1st Unit Data science

Data Science is a combination of multiple disciplines that uses statistics, data analysis, and machine learning to analyze data and to extract knowledge and insights from it.

I.What is Data Science?

Data Science is about data gathering, analysis and decision-making.

Data Science is about finding patterns in data, through analysis, and make future predictions.

By using Data Science, companies are able to make:

- Better decisions (should we choose A or B)
- Predictive analysis (what will happen next?)
- Pattern discoveries (find pattern, or maybe hidden information in the data)

Where is Data Science Needed?

Data Science is used in many industries in the world today, e.g. banking, consultancy, healthcare, and manufacturing.

Examples of where Data Science is needed:

- For route planning: To discover the best routes to ship
- To foresee delays for flight/ship/train etc. (through predictive analysis)
- To create promotional offers
- To find the best suited time to deliver goods
- To forecast the next years revenue for a company
- To analyze health benefit of training
- To predict who will win elections

Data Science can be applied in nearly every part of a business where data is available. Examples are:

- Consumer goods
- Stock markets
- Industry
- Politics
- Logistic companies
- E-commerce

2. Big Data and Data Science

Big Data:

It is huge, large, or voluminous data, information, or the relevant statistics acquired by large organizations and ventures. Many software and data storages is created and prepared as it is difficult to compute the big data manually. It is used to discover patterns and trends and make decisions related to human behavior and interaction technology.

Advantages of Big Data:

- Able to handle and process large and complex data sets that cannot be easily managed with traditional database systems
- Provides a platform for advanced analytics and machine learning applications
- Enables organizations to gain insights and make data-driven decisions based on large amounts of data
- Offers potential for significant cost savings through efficient data management and analysis

Disadvantages of Big Data:

- Requires specialized skills and expertise in data engineering, data management, and big data tools and technologies
- Can be expensive to implement and maintain due to the need for specialized infrastructure and software
- May face privacy and security concerns when handling sensitive data
- Can be challenging to integrate with existing systems and processes

Data Science:

Data Science is a field or domain which includes and involves working with a huge amount of data and using it for building predictive, prescriptive, and prescriptive analytical models. It's about digging, capturing, (building the model) analyzing(validating the model), and utilizing the data(deploying the best model). It is an intersection of Data and computing. It is a blend of the field of Computer Science, Business, and Statistics together.

Advantages of Data Science:

- Provides a framework for extracting insights and knowledge from data through statistical analysis, machine learning, and
- data visualization techniques
- Offers a wide range of applications in various fields such as finance, healthcare, and marketing
- Helps organizations make informed decisions by extracting meaningful insights from data
- Offers potential for significant cost savings through efficient data management and analysis

Disadvantages of Data Science:

- Requires specialized skills and expertise in statistical analysis, machine learning, and data visualization
- Can be time-consuming and resource-intensive due to the need for data cleaning and preprocessing
- May face ethical concerns when dealing with sensitive data
- Can be challenging to integrate with existing systems and processes

Similarities between Big Data and Data Science:

- Both fields deal with large amounts of data and require specialized skills and expertise
- Both aim to extract insights and knowledge from data to inform decision-making
- Both have a wide range of applications in various industries
- Both can lead to significant cost savings and operational efficiencies when applied correctly

Below is a table of differences between Big Data and Data Science:

Data Science	**Big Data**
Data Science is an area.	Big Data is a technique to collect, maintain and process huge information.
It is about the collection, processing, analyzing, and utilizing of data in various operations. It is more conceptual.	It is about extracting vital and valuable information from a huge amount of data.
It is a field of study just like Computer Science, Applied Statistics, or Applied Mathematics.	It is a technique for tracking and discovering trends in complex data sets.
The goal is to build data-dominant products for a venture.	The goal is to make data more vital and usable i.e. by extracting only important information from the huge data within existing traditional aspects.

Data Science	**Big Data**
Tools mainly used in Data Science include SAS, R, Python, etc	Tools mostly used in Big Data include Hadoop, Spark, Flink, etc.
It is a superset of Big Data as data science consists of Data scrapping, cleaning, visualization, statistics, and many more techniques.	It is a sub-set of Data Science as mining activities which is in a pipeline of Data science.
It is mainly used for scientific purposes.	It is mainly used for business purposes and customer satisfaction.
It broadly focuses on the science of the data.	It is more involved with the processes of handling voluminous data.

3. Statistical inference, Population and samples

Samples and Populations

In statistics, we generally want to study a population. You can think of a population as an entire collection of persons, things, or objects under study. To study the larger population, we select a sample. The idea of sampling is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000 to 2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if the manufactured 16 ounce containers does indeed contain 16 ounces of the drink.

From the sample data, we can calculate a **statistic**. A statistic is a number that is a property of the sample. Common sample statistics include sample means, sample proportions, and sample variances. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic.

The statistic can also be used as an estimate of a **population parameter**. A parameter is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

3.1 Confidence Intervals

As described in the previous section, we often want to use sample data to estimate population quantities. Due to the randomness inherent to sampling, an observed sample statistic is almost certainly not equal to the true population parameter. To quantify the variability surrounding the sample statistic, we can compute a **confidence interval** which provides a lower and upper bound for where we think the true population value lies. Note that unless we take a sample which consists of the entire population (often called a **census**), we will never know the true population parameter with absolute certainty.

3.2 Confidence Intervals for Proportions

Suppose that in a survey of one hundred adult cell phone users, 30% switched carriers in the past two years. Based on this sample statistic, what can we conclude about the entire population of adult cell phone users? Our sample estimate of 30% is based on a random sample of one hundred users and not the entire population, so we cannot conclude that the true population parameter is 30%. Instead, we must calculate a confidence interval to understand the range of plausible values for the population proportion.

In R, we can calculate a confidence interval for proportions with the binom.test() command, which uses the following syntax:

```
binom.test(x, n, conf.level = 0.95)
```

- *Required arguments*
  - x: The number of “successes” in the sample.
  - n: The sample size.
- *Optional arguments*
  - conf.level: The confidence level of the interval.

Applying this to our sample of cell phone users, we can run binom.test() with x equal to thirty and n equal to one hundred:

```
binom.test(30, 100)
##
```

```
##  Exact binomial test
##
## data:  30 and 100
## number of successes = 30, number of trials = 100, p-value = 0.0000785
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2124064 0.3998147
## sample estimates:
## probability of success
##                    0.3
```

For now, all we care about in this output is the line that reads 95 percent confidence interval:. We interpret this output as follows: we are 95% confident that the true population proportion of adult cell phone users who switched carriers in the past two years is between 21.24% and 39.98%.

