Statistics for Data Science

by Pankaj - Royal iLearning

TOC

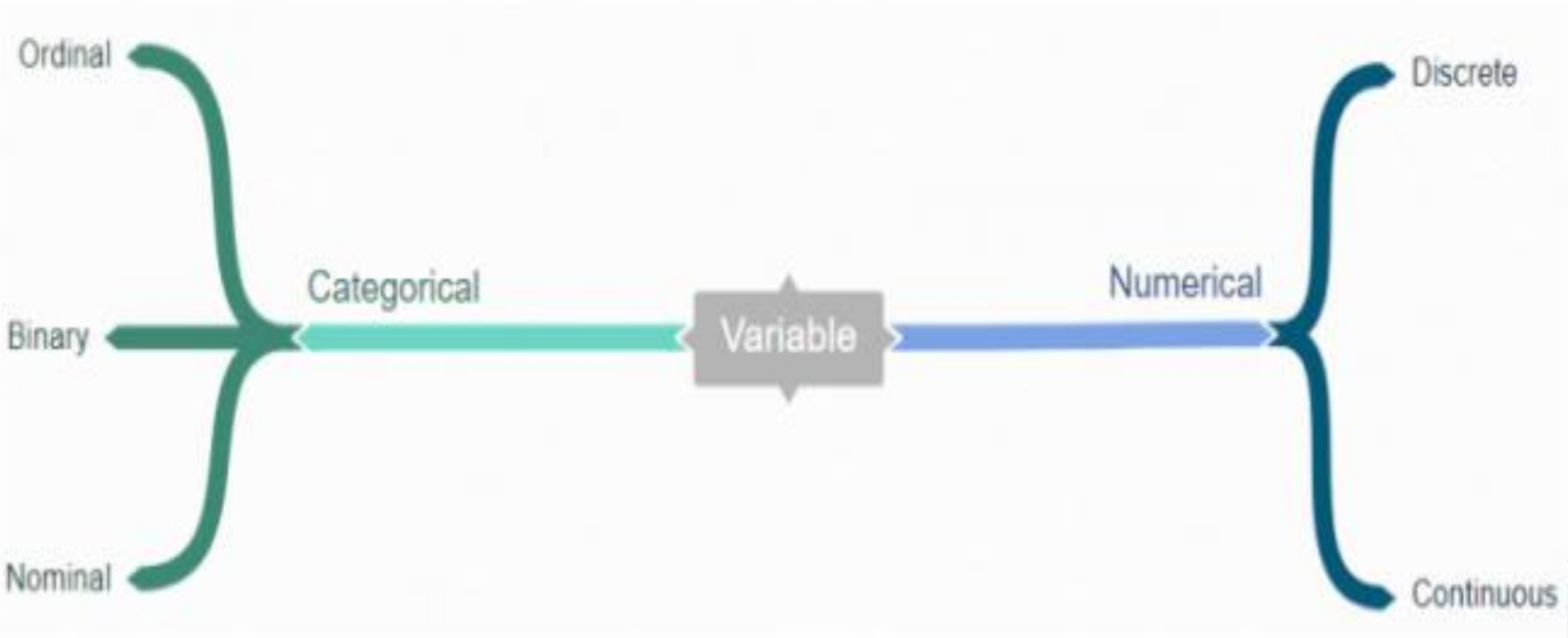
- Types of Data
- Mean, Median, Mode
- Using mean, median, and mode in Python/R
- Variation and Standard Deviation
- Probability Density Function; Probability Mass Function
- Common Data Distributions
- Percentiles and Moments
- Covariance and Correlation
- Conditional Probability
- Bayes’ Theorem

TOC

- Types of Data
- Mean, Median, Mode
- Using mean, median, and mode in Python/R
- Variation and Standard Deviation
- Probability Density Function; Probability Mass Function
- Common Data Distributions
- Percentiles and Moments
- Covariance and Correlation
- Conditional Probability
- Bayes’ Theorem

Types of Data

- Numerical
- Categorical



Numerical

Numerical data is information that is measurable, and it is, of course, data represented as numbers and not words or text.

- **Continuous** numbers are numbers that don’t have a logical end to them.

  Examples - variables that represent money or height.

