a failure.
- The probability of success, denoted by *P*, is the same on every trial.
- The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.
- The experiment continues until *r* successes are observed, where *r* is specified in advance.

Example of Negative Binomial Experiment:
Consider the following statistical experiment. You flip a coin repeatedly and count the number of times the coin lands on heads. You continue flipping the coin until it has landed 5 times on heads. This is a negative binomial experiment

**Geometric Distribution – Special case of Negative Binomial Distribution.**

# Probability and Distributions

## Discrete Variables – Negative Binomial Distribution

**Notations:**

***x*:** The number of trials required to produce *r* successes in a negative binomial experiment.
***r*:** The number of successes in the negative binomial experiment.
***P*:** The probability of success on an individual trial.
***Q*:** The probability of failure on an individual trial. (This is equal to 1 - *P*.)
**b*(*x*; *r, P*):** Negative binomial probability - the probability that an *x*-trial negative binomial experiment results in the *rth* success on the *xth* trial, when the probability of success on an individual trial is *P*.
$_nC_r$**:** The number of combinations of *n* things, taken *r* at a time.

A **negative binomial random variable** is the number *X* of repeated trials to produce *r* successes in a negative binomial experiment.
The negative binomial distribution is also known as the **Pascal distribution**.

# Probability and Distributions

## Discrete Variables – Negative Binomial Distribution

**Example:**
Suppose we flip a coin repeatedly and count the number of heads (successes). If we continue flipping the coin until it has landed 2 times on heads, we are conducting a negative binomial experiment.

| Number of coin flips | Probability |
|---|---|
| 2 | 0.25 |
| 3 | 0.25 |
| 4 | 0.1875 |
| 5 | 0.125 |
| 6 | 0.078125 |
| 7 or more | 0.109375 |

The **negative binomial probability** refers to the probability that a negative binomial experiment results in *r* - 1 successes after trial *x* - 1 and *r* successes after trial *x*.

$$b^*(x; r, P) = {}_{x-1}C_{r-1} * P^r * (1 - P)^{x - r}$$

# Probability and Distributions

## Discrete Variables – Geometric Distribution

The **geometric distribution** is a special case of the negative binomial distribution. It deals with the number of trials required for a single success. Thus, the geometric distribution is negative binomial distribution where the number of successes (*r*) is equal to 1.

**Example:** Tossing a coin until it lands on heads. We might ask: What is the probability that the first head occurs on the third flip? That probability is referred to as a **geometric probability** and is denoted by g(*x*; *P*).

$$g(x; P) = P * Q^{x-1}$$

**Mean of Negative Binomial Distribution:**
If we define the mean of the negative binomial distribution as the average number of trials required to produce r successes, then the mean is equal to:

$$\mu = r / P$$

# Probability and Distributions

## Discrete Variables – Negative Binomial Distribution Examples

**Example 1:**
Avi is a high school basketball player. He is a 70% free throw shooter. That means his probability of making a free throw is 0.70. During the season, what is the probability that Avi makes his third free throw on his fifth shot?

**Sol:**
This is an example of a negative binomial experiment. The probability of success (*P*) is 0.70, the number of trials (*x*) is 5, and the number of successes (*r*) is 3.

To solve this problem, we enter these values into the negative binomial formula.

$$b^*(x; r, P) = {}_{x-1}C_{r-1} * P^r * Q^{x-r}$$

$$b^*(5; 3, 0.7) = {}_4C_2 * 0.7^3 * 0.3^2$$

$$b^*(5; 3, 0.7) = 6 * 0.343 * 0.09 = 0.18522$$

Thus, the probability that Avi will make his third successful free throw on his fifth shot is 0.18522.

# Probability and Distributions

## Discrete Variables – Negative Binomial Distribution Examples

**Example 2:**
What is the probability that Avi makes his first free throw on his fifth shot?

**Sol:**
This is an example of a geometric distribution, which is a special case of a negative binomial distribution.

The probability of success (*P*) is 0.70, the number of trials (*x*) is 5, and the number of successes (*r*) is 1. We enter these values into the negative binomial formula.

$$b^*(x; r, P) = {}_{x-1}C_{r-1} * P^r * Q^{x-r}$$
$$b^*(5; 1, 0.7) = {}_4C_0 * 0.7^1 * 0.3^4$$
$$b^*(5; 3, 0.7) = 0.00567$$