3.3 Hypothesis Testing

An important component of inference is **hypothesis testing**, which allows us to analyze the evidence provided by the sample to assess some claim about the population.

- A **one-sample hypothesis test** compares a single population parameter to a specified value. For example, you may wonder whether your local barista pours a full twelve ounces of coffee in each cup. By drawing a sample of the cups poured by your barista, you could use a one-sample hypothesis test to answer this question.
- A **two-sample hypothesis test** assesses the equality of parameters from two different populations. For example, you may wonder whether the barista near your home and the barista near your work pour similar amounts on average, or if one pours more than the other. By drawing a sample of the cups poured by each barista, you could use a two-sample hypothesis test to answer this question.

3.3.1 Two-Sample Hypothesis Testing

Although analyzing one sample of data is useful for problems like gauging public opinion or testing the stability of a manufacturing process, there are more advanced analyses which involve comparing the responses of two or more groups. This can be in the form of comparing means or comparing proportions.

Note that **A/B testing**, which the reader may be familiar with, is a common application of two-sample hypothesis testing in business settings. Because A/B testing is primarily used to establish causal relationships between variables,

Testing Means

Many business applications involve a comparison of two population means. For instance, a company may want to know if a new logo produces more sales than the previous logo, or a consumer group may want to test whether two major brands of food freezers differ in the average amount of electricity they use. In this section we extend our knowledge of hypothesis testing on one population mean to comparing two population means.

Independent Samples

The independent samples t-test is used to compare the means of two independent samples. It can be used to test whether:

- Biology graduates have a different average annual income than chemistry graduates.
- Length of life, on average, is shorter for never-married persons than for people who are or have been married.
- The mean years of schooling of Republicans is different than the mean years of schooling of Democrats.
- Men average more hours of sleep per night than women.
- The PE (price to earnings) ratio for tech stocks is on average higher than for financial services stocks.

When performing two sample tests of means, the null hypothesis is always that the population means of the two groups are the same.

3.4 Hypothesis Testing with More Than Two Samples

3.4.1 Testing Means (ANOVA)

If we want to compare the means of more than two groups, one procedure available is called **Analysis of Variance (ANOVA)**. The name seems strange because we are comparing means, but the word variance comes from the fact that this procedure makes a relatively strong assumption that the variability in each group we are comparing is the same. A rule of thumb when using ANOVA is that the ratio of the largest standard deviation of the groups to the smallest standard deviation should be no more than three.

The null hypothesis for ANOVA is that the means of all our groups are the same. The alternative is that there are at least two groups that have different means. If we reject the null hypothesis we need to do further analyses to see where the differences exist.

3.4.2 Testing Proportions (Chi-Square)

Research in business often generates frequency (count) data. This is certainly the case in most opinion surveys in which the person interviewed is asked to respond to a question by marking, say "Agree", "Not Sure", or "Disagree", or some other such collection of categories. In a case like this, the investigator might be concerned with determining what proportion of respondents marked each of the choices or whether there is any relationship between the opinion marked and the sex, age, or occupation of the respondent.

Chi-square methods make possible the meaningful analysis of frequency data by permitting the comparison of frequencies actually observed with frequencies which would be expected if the null hypothesis were true. At first glance the chi-square test procedures can be confusing as there are two different tests with very similar names.

- **Chi-Square Goodness of Fit Test:** this is used to test if counts in different categories follow a specified distribution.
- **Chi-Square Test of Independence:** this is used to test if two categorical variables are independent or dependent.

3.4.3 Goodness-of-Fit

Suppose that the Bar Galaxy Chocolate Co. wants to determine if customers have a preference for any of the following four candy bars. From a random sample of 200 people, it was found that:

1. 43 preferred The Frosty Bar
2. 53 preferred Galaxy's Milk Chocolate
3. 60 preferred Galaxy's Special Dark Chocolate
4. 44 preferred Munchies Bar

For the goodness-of-fit test, the null hypothesis states that customers have no preference for any of the four candy bars (1, 2, 3, and 4). That is, all four candy bars are equally preferred. The alternative hypothesis states that the preference probabilities are not all the same

4. Statistical Modeling

Statistical modeling is like a formal depiction of a theory. It is typically described as the mathematical relationship between random and non-random variables.

The science of statistics is the study of how to learn from data. It helps you collect the right data, perform the correct analysis, and effectively present the results with statistical knowledge. Statistical modeling is key to making scientific discoveries, data-driven decisions, and predictions.

Statistical modeling helps you differentiate between reasonable and dubious conclusions based on quantitative evidence. Analyses and predictions made by statisticians are highly trustworthy. A statistician can help investigators avoid various analytical traps along the way.

What is statistical modeling?

The statistical modeling process is a way of applying statistical analysis to datasets in data science. The statistical model involves a mathematical relationship between random and non-random variables.

A statistical model can provide intuitive visualizations that aid data scientists in identifying relationships between variables and making predictions by applying statistical models to raw data.

Examples of common data sets for statistical analysis include census data, public health data, and social media data.

Statistical modeling techniques

Data gathering is the foundation of statistical modeling. The data may come from the cloud, spreadsheets, databases, or other sources. There are two categories of statistical modeling methods used in data analysis. These are:

- **Regression model:** A predictive model designed to analyze the relationship between independent and dependent variables. The most common regression models are logistical, polynomial, and linear. These models determine the relationship between variables, forecasting, and modeling.
- **Classification model:** An algorithm analyzes and classifies a large and complex set of data points. Common models include decision trees, Naive Bayes, the nearest neighbor, random forests, and neural networking models.
- **K-means clustering:** The algorithm combines a specified number of data points into specific groupings based on similarities.
- **Reinforcement learning:** This technique involves training the algorithm to iterate over many attempts using deep learning, rewarding moves that result in favorable outcomes, and penalizing activities that produce undesired effects.

5. Introduction to Probability Distributions

This article will focus on critical Probability Distributions from the data science point of view. We will divide these Probability distributions based on whether the data is discrete or continuous.

No matter what field you are in, Statistics and Probability will be there. Economics, Finance, Trading, Social sciences, Natural sciences and, of course, Data science is indispensable. A significant chunk of Data science is about understanding the behaviours and properties of variables, and this is not possible without knowing what distributions they belong to. Simply put, the probability distribution is a way to represent possible values a variable may take and their respective probability.