- **Discrete** numbers are the opposite; they have a logical end to them.

  Example - variables for days in the month, or number of bugs logged.

Categorical

Categorical data, this is any data that isn’t a number, which can mean a string of text or date.

- **Ordinal** values are values that have a set order to them. Example – Ranking of race “First” or “Third”.
- **Nominal** values are the opposite of ordinal values, and they represent values with no set order to them. Example - variables such as “Country” or “Marital Status”.
- **Binary** data types only have two values – yes or no. Example - “True” and “False” or 1 and 0

TOC

- Types of Data
- Mean, Median, Mode
- Using mean, median, and mode in Python/R
- Variation and Standard Deviation
- Probability Density Function; Probability Mass Function
- Common Data Distributions
- Percentiles and Moments
- Covariance and Correlation
- Conditional Probability
- Bayes’ Theorem

Mean, Median, Mode

The **measures of central tendencies** such as mean, median and mode are measures that gives us some idea about the of centrality of data.

Mean or average is the value obtained by dividing the sum of all the data by the total number of data points.

Median the data situated at the middle position of the set. In a set with odd number of data points the median is the middlemost value while if the number of data points is even then it is the average of the two middle items.

Mode refers the data item that occurs most frequently in a given data set.

Mean, Median, Mode

Example: Data Set - 55, 56, 56, 58, 60, 61, 63, 64, 70, 78

Mean = (55 + 56 + 56 + 58 + 60 + 61 + 63 + 64 + 70 + 78) / 10 = 62.1

Median = (60 + 61)/2 = 60.5

Mode = 56

TOC

- Types of Data
- Mean, Median, Mode
- Using mean, median, and mode in Python/R
- Variation and Standard Deviation
- Probability Density Function; Probability Mass Function
- Common Data Distributions
- Percentiles and Moments
- Covariance and Correlation
- Conditional Probability
- Bayes' Theorem

TOC

- Types of Data
- Mean, Median, Mode
- Using mean, median, and mode in Python/R
- **Variation and Standard Deviation**
- Probability Density Function; Probability Mass Function
- Common Data Distributions
- Percentiles and Moments
- Covariance and Correlation
- Conditional Probability
- Bayes' Theorem

Variation and Standard Deviation

Variance signifies how much the data items are deviating from mean.

Mathematical Formulae

$$Variance(\sigma^2) = \sum \frac{(x-\bar{x})^2}{n-1}$$

:

- Larger variance means the data items deviate more from the mean.
- Smaller variance means the data items are closer to the mean.

Variation and Standard Deviation

Standard Deviation is simply the square root of the variance. In the above formula, σ is the standard deviation and $\sigma2$ is the variance.

TOC

- Types of Data
- Mean, Median, Mode
- Using mean, median, and mode in Python/R
- Variation and Standard Deviation
- **Probability Density Function; Probability Mass Function**
- Common Data Distributions
- Percentiles and Moments
- Covariance and Correlation
- Conditional Probability
- Bayes’ Theorem

Probability Density Function; Probability Mass Function

- What is Probability?

Probability is the likelihood of the occurrence of an event. Examples of events can be : Tossing a coin with the head up.

$$P(A) = \frac{Number\ of\ times\ A\ event\ has\ occurs}{Total\ number\ of\ event\ occurs}$$

A **Probability Density Function (PDF)** is a function that describes the relative likelihood for this random variable to take on a given value.

$$P(a \leq X \leq b) = \int baf(x)dx$$

[a,b]= Interval in which x lies.

$P(a \leq X \leq b)$ = probability that some value x lies within this interval.

dx = b-a

Probability Density Function; Probability Mass Function

- The **Probability Mass Function (PMF)** is a function which describes the probability associated with the random variable x. This function is named $P(x)$ or $P(x=x)$ to avoid confusion. $P(x=x)$ corresponds to the probability that the random variable x take the value x

$$P(a \leq X \leq b) = \int baf(x)dx$$

Let X be a discrete random variable with range R_X={x1,x2,x3,...} (finite or countably infinite). The function

$$P_X(x_k)=P(X=x_k), for\ k=1,2,3,...,$$

is called the probability mass function (PMF) of X.