Using Geometric formula:

$$g(x; P) = P * Q^{x-1}$$
$$g(5; 0.7) = 0.7 * 0.3^4 = 0.00567$$

# Probability and Distributions

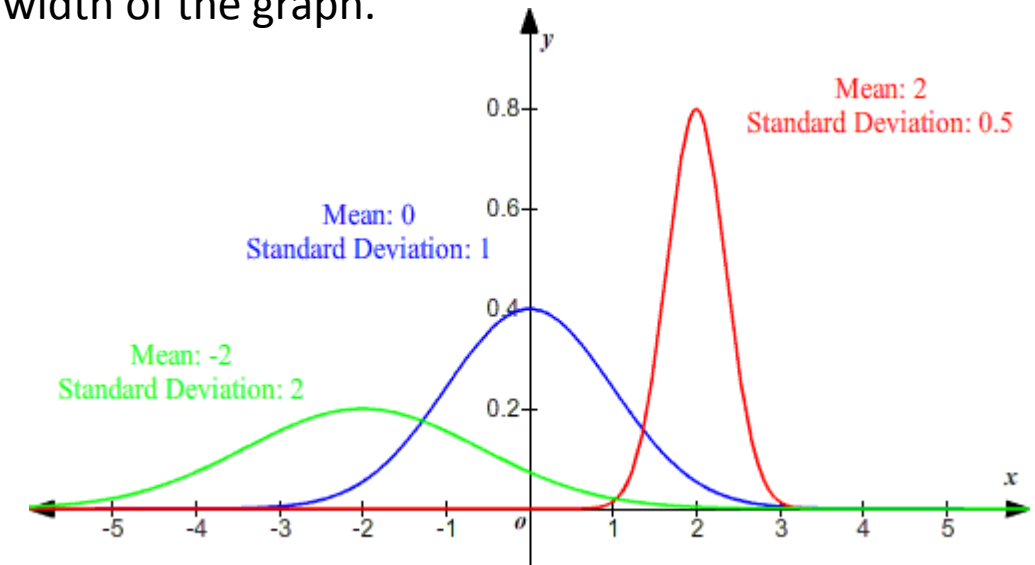
## Continuous Variables – Normal Distribution

The Normal Equation –

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2}$$

where, *X* is a normal random variable,
μ is the mean, σ is the standard deviation,
π is approximately 3.14, and *e* is approximately 2.71.

The random variable *X* in the normal equation is called the **normal random variable**. The normal equation is the probability density function for the normal distribution.
The graph of the normal distribution depends on two factors - **the mean and the standard deviation**.
The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph.

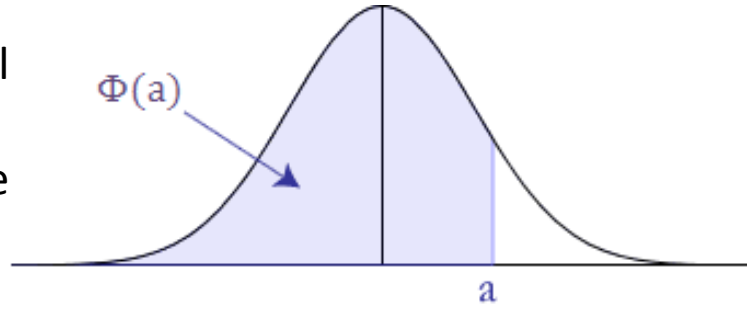


# Probability and Distributions

## Continuous Variables – Normal Distribution

**Probability and the Normal Curve:**
The normal distribution is a continuous probability distribution. This has several implications for probability.

