Statistics

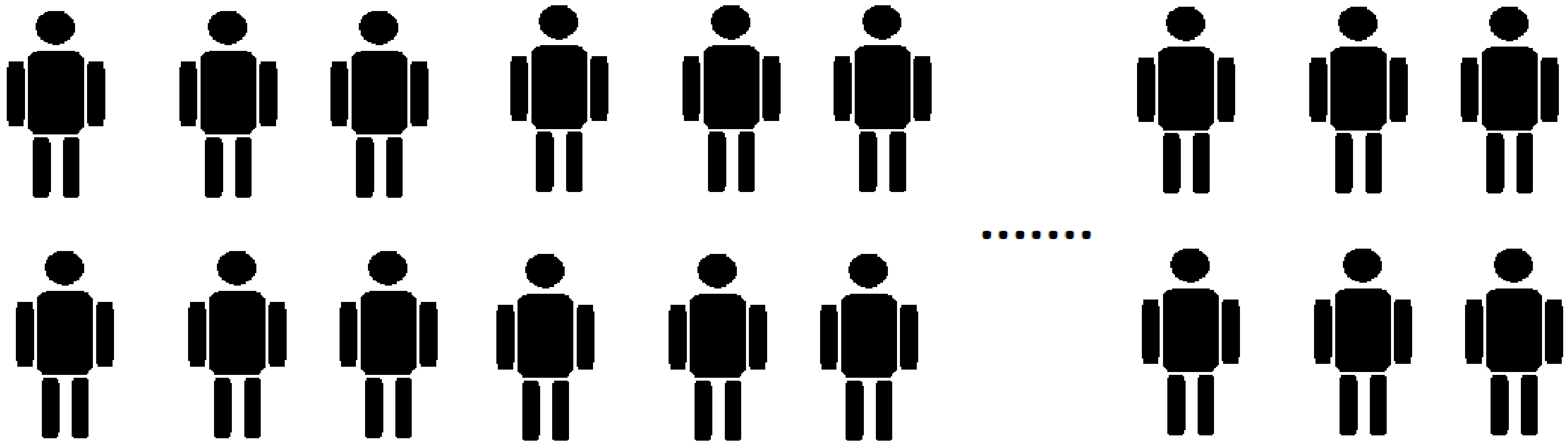
Agenda:

1	Introduction to Statistics
2	Statistical Measures
3	Population VS Sample
4	Statistics using Pandas
5	Random Variable
6	Expected Value
7	Data Distribution
8	Quartiles
9	Covariance & Correlation
10	Sample_Space, Events, Trials, & Experiments
11	Independent & dependent Events

1. Introduction to Statistics

What is Statistics?

- Statistics is the science of summarizing and describing the data.
- For example:
  - Suppose you have a dataset that contains about 100,000,000 observations about Egyptian people height.



What is Statistics?

- If you want to describe how high Egyptian people are, you don’t tell the height of each single person of the 100,000,000 people in the Egyptian population! But instead, you simply say “The average height of the Egyptian people is 170cm”.

- What you have just done is that you summarized the 100,000,000 observations into one number, 170cm, which we call a statistical measure.

2. Statistical Measures

Statistical Measures:

- A Statistical Measure is a number, that is calculated to summarize many records(rows) of information into one single value.
- Statistical measures can be used to get statistical inference about the population.
- Since statistical measures are related to data, let's first understand types of the data. Data can be:
  - Continuous(Numerical).
  - Or Discrete(Categorical).

Continuous Vs Discrete:

Continuous Data	Discrete Data
➢ Is the data that has infinite number of possible values. ➢ Also known as **Numerical data**. ➢ Continuous data could be: ➢ Float dtypes; such as, Salary or Weight. ➢ Int dtypes that have large number of possible unique values; such as, number-of-hours-played.	➢ Is the data that has finite number of possible values ➢ Also known as **Categorical data**. ➢ Continuous data could be: ➢ String dtypes; such as, City-name. ➢ Int dtypes that have small number of possible unique values; such as, number-of-children.

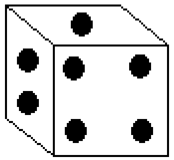
Popular Statistical Measures:

1. Probability.
2. Measures of Central Tendency.
3. Measures of dispersion (Deviation).

Probability:

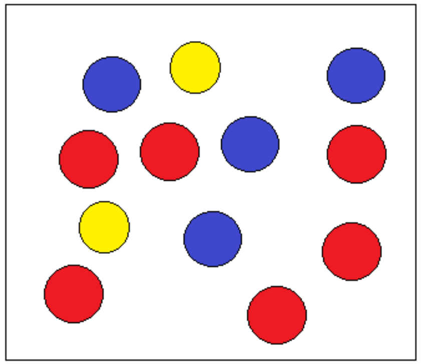
- Is the ratio between frequency of the unique-value & total number of samples.

- Example1, suppose you have a dice:



  - The unique possible values are; 1, 2, 3, 4, 5, 6.
  - Probability of 1 = 1 / 6 = .167

- Example1, suppose you have the box of balls on the right:



  - The unique possible values are; blue, red, yellow.
  - Probability of blue = 4 / 12 = .333

Measures of Central Tendency:

- Are the measures used to represent the average values of the data we have.
- There are three main measures of central tendency:
  - Mean.
  - Median.
  - Model.

- Mean & Median are used to summarize Numerical data, while Mode is used to summarize categorical data.

Measures of Central Tendency (Mean):

- Mean is the ratio between the summation of all values and total number of observation in the data.
- For example, suppose you have the following set of observation:
  - [5, 2, 3, 10, 20].
  - Mean = (5+2+3+10+20) / 5 = 8.
- Mean is used with numerical data that doesn’t contain extreme values (outliers), because mean is sensitive to outliers.
- We use symbol μ to represent the mean.

Measures of Central Tendency (Median):

- Median is the middle value in the data after being sorted.
- Steps:
  - First sort the data, then Find the number in the middle, and this is your Median. If there are two number in the middle, then the Median is the average between them.
- Median is used with numerical data that contains outliers.