TOC

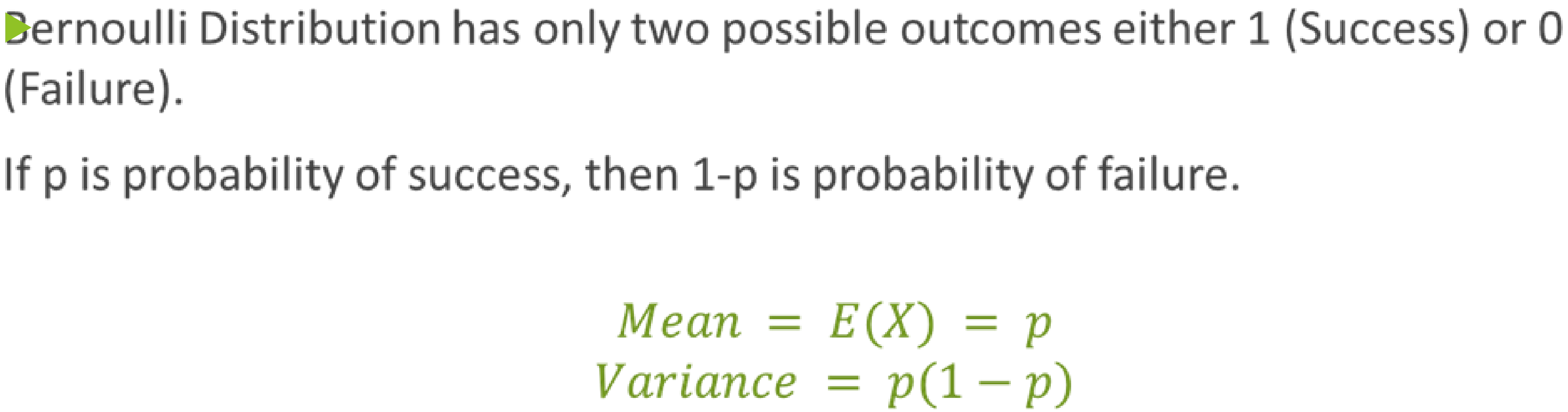
- Types of Data
- Mean, Median, Mode
- Using mean, median, and mode in Python/R
- Variation and Standard Deviation
- Probability Density Function; Probability Mass Function
- **Common Data Distributions**
- Percentiles and Moments
- Covariance and Correlation
- Conditional Probability
- Bayes’ Theorem

Common Data Distribution

There are 6 Common data distribution:

- Bernoulli Distribution
- Uniform Distribution
- Binomial Distribution
- Normal Distribution
- Poisson Distribution
- Exponential Distribution

1. Bernoulli Distribution



Bernoulli Distribution has only two possible outcomes either 1 (Success) or 0 (Failure).

If p is probability of success, then 1-p is probability of failure.

$$Mean = E(X) = p$$
$$Variance = p(1-p)$$

2. Uniform Distribution

Uniform Distribution is in which all outcomes are equally likely; each variable has the same probability that it will be the outcome

Example – Flipping of coin.

3. Binomial Distribution

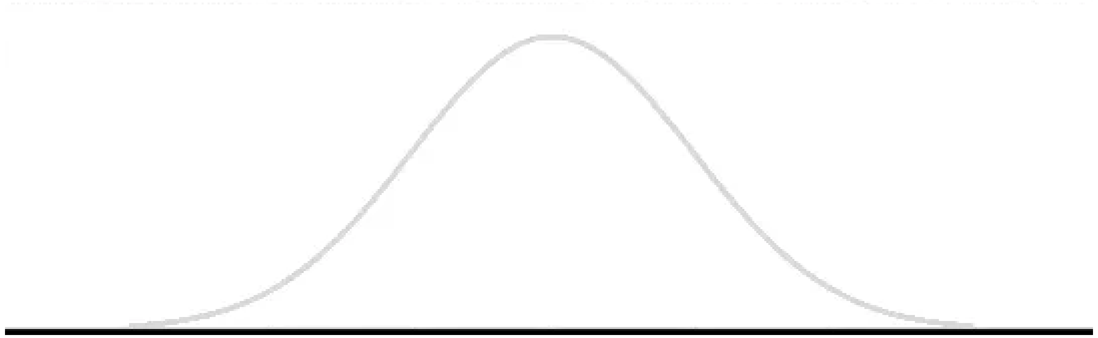
A Binomial distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials.