Discrete

- Bernoulli Distribution
- Binomial Distribution
- Poisson Distribution

Continuous

- Normal Distribution
- chi^2 Distribution
- Student-t Distribution
- Log-Normal Distribution
- Exponential Distribution

Probability Density Function (PDF)

The Probability Density Function is PMF equivalent but for continuous random variables. A continuous distribution is characterised by infinite numbers of the random variables, which means the probability of any random sample at a given point is infinitesimally low. So, a range of values is used to infer the likelihood of a random sample. And doing an integration over the range will fetch our likelihood for the same sample. Mathematically,

$$Pr[a \leq X \leq b] = \int_a^b f_X(x)dx$$

Cumulative Density Function (CDF)

A cumulative density function at x explains the probability of a random variable X taking on values less than or equal to x. It applies to distribution regardless of its type, continuous or discrete. For a constant distribution, all we need to do is integrate the density function from negative infinity to x.

$$f_X(x)=\int_{-\infty}^{x} f_X(t)dt$$

And CDF for a random variable X sampled from a discrete distribution is given by

Finding the CDF of x is equal to the probability of random variable X <=x, which is, in turn, the summation of the possibility of X similar to x_i when x_i is less than or equal to x.

Discrete Distributions

As its name suggests, a Discrete Distribution is a distribution where observation can take only a finite number of values. For example, the rolling of a die can only have resulted from 1 to 6, or the gender of a species. It is fairly common to have discrete variables in a real-world data set, be it gender, age or visitors to a place at a particular time. There are a lot of other discrete distributions, but we will focus on the most common and important of them.

Bernoulli Distribution

Bernoulli Distribution can be safely assumed to be the simplest 0f discrete distributions. Consider an example of flipping an unbiased coin. You either get a Head or a Tail. If we consider either of them as our priority(caring only about Head/Tail), the outcome will only be 0 (failure) or 1 (success). As it is an unbiased coin probability assigned to each outcome is 0.5. Remember, the outcome is always binary True/False, Head/Tail, Success/Failure etc.

The probability mass function or PMF of Bernoulli Distribution is given as

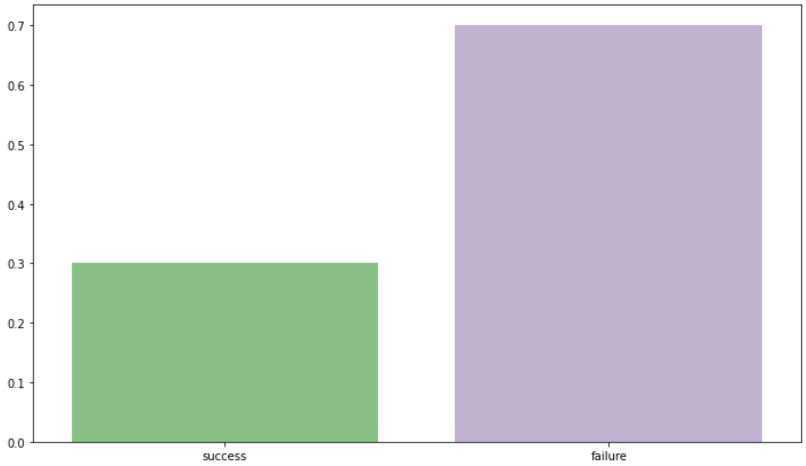
$$P(n)=\begin{cases}1-p & n=0\\ p & n=1\end{cases}$$

It can also be generalised as,

$$P(n)=p^n\times(1-p)^{1-n}$$

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.set_palette('Accent')
```

```
fig,ax = plt.subplots(figsize=(12,7))
sns.barplot(y=[0.3,0.7], x = ['success','failure'],)
```



Binomial Distribution

Binomial Distribution is simply an extension of Bernoulli distribution. If we repeat Bernoulli trials for n times, we will get a Binomial distribution. If we want to model the number of successes in n trials, we use Binomial Distribution. As each unit of Binomial is a Bernoulli trial, the outcome is always binary. The observations are independent of each other.

The Probability Mass Function is given by

$$f(x)=\binom{n}{k}p^k\times(1-p)^k$$

where

and p is the probability of success

As binomial distribution is Bernoulli trials taken n number of times, the mean and variance are given by

$$E(x) = np$$

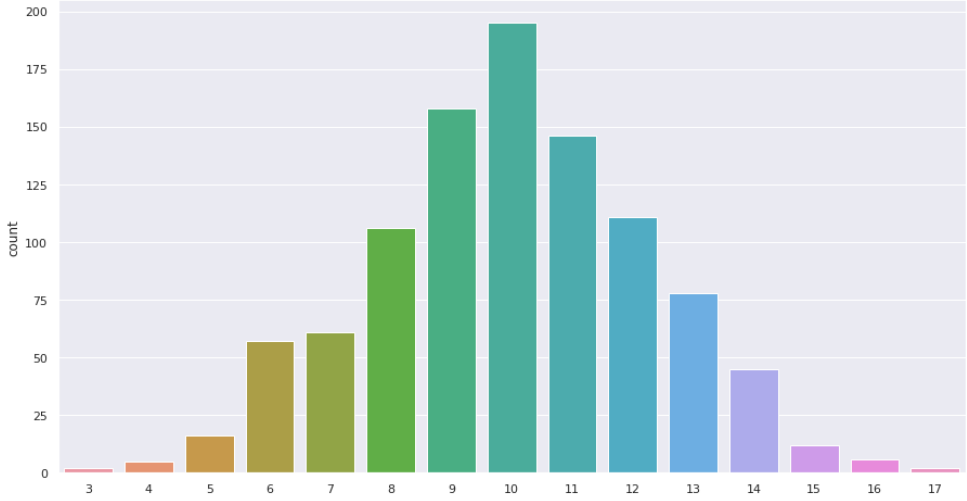
$$Var(x) = np(1-p)$$

The cumulative distribution function is given by

$$p(X\leq k)=\sum_{i=0}^{k}\binom{n}{k}p^i(1-p)^{n-i}$$