Example1	Example2
➢ Suppose you have this set of observations: [5, 2, 3, 10, 20] . ➢ First sort them → [2, 3, 5, 10, 20]. ➢ Median = 5.	➢ Suppose you have this set of observations: [3, 5, 2, 3, 10, 20] . ➢ First sort them → [2 , 3, 3, 5, 10, 20]. ➢ Median = (3+5) / 2 = 4.

Measures of Central Tendency (Mode):

- Mode is the most frequent value in the data.
- Mode is used with categorical data.

Example1	Example2
➢ Suppose you have this set of observations: [5, 2, 3, 3, 2, 3, 1, 5, 9, 8, 3, 1, 7, 6]. ➢ Mode= 5.	➢ Suppose you have this set of observations: [“Cairo”, “Alex”, “Aswan”, “Alex”, “Alex”, “Mansoura”, “Alex”, “Cairo”]. ➢ Mode = “Alex”.

Measures of Dispersion:

- Are measures used to measure the spread of the data.
- Also Called Measures of Deviation.
- For example, suppose you have the following two sets of numbers:
  - Set1 = [5, 5, 5, 5, 5] & Set2 = [-5, 0, 5, 10, 15].
  - The two sets contains the same value of mean = 5.
  - But as you can see Set2 has more spread than set1.
  - So, we need a way to measure the amount of spread.

Measures of Dispersion:

- There are two main measures of Dispersion:
  - Variance.
  - Standard deviation.

- Standard Deviation is the most used as a measure of dispersion, that’s why we call it standard, however variance is a popular measure too and has its applications.

Measures of Dispersion (Variance):

- Is the average of all differences between each value in the data & the mean of this data.
- $\sigma 2$ is used to represent the Variance.
- Formula: $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$, where X_i represents the i^{th} value in the data, and N represents total number of values.

Measures of Dispersion (Variance):

Example1	Example2
➢ Data = [5, 5, 5, 5, 5] . ➢ μ = (5+5+5+5+5) / 5 = 5. ➢ σ2 = $((5-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2) / 5 = 0$. ➢ Variance = 0	➢ Data = [-5, 0, 5, 10, 15]. ➢ μ = (-5+0+5+10+15) / 5 = 5. ➢ σ2 = $((5--5)^2 + (5-0)^2 + (5-5)^2 + (5-10)^2 + (5-15)^2) / 5 = 50$. ➢ Variance = 50.

Measures of Dispersion (Standard Deviation):

- Is the square root of the variance.
- σ is used to represent the Standard deviation.
- Formula: $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$, where X_i represents the i^{th} value in the data, and N represents total number of values.
- Standard deviation is always preferred over variance as a measure of dispersion, and the reason is that unlike variance, standard deviation is not sensitive to outliers.

Measures of Dispersion (Standard Deviation):

Example1	Example2
➢ Data = [5, 5, 5, 5, 5] .	➢ Data = [-5, 0, 5, 10, 15].
➢ $\mu = (5+5+5+5+5) / 5 = 5$.	➢ $\mu = (-5+0+5+10+15) / 5 = 5$.
➢ $\sigma 2 = ((5-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2) / 5 = 0$.	➢ $\sigma 2 = ((5--5)^2 + (5-0)^2 + (5-5)^2 + (5-10)^2 + (5-15)^2) / 5 = 50$.
➢ $\sigma = \surd\sigma 2 = \surd 0 = 0$.	➢ $\sigma = \surd\sigma 2 = \surd 50 = 7.07$
➢ Standard deviation = 0.	➢ Standard deviation = 7.07

3. Population Vs Sample

What is Population?

- Population is the whole complete set of observation.
- For example:
  - In Egypt, we have 100,000,000 people if we could collect 100,000,000 observations about their heights, then the population = heights-of-100,000,000-people.
- But could we really collect this huge number of observations? Do we have the resources(money & time) to do this?!
- The answer is No! and here comes the concept of Sample.

What is Sample?

- A sample is a randomly chosen subset from the population, that represents the whole set of observations without having to actually deal with the whole population.
- For example:
  - In Egypt, we could represent the 100,000,000 people with only 1000,000 observations collected randomly.
- The larger the sample is, the more strongly it represents the population, but the harder to collect and work on.

4. Statistics using Pandas

What is Pandas?

- You can apply statistics using Numpy or Pandas.
- Pandas is a library built on Numpy, which is more suitable for dealing with tabular datasets.
- In Pandas tabular data is read as DataFrame which is the main datatype in pandas that represents matrix.
- In pandas, vectors are represented by a datatype called Series.
- Each row or column in the DataFrame is a Series.

Reading Tabular Data:

➢ Tabular datasets come in two main file formats:

CSV files

```python
import pandas as pd
df = pd.read_csv("file.csv")
df
```

	Length	Width	City	Price
0	20	10	Cairo	5000000
1	15	15	Alex	4000000
2	30	20	Aswan	1500000
3	10	50	Alex	8000000
4	5	15	Giza	800000
5	12	10	Alex	1000000
6	5	30	Luxor	500000
7	7	20	Aswan	700000
8	20	40	Alex	9000000
9	8	20	Cairo	900000
10	6	14	Giza	6000000

XLSX files

```python
import pandas as pd
df = pd.read_excel("file.xlsx")
df
```

	Length	Width	City	Price
0	20	10	Cairo	5000000
1	15	15	Alex	4000000
2	30	20	Aswan	1500000
3	10	50	Alex	8000000
4	5	15	Giza	800000
5	12	10	Alex	1000000
6	5	30	Luxor	500000
7	7	20	Aswan	700000
8	20	40	Alex	9000000
9	8	20	Cairo	900000
10	6	14	Giza	6000000

Pandas for Statistics:

Mean of Length Column	Median of Length Column	Mode of City Column
`df.Length.mean()` 12.545454545454545	`df.Length.median()` 10.0	`df.City.mode()` 0 Alex
Variance of Length Column	**Standard-Deviation of Length column**	
`df.Length.var()` 63.672727272727265	`df.Length.std()` 7.979519238195198	

5. Random Variables

What is Random Variable?

