# 3 PROBABILITY DISTRIBUTION AND APPLICATIONS IN R

***Basic Concepts of Probability***

A probability is a number that reflects the chance or likelihood that a particular event will occur. Probabilities can be expressed as proportions that range from 0 to 1, and they can also be expressed as percentages ranging from 0% to 100%. A probability of 0 indicates that there is no chance that a particular event will occur, whereas a probability of 1 indicates that an event is certain to occur. A probability of 0.45 (45%) indicates that there are 45 chances out of 100 of the event occurring.

The concept of probability can be illustrated in the context of a study of obesity in children 5-10 years of age who are seeking medical care at a particular pediatric practice. The population (sampling frame) includes all children who were seen in the practice in the past 12 months and is summarized below.

| | Age (years) | | | | | | |
|---|---|---|---|---|---|---|---|
| | **5** | **6** | **7** | **8** | **9** | **10** | **Total** |
| **Boys** | 432 | 379 | 501 | 410 | 420 | 418 | 2,560 |
| **Girls** | 408 | 513 | 412 | 436 | 461 | 500 | 2,730 |
| **Totals** | 840 | 892 | 913 | 846 | 881 | 918 | 5,290 |

**Unconditional Probability**

If we select a child at random (by simple random sampling), then each child has the same probability (equal chance) of being selected, and the probability is 1/N, where N=the population size. Thus, the probability that any child is selected is 1/5,290 = 0.0002. In most sampling situations we are generally not concerned with sampling a specific individual but instead, we concern ourselves with the probability of sampling certain types of individuals. For example, what is the probability of selecting a boy or a child 7 years of age? The following formula can be used to compute probabilities of selecting individuals with specific attributes or characteristics.

**P(characteristic) = # persons with characteristic / N**

Try to figure these out before looking at the answers:

1. What is the probability of selecting a boy?
    a. If we select a child at random, the probability that we select a boy is computed as follows P(boy) = 2,560/5,290 = 0.484 or 48.4%.
2. What is the probability of selecting a 7-year-old?
    a. The probability of selecting a child who is 7 years of age is P(7 years of age) = 913/5,290 = 0.173.
3. What is the probability of selecting a boy who is 10 years of age?
    a. P(boy who is 10 years of age) = 418/5,290 = 0.079.

4. What is the probability of selecting a child (boy or girl) who is at least 8 years of age?
    a. P(at least 8 years of age) = (846 + 881+ 918)/5,290 = 2,645/5,290 = 0.500.

**Conditional Probability**

Each of the probabilities computed in the previous section (e.g., P(boy), P(7 years of age)) is an unconditional probability because the denominator for each is the total population size (N=5,290) reflecting the fact that everyone in the entire population is eligible to be selected. However, sometimes it is of interest to focus on a particular subset of the population (e.g., a sub-population). For example, suppose we are interested just in the girls and ask the question, what is the probability of selecting a 9-year-old from the sub-population of girls? There is a total of NG=2,730 girls (here NG refers to the population of girls), and the probability of selecting a 9-year-old from the sub-population of girls is written as follows:

**P(9 year old | girls) = # persons with characteristic / N**

where | girls indicate that we are conditioning the question to a specific subgroup, i.e., the subgroup specified to the right of the vertical line.

The conditional probability is computed using the same approach we used to compute unconditional probabilities. In this case:

**P(9 year old | girls) = 461/2,730 = 0.169.**

This also means that 16.9% of the girls are 9 years of age. Note that this is not the same as the probability of selecting a 9-year old girl from the overall population, which is P(girls who are 9 years of age) = 461/5,290 = 0.087.

What is the probability of selecting a boy from among the 6-year-olds?
P(boy | 6 years of age) = 379/892 = 0.425. Thus 42.5% of the 6-year-olds are boys (57.5% of the 6 year olds are girls).

***What is a Discrete Probability Distribution?***

In statistics, you'll come across dozens of different types of probability distributions, like the binomial distribution, normal distribution, and Poisson distribution. All of these distributions can be classified as either a continuous or a discrete probability distribution.
A discrete probability distribution is made up of discrete variables. Specifically, if a random variable is discrete, then it will have a discrete probability distribution.