```
import numpy as np
import seaborn as sns
```

```
sns.set(style="darkgrid", palette="muted")
fig,ax = plt.subplots(figsize=(15,8))
binomial = np.random.binomial(20,0.5,1000)
sns.countplot(binomial)
```



Poisson Distribution

Poisson Distribution describes the probability of a given number of events occurring in a fixed interval, for example, the number of unique pageviews on an article on a given day or the number of customers visiting a florist shop at a particular time. It is not just limited to time intervals, and we can also extend its use to the area, length and volume intervals. For example, total rainfalls in a particular area.

The Probability Mass Function for Poisson distribution is given by

$$p(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{for} \quad x = 0, 1, 2, \cdots$$

Here, lambda is the shape parameter that describes the average number of events in that interval.

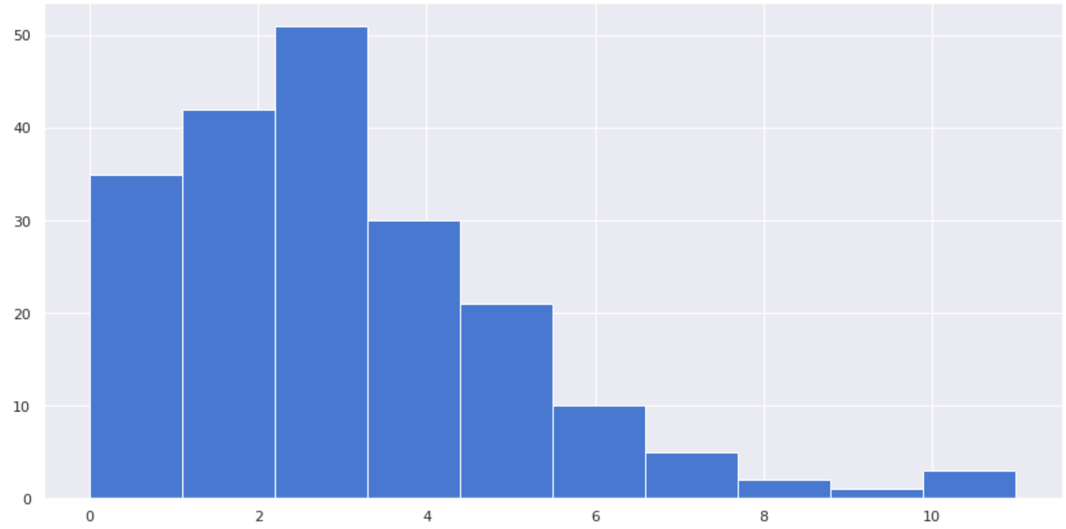
lambda is also the mean and variance of the distribution.

The cumulative distribution function is given as

$$F(x;\lambda) = \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}$$

We will use scipy to simulate the Poisson distribution.

```
from scipy.stats import poisson
sns.set(style="darkgrid", palette="muted")
res = poisson.rvs(mu=3, size=200)
fig,ax = plt.subplots(figsize=(14,7))
plt.hist(x=res,)
plt.show()
```



Continuous Distributions

Continuous distributions, unlike discrete distributions, have smooth curves consisting of an infinite number of samples which means any random variable virtually can take any value. Weight, Height, stock prices, and the half-life of an object have continuous distributions. Unlike discrete statistical distributions, it is impossible to calculate the probability of a random variable equating to a certain value. For example, finding the probability that someone’s height is approximately 167cm is impossible. We are not talking about 167.0002 or 167.00001 but exact 167 cm, which is impossible to find in a continuous distribution. A range is used to find the probability of any value within the same range.

Normal Distribution

The most common and naturally occurring distribution is Normal Distribution. It is otherwise also known as Gaussian Distribution. There is no field where this distribution is not seen. Finance, Statistics, Chemistry, you name it. It is an omnipresent distribution. A classic example could be the distribution of SAT scores higher number of students will score around the mean. As the distance increases from either side of the mean, the probability decreases.

The shape of the distribution resembles that of a classic bell, and hence it is called a bell-shaped curve. The curve is symmetric about its mean, and the variance defines the thickness of the distribution.

The mean, median and mode are the same in the case of Normal distribution.

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

The Probability Distribution Function is given as

- Here, x = any random variable
- mu = mean
- sigma = standard deviation

When mu is 0 and variance/standard deviation is 1, it becomes Standard Normal Distribution.

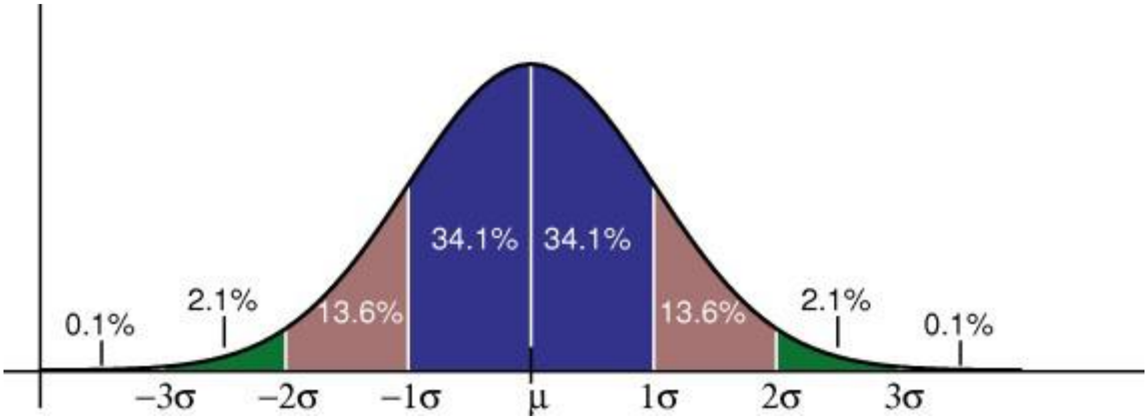
$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

The cumulative distribution function like we discussed, is the integral of PDF from negative infinity to the point for a standard normal distribution is given by

$$F(x) = p(X \le x) = \int_{-\infty}^{x} f(t)\,dt$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\,dt.$$

The Empirical Rule in normal distribution explains the probability of any random variable based on its standard deviation distance from the mean.



source: statistics

Chi^2 Distribution

A chi-square distribution can be defined as the distribution of the sum of the squares of v random variables drawn from a Gaussian or Normal distribution. The value v is also the degree of freedom of the distribution and the only parameter. For example, if 20 random variables are drawn from a normal distribution, the degree of freedom will be 20.

Unlike other distributions we have studied so far, the chi-square distribution doesn’t occur naturally. It is a theoretical distribution where the observations are calculated from the observations of a Normal distribution. The chi-square distribution is typically used in statistical significance testing where the underlying distribution is Normal.

The Probability Density Function is given by

$$f(x) = \frac{e^{\frac{-x}{2}} x^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} \qquad \text{for } x \geq 0$$

Where the capital letter gamma is the gamma function.

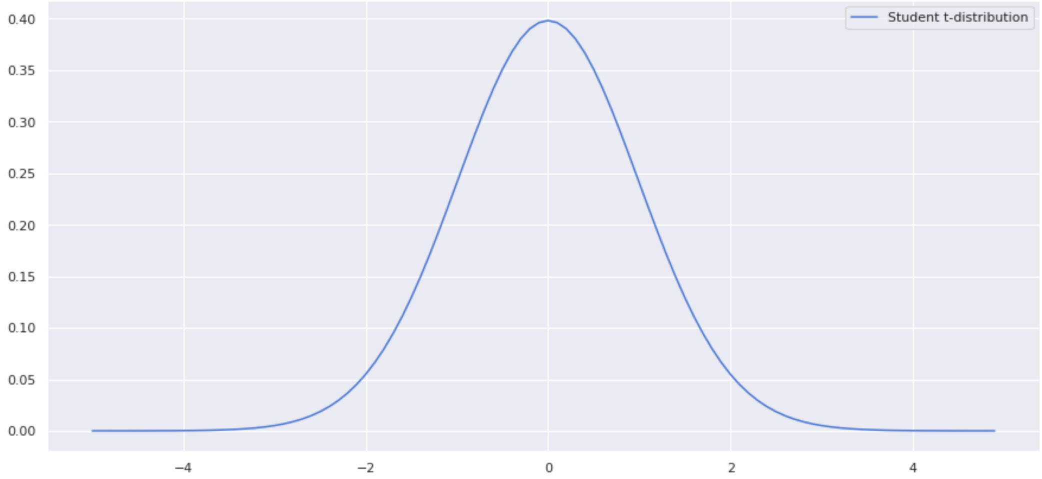
And the cumulative density function is given by

$$F(x) = \frac{\gamma(\frac{\nu}{2}, \frac{x}{2})}{\Gamma(\frac{\nu}{2})} \qquad \text{for } x \geq 0$$

Student’s t-Distribution

In general, students’ t-distribution or t distribution is predominantly used in significance testing and construction of confidence intervals. Just as chi-squared distribution, t-distribution also doesn’t occur naturally. This distribution arises while estimating the mean of a normal distribution when the population parameters are unknown, and the sample size is relatively small.