- A Random Variable is a writing style we use to write the data in a way that helps us make good notation.
- For example:
  - suppose you have data about number of children; [180, 200, 150, 160, 152, 179, 168].
  - We could just say X = [3, 1, 3, 4, 3, 2, 3, 1, 5, 3], where X is a Random Variable.
- Random Variable could be Numeric or Categorical.

Why is Random Variable Useful?

- It helps us to simplify the writing style.
- Now we can just say "P(X=3) = .5", instead of having to say "probability-of-number-of-children = 3 is .5".
- Or, we can just say "Mean(X) = 2.8", instead of having to say "mean-of-number-of-children = 2.8".

Random Variable Real-Life Example:

- You can consider each column in the dataset to be a random variable.

- For example, in the Dataset on the right, Length column could be considered a random variable.

	Length	Width	City	Price
0	20	10	Cairo	5000000
1	15	15	Alex	4000000
2	30	20	Aswan	1500000
3	10	50	Alex	8000000
4	5	15	Giza	800000
5	12	10	Alex	1000000
6	5	30	Luxor	500000
7	7	20	Aswan	700000
8	20	40	Alex	9000000
9	8	20	Cairo	900000
10	6	14	Giza	6000000

Pandas DataFrame Columns:

Read the DataSet

```python
import pandas as pd
df = pd.read_csv("file.csv")
df
```

	Length	Width	City	Price
0	20	10	Cairo	5000000
1	15	15	Alex	4000000
2	30	20	Aswan	1500000
3	10	50	Alex	8000000
4	5	15	Giza	800000
5	12	10	Alex	1000000
6	5	30	Luxor	500000
7	7	20	Aswan	700000
8	20	40	Alex	9000000
9	8	20	Cairo	900000
10	6	14	Giza	6000000

Access the column

```python
random_variable1 = df.Length
print(random_variable1)
print()
print(random_variable1.mean())
print(random_variable1.median())
print(random_variable1.std())
```

```
0     20
1     15
2     30
3     10
4      5
5     12
6      5
7      7
8     20
9      8
10     6
Name: Length, dtype: int64

12.545454545454545
10.0
7.979519238195198
```

6. Expected Value

What is Expected Value?

- Is the same as Mean but calculated in a different way.
- E is used to represented Expected value.
- Expected value of a Random Variable X is calculated by multiplying each value of the random variable by its probability and add the products.
- The formula is: $\sum_i X_i * P(X_i)$, where X_i is the i^{th} value in the random variable X.

Expected Value Example:

- Suppose the following Random Variable:
  - X = [0, 5, 5, 5, 10, 0, 5, 10, 5].
  - $P(X=0) = 2/9 = .222$
  - $P(X=5) = 5/9 = .556$
  - $P(X=10) = 2/9 = .222$
  - $E(X) = 0 * .222 + 5 * .556 + 10 * .222 = 5.$

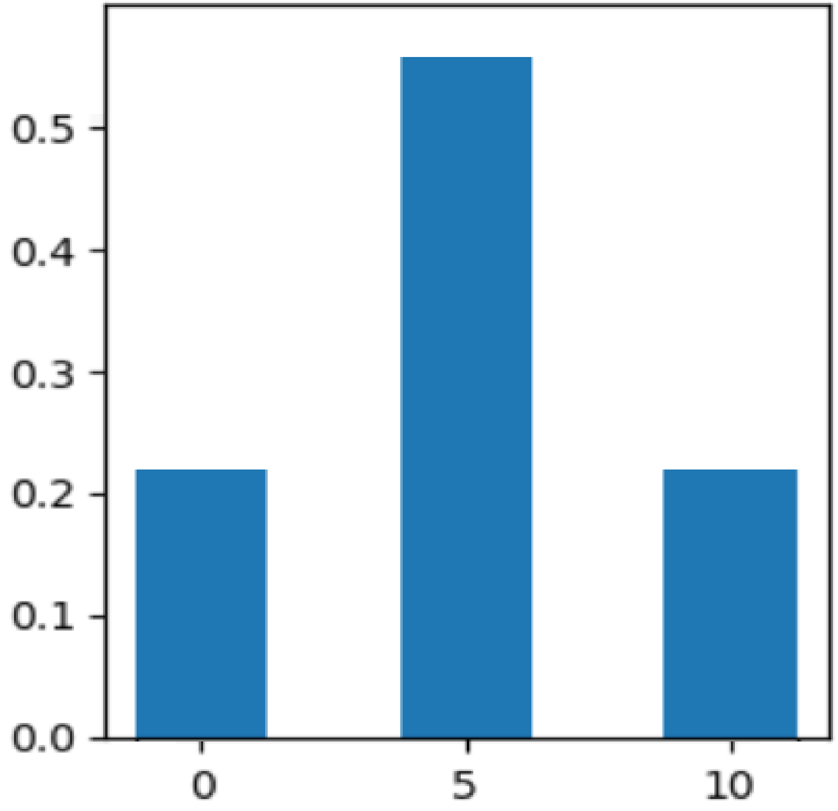
7. Data Distribution

What is Data Distribution?

- Data Distribution is a way to describes how the observations are distributed or spread across the unique values of the data.
- In other words, Data Distribution represents how much each unique value occurs in the data or how frequent each unique value is.