**Discrete Probability Distribution Examples**

For example, let's say you had the choice of playing two games of chance at a fair.
Game 1: Roll a die. If you roll a six, you win a prize.
Game 2: Guess the weight of the man. If you guess within 10 pounds, you win a prize.
One of these games is a discrete probability distribution and one is a continuous probability distribution. Which is which?

For game 1, you could roll a 1,2,3,4,5 or 6. All of the die rolls have an equal chance of being rolled (one out of six, or 1/6). This gives you a discrete probability distribution of:

| Roll | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Odds | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

For the guess the weight game, you could guess that the mean weighs 150 lbs. Or 210 pounds. Or 185.5 pounds. Or any fraction of a pound (172.566 pounds). Even if you stick to, say, between 150 and 200 pounds, the possibilities are endless:
160.1 lbs.
160.11 lbs.
160.111 lbs.
160.1111 lbs.
160.111111 lbs.

In reality, you probably wouldn't guess 160.111111 lbs…that seems a little ridiculous. But it doesn't change the fact that you could (if you wanted to), so that's why it's a continuous probability distribution.

***What is a Factorial?***

Factorials (!) are products of every whole number from 1 to n. In other words, take the number and multiply through by 1.

For example:

If n is 3, then 3! is 3 x 2 x 1 = 6.
If n is 5, then 5! is 5 x 4 x 3 x 2 x 1 = 120.
It's a shorthand way of writing numbers. For example, instead of writing 479001600, you could write 12! instead (which is 12 x 11 x 10 x 9 x 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1).

**What is a factorial used for in stats?**

In algebra, you probably encountered ugly-looking factorials like (x – 10!)/(x + 9!). Don't worry; You won't be seeing any of these in your beginning stats class. Phew! The only time you'll see them is for permutation and combination problems.

The equations look like this:

$$\frac{100}{5!(100-5)!}.$$

Permutations and Combinations are the various ways in which objects from a set may be selected, generally without replacement, to form subsets. This selection of subsets is called a permutation when the order of selection is a factor, a combination when order is not a factor. By considering the ratio of the number of desired subsets to the number of all possible subsets for many games of chance in the 17th century, the French mathematicians Blaise Pascal and Pierre de Fermat gave impetus to the development of combinatorics and probability theory.

The concepts of and differences between permutations and combinations can be illustrated by an examination of all the different ways in which a pair of objects can be selected from five distinguishable objects—such as the letters A, B, C, D, and E. If both the letters selected and the order of selection are considered, then the following 20 outcomes are possible:

```
AB   BA   AC   CA   AD
DA   AE   EA   BC   CB
BD   DB   BE   EB   CD
DC   CE   EC   DE   ED
```

***R Applications – Real-world Use Cases of R programming***

**What is R used for R ? R Applications across these sectors**

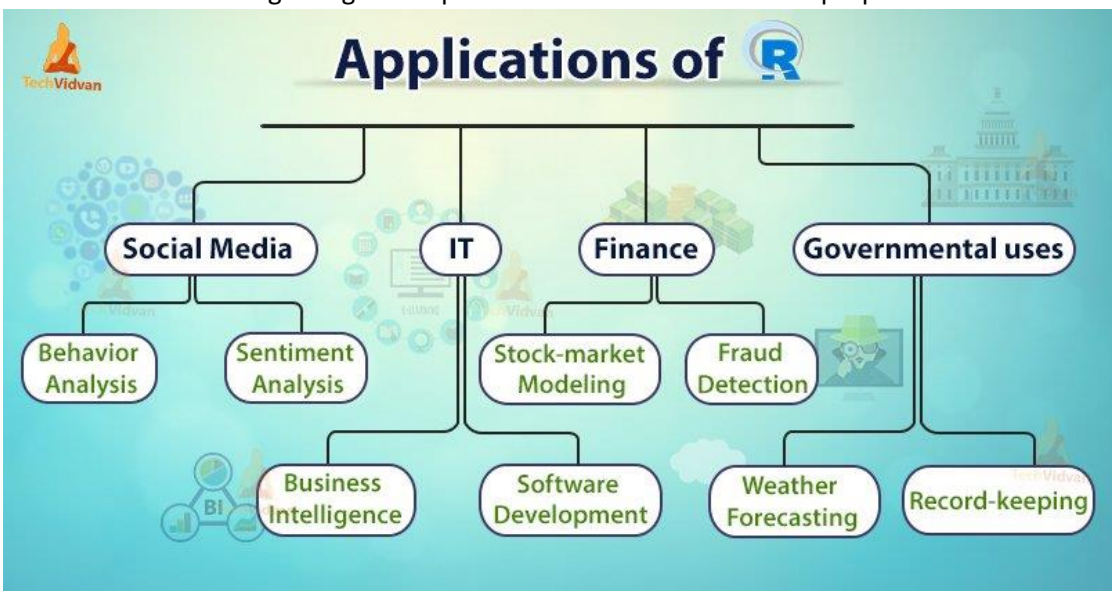
Data Science and Big Data have proved themselves useful and even necessary in many different fields and industries today. It helps them to keep up with the trends and capitalize on every opportunity.