```
from scipy.stats import norm
from scipy.stats import t
sample_space = np.arange(-5, 5, 0.1)
dof = len(sample_space) - 1
fig,ax = plt.subplots(figsize=(15,7))
sns.lineplot(x = sample_space, y = t.pdf(sample_space,dof), label ='Student t-distribution')
```



Log-Normal Distribution

The log-Normal distribution is a continuous distribution of random variables, whereas the natural logarithm of these random variables is a Normal distribution. So, if X is any log-normally distributed random variable, then ln(X) follows a Normal distribution. A Log-Normal distribution always yields positive values as opposed to Normal distribution. The log-normal distribution is used where we do not want to let go of the convenience of Normal Distribution yet want only positive outcomes. Height, Weight, Amount of milk production, Amount of rainfall etc., are cases where we can use the Log-Normal distribution.

The Probability Density Function is given by

$$f(x) = \frac{e^{-((\ln((x-\theta)/m))^2/(2\sigma^2))}}{(x-\theta)\sigma\sqrt{2\pi}} x > \theta; m, \sigma > 0$$

- here, the mu = location parameter tells about the location of the x-axis
- sigma = standard deviation
- m = the scale parameter responsible for shrinking of distributions
- When the theta=0 and m=1, it is called the Standard log-normal distribution.

When the theta = 0 and m = 1, it is called standard log-normal distribution.

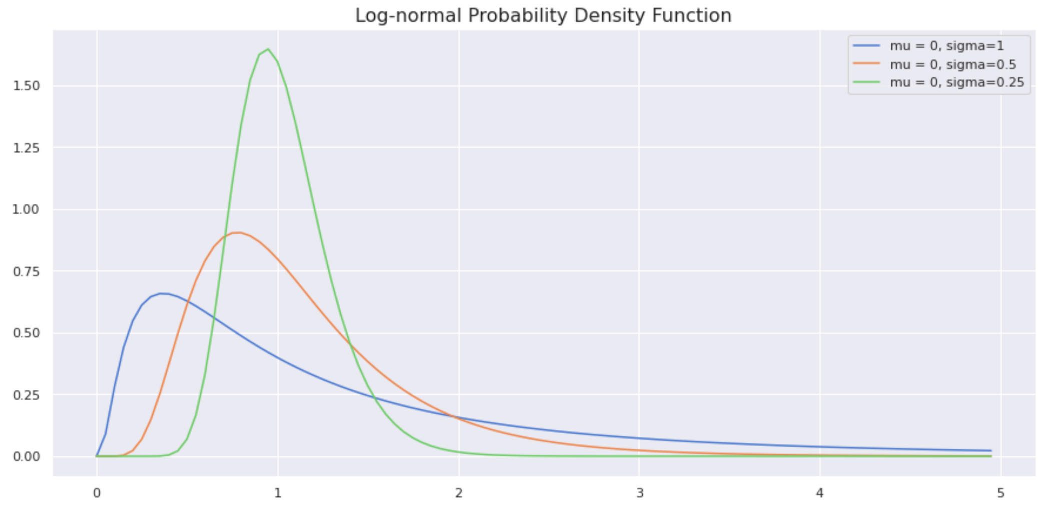
$$f(x) = \frac{e^{-((\ln x)^2/2\sigma^2)}}{x\sigma\sqrt{2\pi}} x > 0; \sigma > 0$$

The Cumulative Distribution Function is given as

$$F(x) = \Phi(\frac{\ln(x)}{\sigma})x \geq 0; \sigma > 0$$

Where phi is the CDF of normal distribution