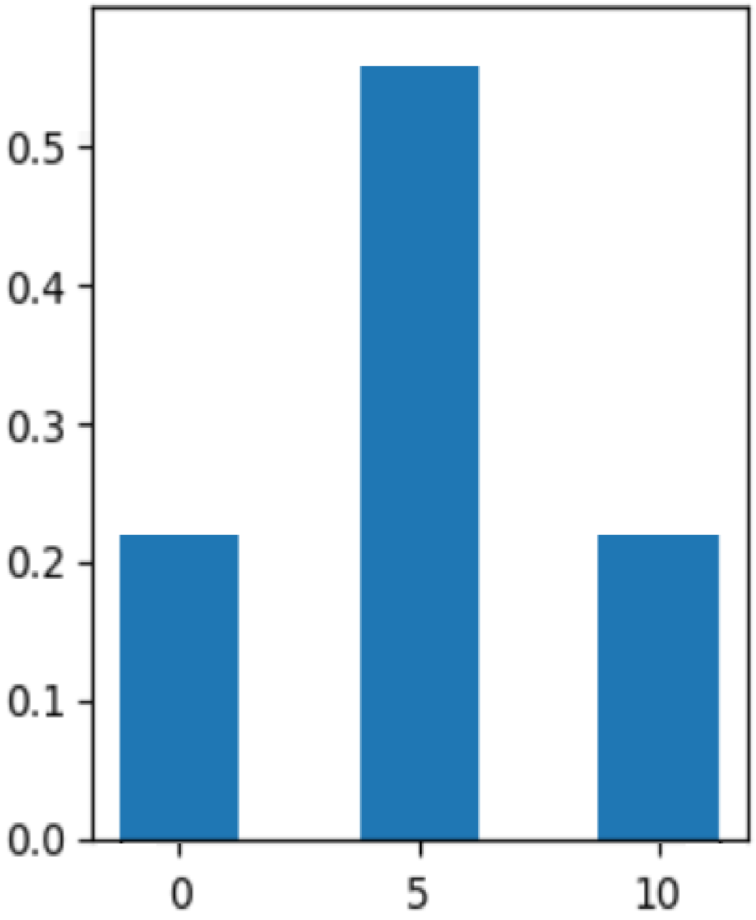
Data Distribution Example:

- ➢ If you have Random Variable X = [0, 5, 5, 5, 10, 0, 5, 10, 5].
- ➢ Then the data distribution of this random variable is distributed as following:
  - ➢ 22.2% of the data belong to (X=0).
  - ➢ 55.6% of the data belong to (X=5).
  - ➢ 22.2% of the data belong to (X=0).



Data Distribution Histogram:

- It's common to represent the data distribution as a graph called Histogram.
- A histogram is a 2-dimensional graph, where:
  - X-axis represents the unique values in the Random Variable.
  - Y-axis represents the probability of each unique value.
  - Each unique value has a bar (rectangle) whose height is equal to the probability.

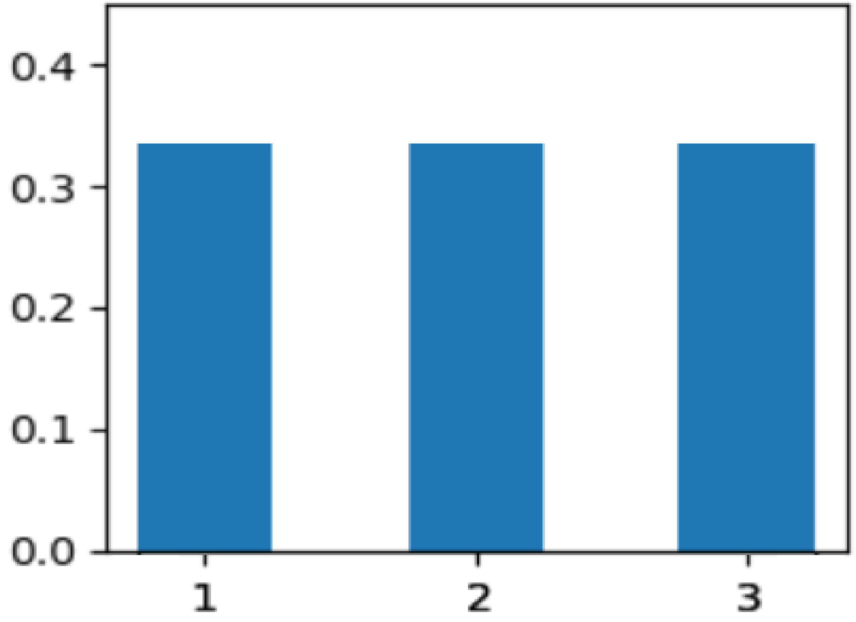


Data Distribution Types:

- There are so many types of data distribution, however we will cover the most important & most popular ones:
  - Uniform Distribution.
  - Normal Distribution.
  - Right-Skewed Distribution.
  - Left-Skewed Distribution.

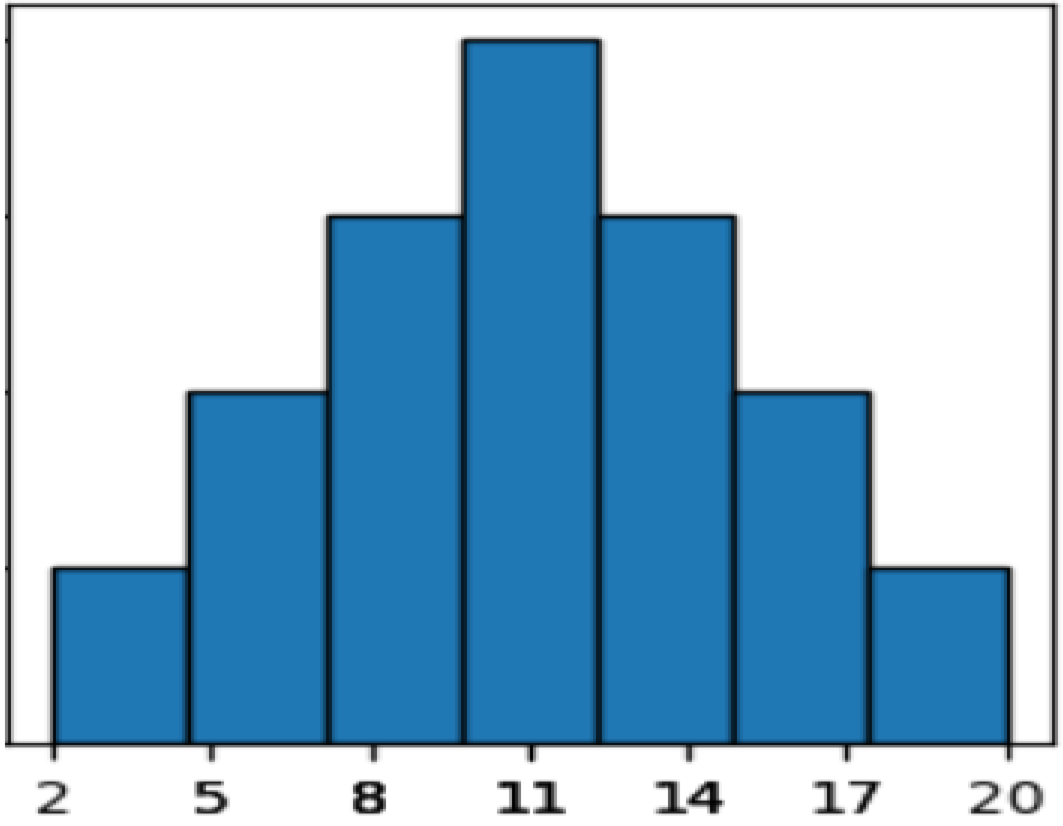
Uniform Distribution:

- Is Data Distribution where observations are equally distributed among the unique values. In other words, all the unique values occur equally with the same frequency.
- For example, Suppose you have X = [1, 2, 2, 3, 1, 3], then the distribution is:
  - 33.3% of the data belong to (X=1).
  - 33.3% of the data belong to (X=2).
  - 33.3% of the data belong to (X=3).



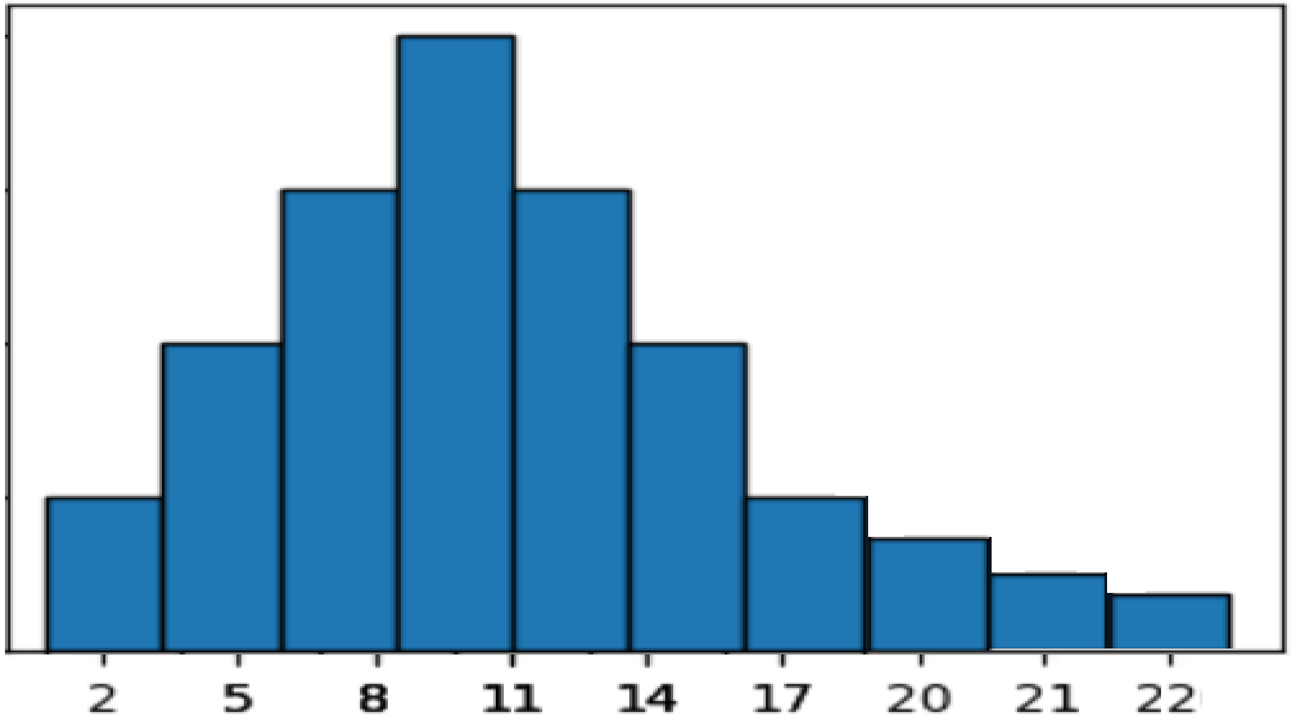
Normal Distribution:

- Is Data Distribution where observations are distributed around the mean the most, with fewer values occurring farther away from the mean in both directions.
- The distribution histogram takes a shape of symmetric bell.



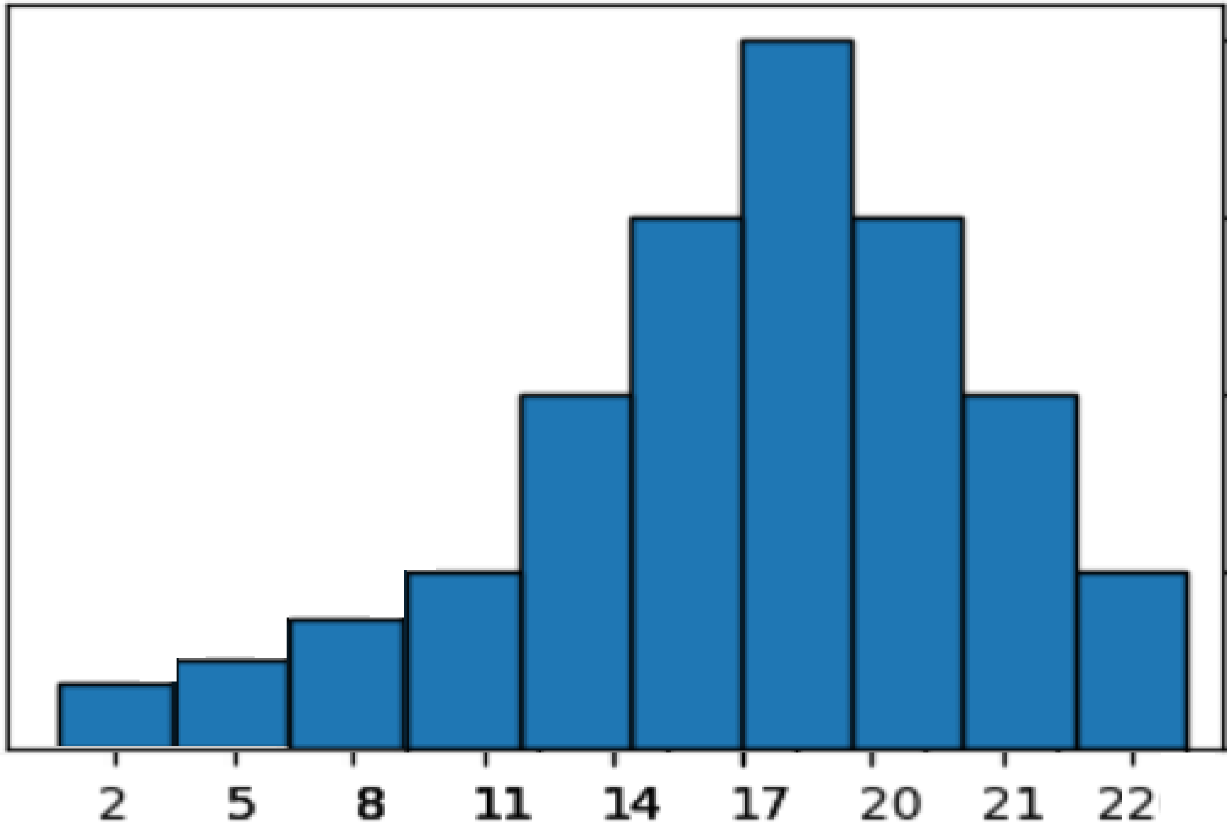
Right-Skewed Distribution:

- Is Data Distribution where observation are mostly distributed around mean and left side to the mean, with few observations at the extreme right to the mean.



Left-Skewed Distribution:

- Is Data Distribution where observation are mostly distributed around mean and right side to the mean, with few observations at the extreme left to the mean.



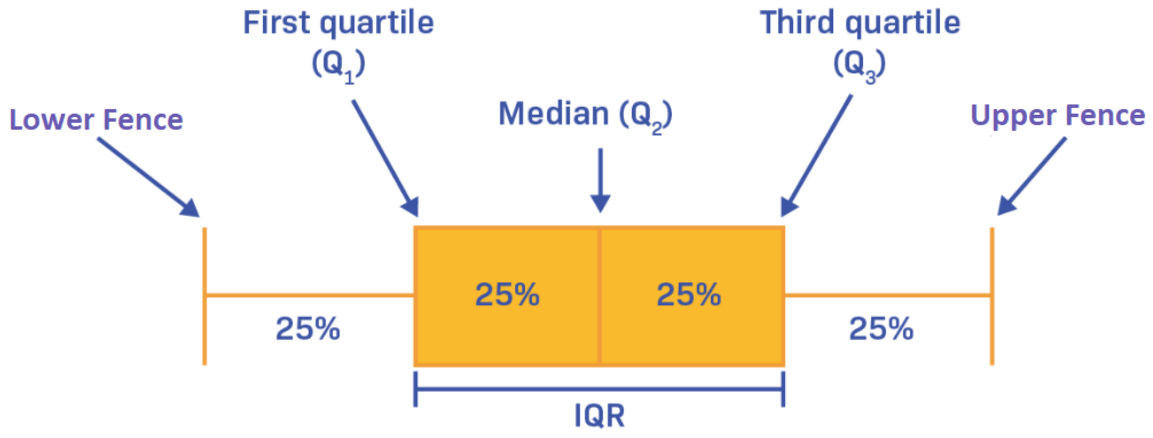
8. Quartiles

What are Quartiles?

- Is a technique used to identify outliers, which are extreme values that occur in the data.
- For example:
  - Suppose you have a random variable X=[20, 30, 10, 50, 180] where X represents people ages.
  - The value 180 is an outlier because it’s a strange or extreme value, since it’s no common to see a 180 years-old person.

What are Quartiles?

- Quartiles are numbers used to detect fences or thresholds, where if a number exceeds these fences, then this number is considered to be an outlier.
- There are three types of quartiles to calculate to be able to calculate the fences. These three quartiles are:
  - First Quartile (Q1).
  - Second Quartile (Q2).
  - Third Quartile (Q3).



How to Calculate Quartiles?

Steps:

1. Sort the Random Variable data.
2. Calculate the median of the Random Variable, and this is your Q2.
3. Calculate the median of the subset right to Q2, and this is your Q1.
4. Calculate the median of the subset left to Q2, and this is your Q3.

Calculate Quartiles Example:

90 33 47 -50 10 19 11 13 16 28 15 19 23 21 44 30 34 36 10 45

1- Sort: -50 10 10 11 13 15 16 19 19 21 23 28 30 33 34 36 44 45 47 90

2- Find Q2: -50 10 10 11 13 15 16 19 19 $\underbrace{21\ \ 23}_{Q2\ =\ 22}$ 28 30 33 34 36 44 45 47 90

Q2 = 22

3- Find Q1 & Q3: -50 10 10 11 $\underbrace{13\ \ 15}_{Q1\ =\ 14}$ 16 19 19 21 | 23 28 30 33 $\underbrace{34\ \ 36}_{Q3\ =\ 35}$ 44 45 47 90

Q1 = 14 Q2 = 22 Q3 = 35

-50 10 10 11 13 | 15 16 19 19 21 | 23 28 30 33 34 | 36 44 45 47 90

Outlier fences:

- There are two fences we need to calculate so that if a number exceed these fences, then it is considered an outlier.
- These two fences are:
  - Upper Fence:
    - If a number is larger than the upper fence, then it is considered an outlier.
  - Lower Fence:
    - If a number is smaller than the lower fence, then it is considered an outlier.

How to Calculate Outlier fences?