Difference between Bernoulli v/s Binomial

A Bernoulli random variable has two possible outcomes: 0 or 1. A binomial distribution is the sum of independent and identically distributed Bernoulli random variables.

4. Normal Distribution

A standard normal distribution is defined as the distribution with mean 0 and standard deviation 1.



5. Poisson Distribution

Poisson Distribution is applicable in situations where events occur at random points of time and space wherein our interest lies only in the number of occurrences of the event.

$$P(X = x) = e^{-\mu}\frac{\mu^{x}}{x!} \qquad for\ x\ =\ 0, 1, 2, \ldots\ldots$$

6. Exponential Distribution

Exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.

$f(x) = \{ \lambda e - \lambda x, \;\; x \geq 0$

TOC

- Types of Data
- Mean, Median, Mode
- Using mean, median, and mode in Python/R
- Variation and Standard Deviation
- Probability Density Function; Probability Mass Function
- Common Data Distributions
- **Percentiles and Moments**
- Covariance and Correlation
- Conditional Probability
- Bayes’ Theorem

Percentiles and Moments

When a value is given x percentile, this means that x percentage of values in the distribution is below that value.

Moments try to measure the shape of the probability distribution function. The zeroth moment is the total probability of the distribution which is 1.

- The first moment is the mean.
- The second moment is the variance.
- The third moment is the skew which measures how lopsided the distribution is.
- The fourth moment is kurtosis which is the measure of how sharp is the peak of the graph.

TOC

- Types of Data
- Mean, Median, Mode
- Using mean, median, and mode in Python/R
- Variation and Standard Deviation
- Probability Density Function; Probability Mass Function
- Common Data Distributions
- Percentiles and Moments
- **Covariance and Correlation**
- Conditional Probability
- Bayes’ Theorem

Covariance and Correlation

Covariance measures how two variables vary in tandem to their means. The formula to calculate covariance is shown below.

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y))$$

where x and y are the individual values of X and Y ranging from i = 1,2, .., n where the probability that each value may occur is equal and is equal to (1/n). E(x) and E(y) are the means of X and Y.

Covariance and Correlation

Correlation also measures how two variables move with respect to each other.

A perfect positive correlation means that the correlation coefficient is 1.

A perfect negative correlation means that the correlation coefficient is -1.

A correlation coefficient of 0 means that the two variables are independent of each other.

The formula for finding the correlation coefficient can be found using the following formula.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}}$$