```
from scipy.stats import lognorm
sample_space = np.arange(0,5,0.05)
sns.set(style="darkgrid", palette="muted",)
fig,ax = plt.subplots(figsize=(15,7))
sns.lineplot(x=sample_space, y = lognorm.pdf(sample_space, 1,),
      label='mu = 0, sigma=1')
sns.lineplot(x = sample_space, y = lognorm.pdf(sample_space, 0.5,),
      label='mu = 0, sigma=0.5')
sns.lineplot(x = sample_space, y=lognorm.pdf(sample_space, 0.25,),
      label='mu = 0, sigma=0.25')
plt.title('Log-normal Probability Density Function',fontdict = {'size':16})
plt.plot()
```



6. Fitting a model

What is a model

In the physical world, “models” are generally simplifications of things in the real world that nonetheless convey the essence of the thing being modeled. A model of a building conveys the structure of the building while being small and light enough to pick up with one’s hands; a model of a cell in biology is much larger than the actual thing, but again conveys the major parts of the cell and their relationships

6.1 Statistical modeling: An example

Let’s look at an example of building a model for data, using the data from NHANES. In particular, we will try to build a model of the height of children in the NHANES sample. First let’s load the data and plot them (see Figure 5.1).

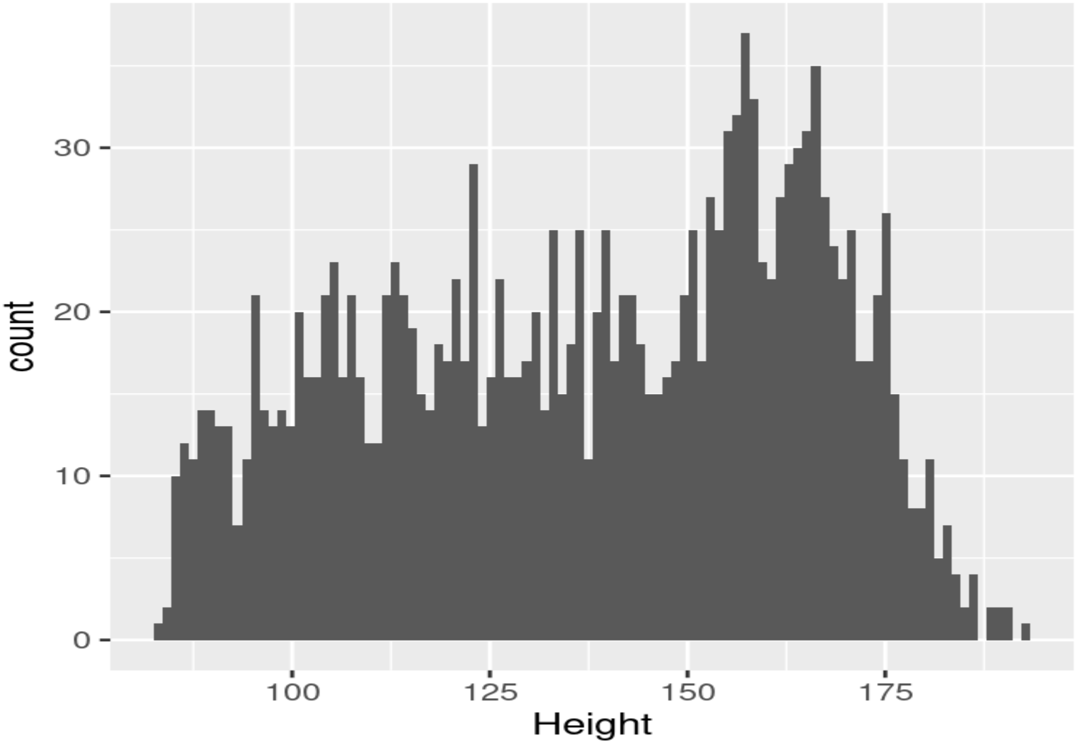


Figure 5.1: Histogram of height of children in NHANES.

The subscript doesn’t appear on the right side of the equation, which means that the prediction of the model doesn’t depend on which observation we are looking at — it’s the same for all of them. The question then becomes: how do we estimate the best values of the parameter(s) in the model?
One very simple estimator that we might imagine is the *mode*, which is simply the most common value in the dataset. This redescribes the entire set of 1691 children

How good of a model is this? In general we define the goodness of a model in terms of the magnitude of the error, which represents the degree to which the data

diverge from the model's predictions; all things being equal, the model that produces lower error is the better model.
How might we find a better estimator for our model parameter? We might start by trying to find an estimator that gives us an average error of zero.

It turns out that if we use the arithmetic mean as our estimator then the average error will indeed be zero (see the simple proof at the end of the chapter if you are interested). Even though the average of errors from the mean is zero,

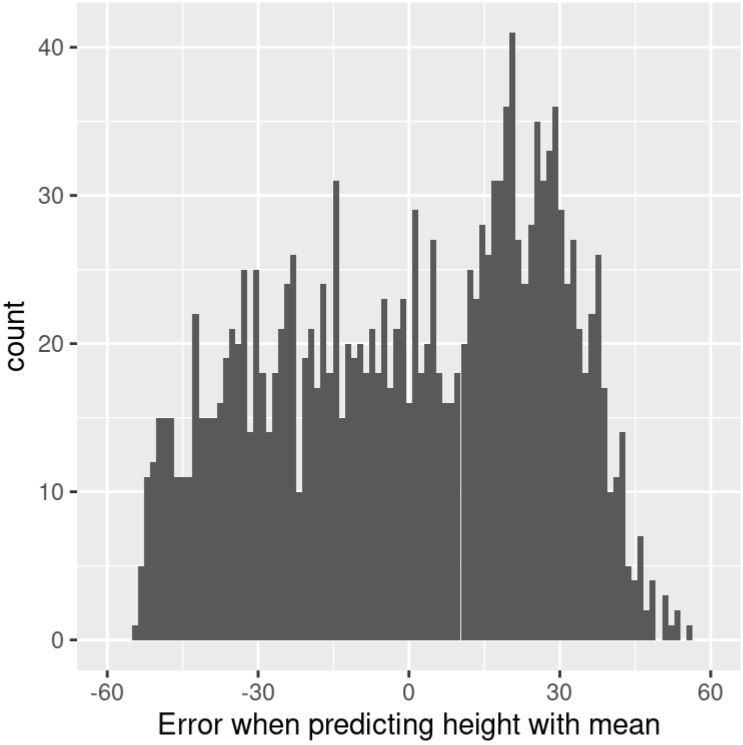


Figure 5.2: Distribution of errors from the mean.

The fact that the negative and positive errors cancel each other out means that two different models could have errors of very different magnitude in absolute terms, but would still have the same average error. This is exactly why the average error is not a good criterion for our estimator; we want a criterion that tries to minimize the overall error regardless of its direction. For this reason, we generally summarize errors in terms of some kind of measure that counts both positive and negative errors as bad. We could use the absolute value of each error value, but it's more common to use the squared errors, for reasons that we will see later in the book.