- Steps:
  1. Calculate IQR, where IQR = Q3 – Q1.
  2. Calculate Lower-Fence where, Lower-Fence = Q1 – 1.5*IQR.
  3. Calculate Upper-Fence where, Upper-Fence = Q3 + 1.5*IQR.
- Example:

90 33 47 -50 10 19 11 13 16 28 15 19 23 21 44 30 34 36 10 45

	Q1 = 14		Q2 = 22		Q3 = 35	
-50 10 10 11 13	|	15 16 19 19 21	|	23 28 30 33 34	|	36 44 45 47 90

IQR = Q3 - Q1 = 35 - 14 = 21

Lower-Fence = Q1 - 1.5 * IQR = 14 - 1.5 * 21 = -17.5

Upper-Fence = Q3 + 1.5 * IQR = 35 + 1.5 * 21 = 66.5

-50 is an outlier, because it is < Lower-Fence ==> (-50 < -17.5)

90 is an outlier, because it is > Upper-Fence ==> (90 > 66.5)

Detect Outliers Using Pandas:

```
random_variable1 = df.Length
Q1 = random_variable1.quantile(.25)
Q3 = random_variable1.quantile(.75)
IQR = Q3 - Q1
Lower_Fence = Q1 - 1.5 * IQR
Upper_Fence = Q3 + 1.5 * IQR
print(f"Any number (< {Lower_Fence}, or > {Upper_Fence}) is an outlier")
```

```
Any number (< -10.0, or > 34.0) is an outlier
```

	Length	Width	City	Price
0	20	10	Cairo	5000000
1	15	15	Alex	4000000
2	30	20	Aswan	1500000
3	10	50	Alex	8000000
4	5	15	Giza	800000
5	12	10	Alex	1000000
6	5	30	Luxor	500000
7	7	20	Aswan	700000
8	20	40	Alex	9000000
9	8	20	Cairo	900000
10	6	14	Giza	6000000

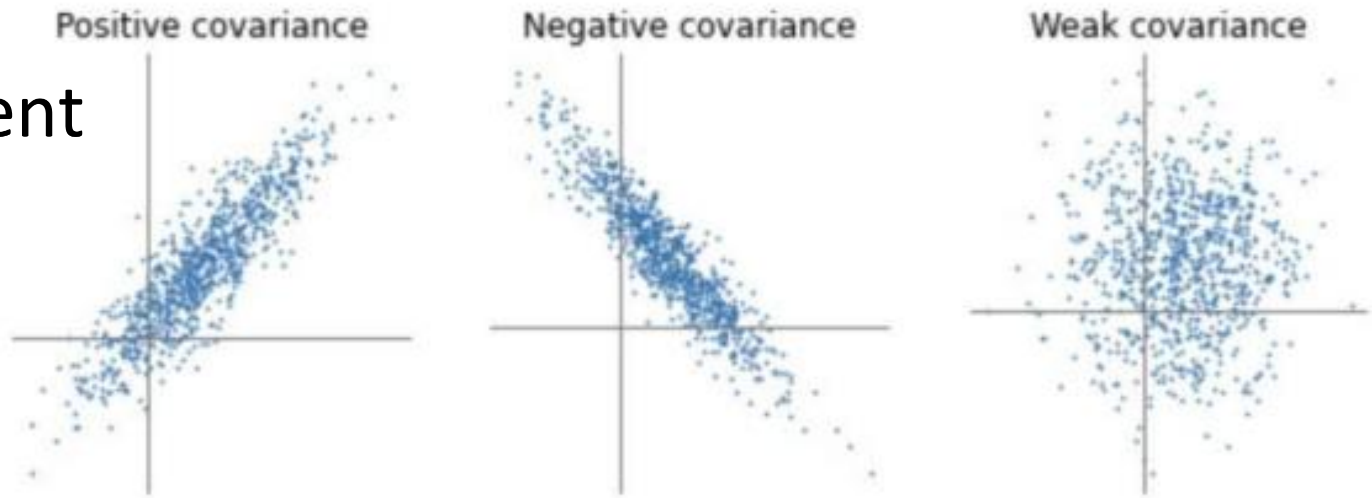
9. Covariance & Correlation

What is Covariance?

- Is a Statistical measure used to describe how much two variables change together.
- For example, suppose you have two random variables X & Y:
  - If Covariance is highly positive, then the relation between them is Positive, which means if X increases, then Y increases also.
  - If Covariance is highly negative, then the relation between them is Negative, which means if X increases, then Y decreases.
  - If Covariance is near to zero, then the relation is weak or there is no relation.

What is Covariance?

- Covariance can also be defined as "How much the deviation of one variable(X) from its mean is (related/or similar) to the deviation of another variable(Y) from its mean".
- The deviation of a random variable from its mean represent the amount of change and the direction of this change also.
- Cov(X, Y) is used to represent covariance between X & Y.



How to Calculate Covariance?

- Formula:
  - $Cov(X, Y) = \sum_{i=1}^{n}((X_i - \mu_x) * (Y_i - \mu_y)) \,/\, n$.
  - n is the number of samples.
  - μ_x is the mean of Random Variable X.
  - μ_y is the mean of Random Variable Y.
- Example:

X = [1, 2, 3, 4, 5, 6, 7, 8, 9] **Y = [9, 8, 7, 6, 5, 4, 3, 2, 1]**

$\mu_x = 5$ $\mu_y = 5$ $n = 9$

$$Cov(X, Y) = ((1-5)*(9-5) + (2-5)*(8-5) + (3-5)*(7-5) + (4-5)*(6-5) + (5-5)*(5-5) + (6-5)*(4-5) + (7-5)*(3-5) + (8-5)*(2-5) + (9-5)*(1-5) +)/n$$
$$= -6.667$$

Result: $Cov(X, Y) = -6.667 < 0$.

Conclusion: The relation between X & Y is Negative.

What is Correlation?

- Is a Statistical measure that is the same as Covariance, except that Correlation is normalized, which give us sense about the relation strength.
- Normalized means that Correlation has values in range = [-1:1].
- For example, suppose you have two random variables X & Y:
  - If Correlation is near to 1, then the relation between them is Strong Positive. While If Correlation is near to 2, then the relation between them is Strong Negative.
  - If Correlation is near to 0, then the relation is weak.