TOC

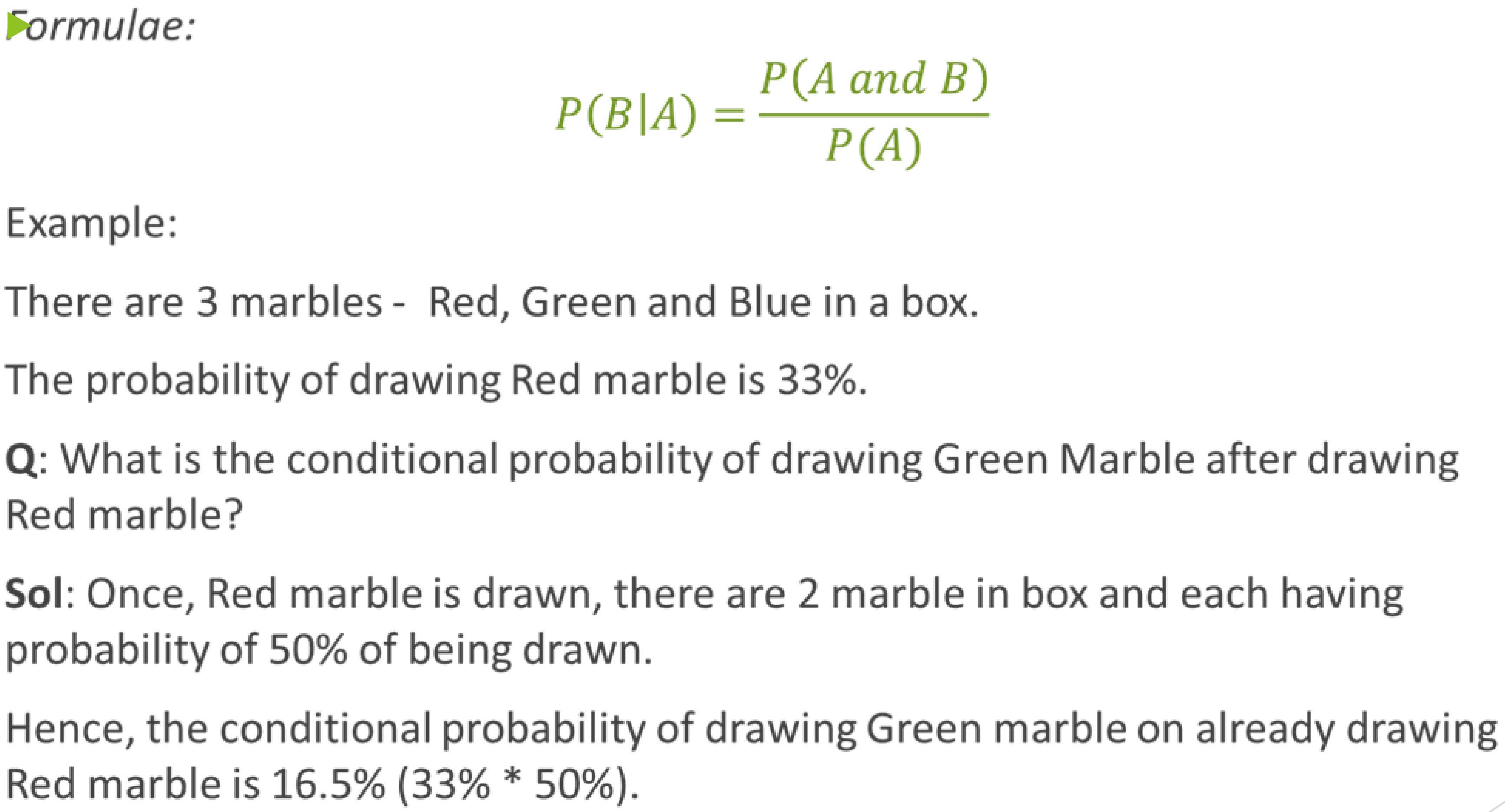
- Types of Data
- Mean, Median, Mode
- Using mean, median, and mode in Python/R
- Variation and Standard Deviation
- Probability Density Function; Probability Mass Function
- Common Data Distributions
- Percentiles and Moments
- Covariance and Correlation
- Conditional Probability
- Bayes’ Theorem

Conditional Probability

- Conditional probability is the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event.

$$P(B|A) = \frac{P(A\ and\ B)}{P(A)}$$

Conditional Probability



Formulae:

$$P(B|A) = \frac{P(A\ and\ B)}{P(A)}$$

Example:

There are 3 marbles - Red, Green and Blue in a box.

The probability of drawing Red marble is 33%.

Q: What is the conditional probability of drawing Green Marble after drawing Red marble?

Sol: Once, Red marble is drawn, there are 2 marble in box and each having probability of 50% of being drawn.

Hence, the conditional probability of drawing Green marble on already drawing Red marble is 16.5% (33% * 50%).

TOC

- Types of Data
- Mean, Median, Mode
- Using mean, median, and mode in Python/R
- Variation and Standard Deviation
- Probability Density Function; Probability Mass Function
- Common Data Distributions
- Percentiles and Moments
- Covariance and Correlation
- Conditional Probability
- **Bayes’ Theorem**

Bayes’ Theorem

- Bayes' theorem is a formula that describes how to update the probabilities of hypotheses when given evidence.

It follows simply from the axioms of conditional probability, but can be used to powerfully reason about a wide range of problems involving belief updates.

Given a hypothesis *H* and evidence *E*, Bayes' theorem states that the relationship between the probability of the hypothesis before getting the evidence *P(H)* and the probability of the hypothesis after getting the evidence *P(H/E)* is

$$P(H|E) = \frac{P(E|H)}{P(E)} P(H)$$