R acts as a tool, used to make sense of big data and to gain use from it. R has also proved itself useful in research by processing large amounts of data in less time.

Let's take a look at an interesting fact. "According to O'Reilly, R is the most-used data science language after SQL."

In this article, we will explore the various R applications in the real world.

Applications of R Programming
Let's start from the beginning and explore the uses of R for research purposes:




**R in Research and Academics**
R is a statistical research tool. It is still used by statisticians and students to perform various statistical computations and analyses. Statistical techniques like linear and non-linear modeling, time-series analysis, classification, classical statistic tests, clustering, and others are all implemented by R and its libraries.

R is also used for machine learning research and deep learning as well. With libraries that facilitate monitored and unmonitored learning, R is one of the most commonly used languages for machine learning.

Other research involving large data sets like big data, finding genetic anomalies and patterns, various drug compositions, all use R to sift through a large collection of relevant data and to draw meaningful conclusions from it.

**R Use Cases in Research & Academics**

- Cornell University: Cornell recommends their researchers and students use R for all their research involving statistical computing.
- UCLA: The University of California, Los Angeles uses R to teach statistics and data analysis to its students.

Apart from research, R also has its applications in IT companies.

**R in IT Sector**

IT companies not only use R for their business intelligence but offer such services to other small, medium, and large scale businesses as well. They use it for their machine learning products too.

They use R to build statistical computing tools and data handling products and to create other data manipulation services.

Some big IT companies that use R:

- Accenture
- IBM
- Infosys
- Paytm
- Tata Consultancy Services
- Wipro

**R Use Cases in IT Sector**

- Mozilla: Mozilla uses R to visualize web activity for their browser firefox.
- Microsoft: Microsoft uses R as a statistical engine within the Azure Machine Learning framework. They also use it for the Xbox matchmaking service.
- Foursquare: R works behind the scenes on Foursquare's recommendation engine.
- Google: Google uses R to improve its search results, to provide better search suggestions, to calculate the ROI of its advertising campaigns, to increase the efficiency of online advertising, and to predict their economic activity.

**R in Finance**

Other than the finance sector which industry will be dealing more with statistics as R is a statistical programming language.

R and data science find widespread use in the finance sector. R provides an advanced statistical suite for all the financial tasks and computations. Moving averages, auto-regression, time-series analysis, stock-market modeling, financial data mining, downside risk assessment are all easily done through R and its libraries.

R is also used to support the business decision-making process. R's data visualization powers can represent the findings of data analysis in multiple graphical formats like candlestick charts, density plots, and drawdown plots of high quality.

This helps the business minds to connect with the technical aspect of data analyses and their results. Companies like American Express, Bajaj Allianz Insurance, JP Morgan, and Standard Chartered use R.

**R Use Cases in Finance**

- Lloyds of London: Lloyds of London use R for risk analysis.
- Bajaj Allianz Insurance: Bajaj Allianz uses R to make their upsell propensity models and recommendation engines. They also use it to mine data and generate actionable insights to improve customer experience.

The digital revolution has changed the world drastically. One of the most prominent changes is the fact that marketplaces have moved to the internet. The E-commerce industry makes heavy usage of R for varying purposes.