There are several common ways to summarize the squared error that you will encounter at various points in this book, so it's important to understand how they relate to one another. First, we could simply add them up; this is referred to as the *sum of squared errors*.

6.2 Improving our model

Can we imagine a better model? Remember that these data are from all children in the NHANES sample, who vary from 2 to 17 years of age. Given this wide age range, we might expect that our model of height should also include age. Let's plot the data for height against age, to see if this relationship really exists.

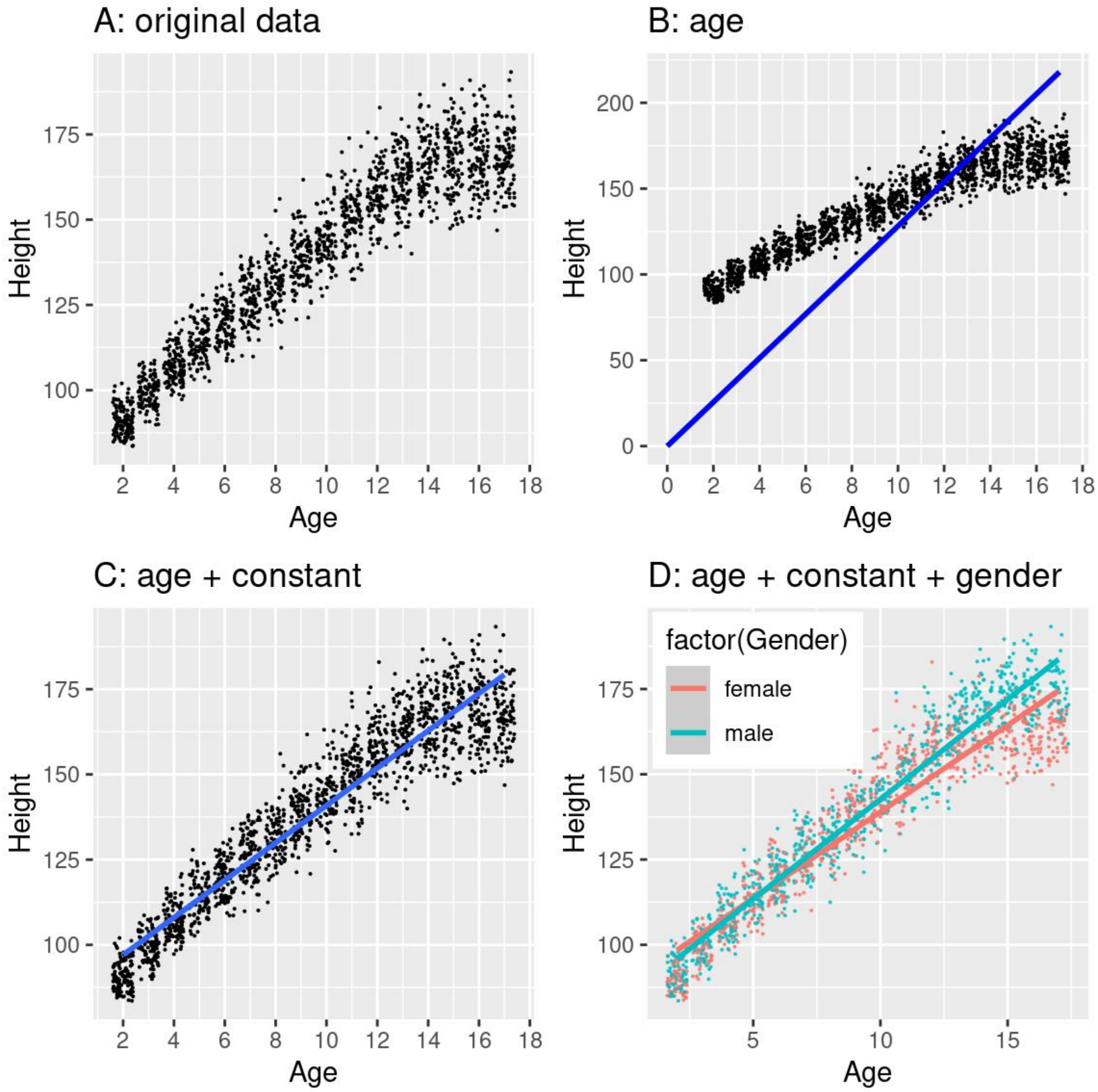


Figure 5.3: Height of children in NHANES, plotted without a model (A),

with a linear model including only age (B)

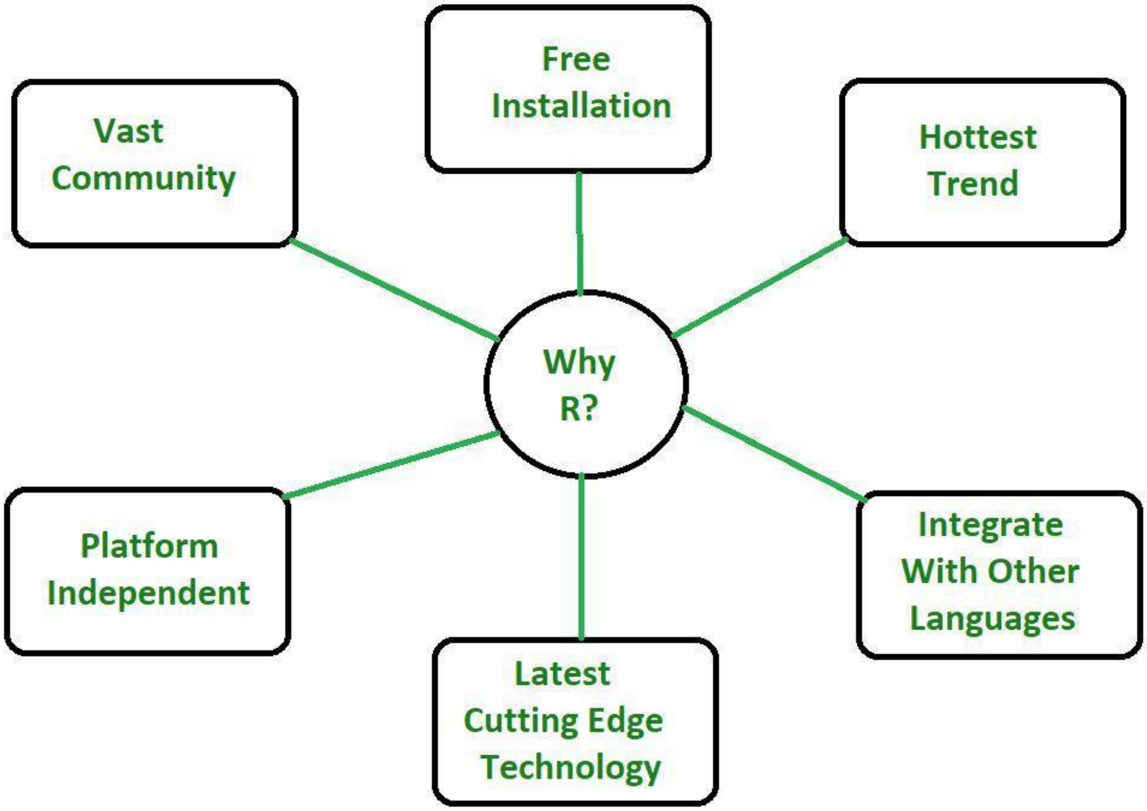
or age and a constant (C),

and with a linear model that fits separate effects of age for males and females (D).

7. R Programming Language – Introduction

R is an open-source programming language that is widely used as a statistical software and data analysis tool. R generally comes with the Command-line interface. R is available across widely used platforms like Windows, Linux, and macOS. Also, the R programming language is the latest cutting-edge tool.

R programming language is an implementation of the S programming language. It also combines with lexical scoping semantics inspired by Scheme. Moreover, the project was conceived in 1992, with an initial version released in 1995 and a stable beta version in 2000.



R Programming Language

What is R Programming Language?

- R programming is used as a leading tool for machine learning, statistics, and data analysis. Objects, functions, and packages can easily be created by R.
- It's a platform-independent language. This means it can be applied to all operating systems.
- It's an open-source free language. That means anyone can install it in any organization without purchasing a license.
- R programming language is not only a statistic package but also allows us to integrate with other languages (C, C++). Thus, you can easily interact with many data sources and statistical packages.
- The R programming language has a vast community of users and it's growing day by day.

- R is currently one of the most requested programming languages in the Data Science job market which makes it the hottest trend nowadays
- It was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently being developed by the R Development Core Team.
- R programming language is an implementation of the S programming language. It also combines with lexical scoping semantics inspired by Scheme. Moreover, the project was conceived in 1992, with an initial version released in 1995 and a stable beta version in 2000.