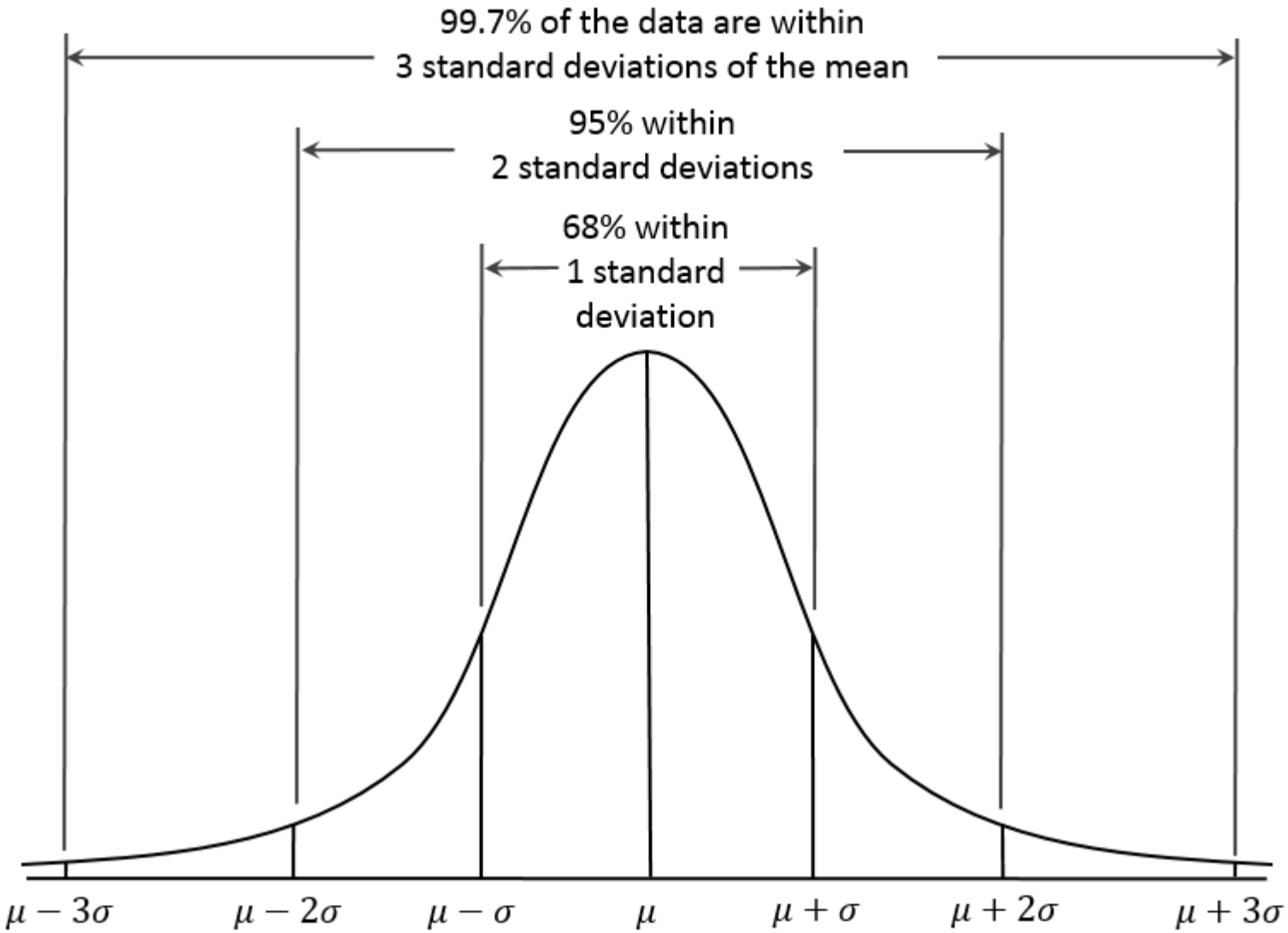
- The total area under the normal curve is equal to 1.
- The probability that *X* is greater than *a* equals the area under the normal curve bounded by *a* and plus infinity (as indicated by the *non-shaded* area).
- The probability that *X* is less than *a* equals the area under the normal curve bounded by *a* and minus infinity (as indicated by the *shaded* area).




Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following rules:

1. About 68% of the area under the curve falls within 1 standard deviation of the mean.
2. About 95% of the area under the curve falls within 2 standard deviations of the mean.
3. About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

# Probability and Distributions



# Probability and Distributions

## Continuous Variables – Normal Distribution

**Exercise 1:**
An average light bulb manufactured by the Acme Corporation lasts 300 days with a standard deviation of 50 days. Assuming that bulb life is normally distributed, what is the probability that an Acme light bulb will last at most 365 days?

**Sol:** The value of the normal random variable is 365 days, the mean is equal to 300 days and the standard deviation is equal to 50 days.

We enter these values into the Normal Distribution Calculator and compute the cumulative probability. The answer is: $P( X \leq 365) = 0.90$. Hence, there is a 90% chance that a light bulb will burn out within 365 days.

**Exercise 2:**
Suppose scores on an IQ test are normally distributed. If the test has a mean of 100 and a standard deviation of 10, what is the probability that a person who takes the test will score between 90 and 110?

**Sol:** $P( 90 < X < 110 ) = P( X < 110 ) - P( X < 90 )$

# Probability and Distributions

## Continuous Variables – Standard Normal Distribution

The **standard normal distribution** is a special case of the normal distribution when a normal random variable has a mean of zero and a standard deviation of one.

The normal random variable of a standard normal distribution is called a **standard score** or a **z-score**. Every normal random variable *X* can be transformed into a *z* score via the following equation:

$$z = (X - \mu) / \sigma$$

A **standard normal distribution table** shows a cumulative probability associated with a particular z-score. Table rows show the whole number and tenths place of the z-score. Table columns show the hundredths place. The cumulative probability (often from minus infinity to the z-score) appears in the cell of the table.