Correlation Vs Covariance:

- Correlation has values in range [-1 : 1]. While Covariance had values between [∞, -∞].
- Having a range between -1 & 1 is very useful since this helps us know how much strong is the relation between the two variables.
- This is useful if I want to compare two relations. While in covariance this is not possible.
- Example:

Correlation	Covariance
➢ Relation1 = .5 ➢ Relation2 = .25 ➢ Relation1 is twice strong as Relation2.	➢ Relation1 = 5 ➢ Relation2 = 2.5 ➢ You can’t tell how much Relation1 is stronger than Relation2.

How to Calculate Correlation?

- Formula:
  - Corr(X, Y) = Cov(X, Y)/($\sigma_x * \sigma_y$).
  - σ_x is the Standard-deviation of Random Variable X.
  - σ_y is the Standard-deviation of Random Variable Y.
- Example:

X = [1, 2, 3, 4, 5, 6, 7, 8, 9] Y = [9, 8, 7, 6, 5, 4, 3, 2, 1]

$\sigma_X = 2.582$ $\sigma_Y = 2.582$

Cov(X, Y) = -6.667

Corr(X, Y) = Cov(X, Y) / ($\sigma_X * \sigma_Y$) = -6.667 / (2.582 * 2.582) = -1

Result: Corr(X, Y) = -1.

Conclusion: The relation between X & Y is Negative.

Covariance & Correlation using Pandas:

Covariance Matrix

- Get the covariance between all the pairs of columns in the DataFrame.

```
random_variable1 = df.Length
random_variable2 = df.Width
df.cov()
```

	Length	Width	Price
Length	6.367273e+01	-7.090909e-01	6.240000e+06
Width	-7.090909e-01	1.633636e+02	2.234000e+07
Price	6.240000e+06	2.234000e+07	1.002800e+13

Correlation Matrix

- Get the correlation between all the pairs of columns in the DataFrame.

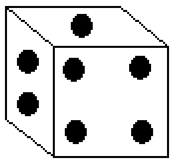
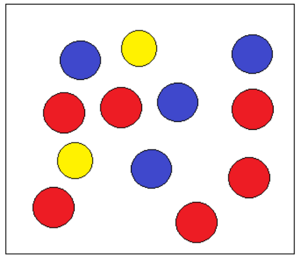
```
random_variable1 = df.Length
random_variable2 = df.Width
df.corr()
```

	Length	Width	Price
Length	1.000000	-0.006953	0.246945
Width	-0.006953	1.000000	0.551948
Price	0.246945	0.551948	1.000000

10. Sample_Space, Events, Trials, & Experiments

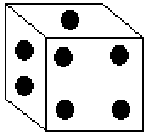
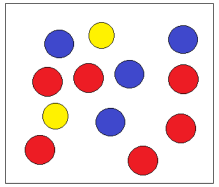
What is Sample Space?

- Is a set of all possible unique values of a Random Variable.
- We represent the sample space using S.
- Examples:

Example1	Example2
➢ suppose you are rolling a six-sided die. ➢ S = [1, 2, 3, 4, 5, 6]. 	➢ Suppose you have the following box of balls. ➢ S = [red, blue, yellow]. 

What are Events?

- An event is a subset of the sample space S.

Example1	**Example2**
- suppose you are rolling a six-sided die.  - S = [1, 2, 3, 4, 5, 6]. - The possible events are: E1={1}, E2={2}, E3={3}, E4={4}, E5={5}, E6={6}, E7={1, 2}, …,E11={1, 3, 5}, etc. - P(E7) means probability that die roll is 1 or 2. - P(E11) means probability that die roll is an odd number.	- Suppose you have the following box of balls.  - S = [red, blue, yellow]. - The possible events are: E1={red}, E2={red}, E3={red}, E4={red, blue}, E5={red, yellow}, E6={blue, yellow}, and E7={red, yellow, blue}. - P(E6) means probability that you draw a blue ball or a yellow ball.

What are Trials?

- A trial is the act or the process we are doing, for example:
  - Flipping a coin is a trial.
  - Rolling a dice is a trial.
- The result of a trial is an event.
- For example, Suppose that a dice is rolled, and 5 appears:
  - Sample-Space = {1, 2, 3, 4, 5, 6}.
  - Trial = rolling the dice.
  - Event = {5}.

What are Experiments?

- An experiment is a series of trials.

Example1	Example2
Flipping a coin twice is one experiment (two trials) 	Rolling three dice is one experiment (three trials). 

11. Independent & dependent Events

Independent Events:

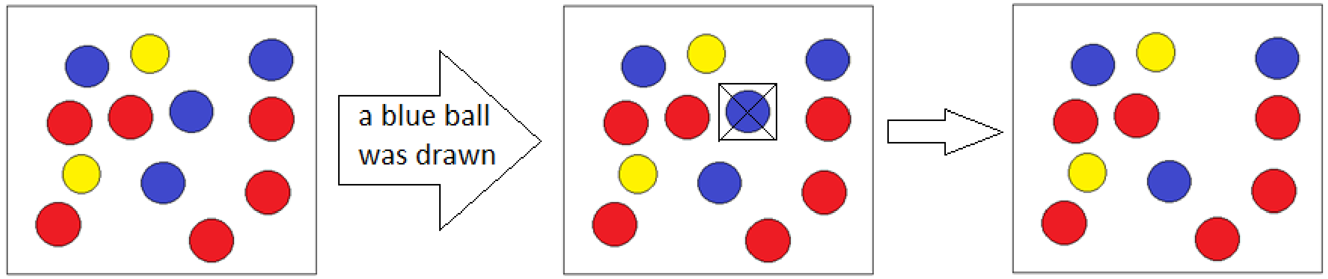
- Independent events occur when the outcome of one trial has no effect on the outcome of another.
- For example, if you flip a fair coin twice, then the chance of getting heads on the second toss (trial) is independent of the result of the first toss.

1st Toss	2nd Toss
H	H
H	T
T	H
T	T

Dependent Events:

- Dependent events occur when the outcome of a trial is affected by the outcome of previous trials.
- An example is drawing balls from a box with replacement.

1st trial 2nd trial



P(E={blue}) = 4 / 12 = .333 P(E={blue}) = 3 / 11 = .273

The outcome of the 1st trial = blue ball.

The event E={blue} in the 2nd trial was affected by the outcome of the 1st trial.

Thank You