**R in E-commerce**

In the field of finance and retail, analytics is useful for risk assessment and for devising a marketing strategy. E-commerce goes beyond that in its usage of data science. E-commerce companies use R to improve the user's experience on their site as well as for marketing and finance purposes. They use R to improve cross-product selling. When a customer is buying a product the site suggests additional products that complement their original purchase. These suggestions also work for products purchased by the customer in the past. Internet-based companies like various e-commerce sites gather and process structured and unstructured data from varying sources. R proves to be highly useful for this.

Apart from this, R is also used to help with marketing strategy, targeted advertising, sales modeling, and financial data processing.

**R Use Cases in E-commerce**

Amazon: Amazon uses R and data analysis to improve its cross-product suggestions.
Flipkart: Flipkart uses R for predictive analysis which helps them with targetted advertisements.
Social media is the most common generator of big data today. Therefore, the most advanced and cutting-edge uses of data science can be found in the social media industry.

If you have any queries in R applications till now, mention them in the comment section.

**R in Social Media**

Social media companies like Facebook use R for behavior analysis and sentiment analysis. They can alter and improve their suggestions to users based on the user's history, and the mood and tone of their recent posts and viewed content. The advertisements shown to the user are also adjusted according to user sentiment and history. R is also used to analyze traffic, user sessions, and content, all to improve user experience. R Use Cases in Social Media
Facebook: Facebook uses R to predict colleague interactions and update its social network graph.
Twitter: Twitter uses R for semantic clustering. They also use it for data visualization.
Another sector making great use of R's statistical computation abilities is the banking sector.

**R in Banking**

Banking firms use R for credit risk modeling and other forms of risk analytics. Banks often use R along with other proprietary software like SAS. It is also used for fraud detection, mortgage haircut modeling, stat modeling, volatility modeling, loan stress test simulation, client assessment, and much more. Apart from statistics, banks also use R for business intelligence and data visualization.

Another use of R is in the calculation of customer segmentation, customer quality, and customer retention.

**R Use Cases in Banking**

- ANZ: ANZ bank uses R for credit risk modeling and also in models for mortgage loss.
- Bank of America: Bank of America uses R for financial reporting and to calculate financial losses.

The healthcare industry is not one to be left behind when it comes to cutting-edge technologies:

**R in Healthcare**

With R, you can crunch data and process information, providing an essential backdrop for further analysis and data processing. Genetics, drug discovery, bioinformatics, epidemiology, etc. are some fields in the healthcare industry that

use R heavily. It is used to analyze and predict the spreading of various diseases, for analyzing genetic sequences, to analyze drug-safety data, and to analyze various permutations and combinations of drugs and chemicals as well. R's Bioconductor package provides facilities for analyzing genomic data. Lastly, R is a god-send for pre-clinical trials of all new drugs and medical techniques.

**R Use Case in Healthcare**
Merck: Merck & co. use the R programming language for clinical trials and drug testing.

Manufacturing companies also use R to make use of big data and to be ahead of the curve.

**R in Manufacturing**
Various manufacturing companies use R to complement their marketing and business strategies. They analyze customer feedback to help streamline and improve their products. They also use the data to support their marketing strategies. Predicting demand and market trends to adjust their manufacturing practices is yet another use of R and data analytics.

**R Use Cases in Manufacturing**
Ford Motor Company: Ford uses R for statistical analyses to support its business strategy and to analyze customer sentiment about its product which helps them in improving their future designs.
John Deere: John Deere uses R to forecast demand for their products and spare parts. They also use it to forecast crop yield and use that data for their business strategy and to meet market demand and downturns.
Every government has to handle a large amount of data. A country's worth of data! Many governmental departments across the world use R as well.

**R in Governmental Use**
Many governmental departments use R for record-keeping and processing their censuses. This helps them in effective law-making and governance. They also use it for essential services like drug regulation, weather forecasting, disaster-impact analysis, and much more.

**R Use Cases in Governmental Activities**
Food and Drug Administration: FDA uses R for drug evaluation and to perform pre-clinical trials. It also uses R to predict possible reactions and medical issues caused by various food products.
National Weather Service: The National Weather Service uses R for weather forecasts and disaster prediction. They also use it to visualize their forecasts and predictions to analyze the areas affected.