We may not be interested in the probability that a standard normal random variable falls between minus infinity and a given value. We may want to know the probability that it lies between a given value and plus infinity. Or we may want to know the probability that a standard normal random variable lies between two given values. These probabilities are easy to compute from a normal distribution table.

# Probability and Distributions

## Continuous Variables – Standard Normal Distribution

**Calculating probabilities from a standard normal table**:
A section of the standard normal table is reproduced below.

To find the cumulative probability of a z-score equal to -1.31, cross-reference the row of the table containing -1.3 with the column containing 0.01. The table shows that the probability that a standard normal random variable will be less than -1.31 is 0.0951; that is, P(Z < -1.31) = 0.0951.

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|---|---|---|---|---|---|---|
| -3.0 | 0.0013 | 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 |
| ... | ... | ... | ... | ... | ... | ... |
| -1.4 | 0.0808 | 0.0793 | 0.0778 | 0.0764 | 0.0749 | 0.0735 |
| -1.3 | 0.0968 | 0.0951 | 0.0934 | 0.0918 | 0.0901 | 0.0885 |
| -1.2 | 0.1151 | 0.1131 | 0.1112 | 0.1093 | 0.1075 | 0.1056 |
| ... | ... | ... | ... | ... | ... | ... |
| 3.0 | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 |

# Probability and Distributions

## Continuous Variables – Standard Normal Distribution

**Calculating probabilities from a standard normal table**:
Finding P(Z > a). The probability that a standard normal random variable (z) is greater than a given value (a) is easy to find. The table shows the P(Z < a). P(Z > a) = 1 - P(Z < a).

Example, that we want to know the probability that a z-score will be greater than 3.00. From the table, we find that P(Z < 3.00) = 0.9987. Therefore, P(Z > 3.00) = 1 - P(Z < 3.00) = 1 - 0.9987 = 0.0013.

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|---|---|---|---|---|---|---|
| -3.0 | 0.0013 | 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 |
| ... | ... | ... | ... | ... | ... | ... |
| -1.4 | 0.0808 | 0.0793 | 0.0778 | 0.0764 | 0.0749 | 0.0735 |
| -1.3 | 0.0968 | 0.0951 | 0.0934 | 0.0918 | 0.0901 | 0.0885 |
| -1.2 | 0.1151 | 0.1131 | 0.1112 | 0.1093 | 0.1075 | 0.1056 |
| ... | ... | ... | ... | ... | ... | ... |
| 3.0 | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 |

# Probability and Distributions

## Continuous Variables – Standard Normal Distribution

**Calculating probabilities from a standard normal table**:
Finding P(a < Z < b). The probability that a standard normal random variables lies between two values is also easy to find. The P(a < Z < b) = P(Z < b) - P(Z < a).

Example, we want to know the probability that a z-score will be greater than -1.40 and less than -1.20. From the table, we find that P(Z < -1.20) = 0.1151; and P(Z < -1.40) = 0.0808.
Therefore, P(-1.40 < Z < -1.20) = P(Z < -1.20) - P(Z < -1.40) = 0.1151 - 0.0808 = 0.0343.

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|---|---|---|---|---|---|---|
| -3.0 | 0.0013 | 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 |
| ... | ... | ... | ... | ... | ... | ... |
| -1.4 | 0.0808 | 0.0793 | 0.0778 | 0.0764 | 0.0749 | 0.0735 |
| -1.3 | 0.0968 | 0.0951 | 0.0934 | 0.0918 | 0.0901 | 0.0885 |
| -1.2 | 0.1151 | 0.1131 | 0.1112 | 0.1093 | 0.1075 | 0.1056 |
| ... | ... | ... | ... | ... | ... | ... |
| 3.0 | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 |

# Probability and Distributions

## Continuous Variables – Standard Normal Distribution

**Exercise 1.**
Molly earned a score of 940 on a national achievement test. The mean test score was 850 with a standard deviation of 100. What proportion of students had a higher score than Molly? (Assume that test scores are normally distributed.)

**Sol.** First, we transform Molly's test score into a z-score, using the z-score transformation equation.
$z = (X - \mu) / \sigma = (940 - 850) / 100 = 0.90$
Then, using the standard normal distribution table, we find the cumulative probability associated with the z-score. In this case, we find $P(Z < 0.90) = 0.8159$.

Therefore, the $P(Z > 0.90) = 1 - P(Z < 0.90) = 1 - 0.8159 = 0.1841$.
Thus, we estimate that 18.41 percent of the students tested had a higher score than Molly.