Why Use R?

- **Statistical Analysis:** R is designed for analysis and It provides an extensive collection of graphical and statistical techniques, By making a preferred choice for statisticians and data analysts.
- **Open Source:** R is an open – source software, which means it is freely available to anyone. It can be accessble by a vibrant community of users and developers.
- **Data Visulaization :** R boasts an array of libraries like ggplot2 that enable the creation of high-quality, customizable data visualizations.
- **Data Manipulation :** R offers tools that are for data manipulation and transformation. For example: IT simplifies the process of filtering , summarizing and transforming data.
- **Integration :** R can be easily integrate with other programming languages and data sources. IT has connectors to various databases and can be used in conjunction with python, SQL and other tools.
- **Community and Packages:** R has vast ecosystem of packages that extend its functionality. There are packages that can help you accomplish needs of analytics.

Features of R Programming Language

- **R Packages:** One of the major features of R is it has a wide availability of libraries. R has CRAN(**Comprehensive R Archive Network**), which is a repository holding more than 10, 0000 packages.
- **Distributed Computing:** Distributed computing is a model in which components of a software system are shared among multiple computers to improve efficiency and performance. Two new packages ddR and multidplyr used for distributed programming in R were released in November 2015.

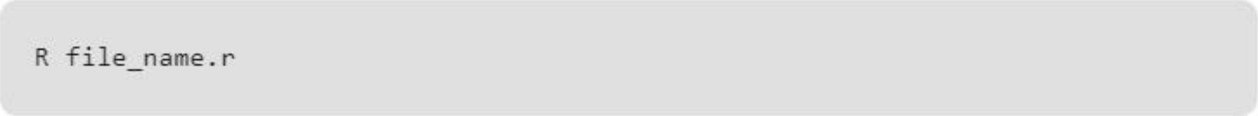
Statistical Features of R

- **Basic Statistics:** The most common basic statistics terms are the mean, mode, and median. These are all known as “Measures of Central Tendency.” So using the R language we can measure central tendency very easily.
- **Static graphics:** R is rich with facilities for creating and developing interesting static graphics. R contains functionality for many plot types including graphic maps, mosaic plots, biplots, and the list goes on.
- **Probability distributions:** Probability distributions play a vital role in statistics and by using R we can easily handle various types of probability distributions such as Binomial Distribution, Normal Distribution, Chi-squared Distribution, and many more.

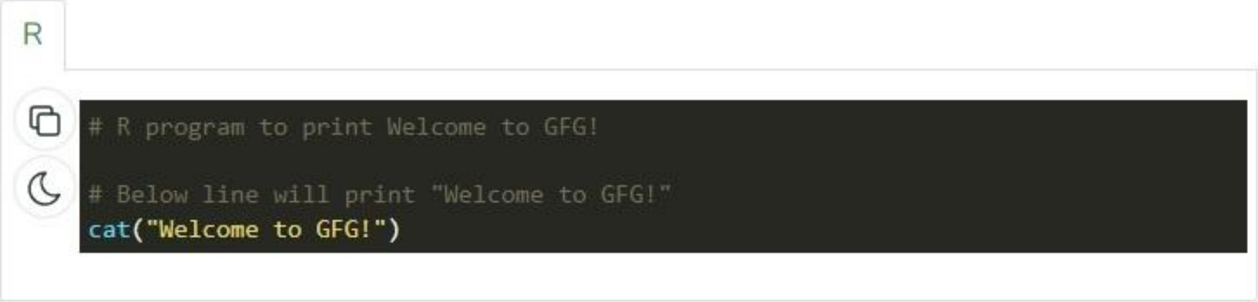
- **Data analysis:** It provides a large, coherent, and integrated collection of tools for data analysis.

Basic R program

Since R is much similar to other widely used languages syntactically, it is easier to code and learn in R. Programs can be written in R in any of the widely used IDE like **R Studio**, **Rattle, Tinn-R**, etc. After writing the program save the file with the extension **.r**. To run the program use the following command on the command line:

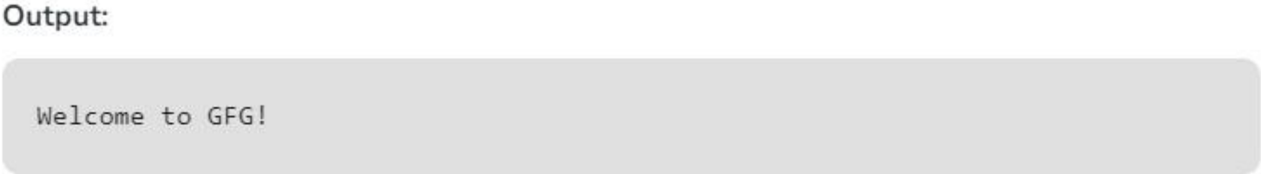


```
R file_name.r
```



```r
# R program to print Welcome to GFG!

# Below line will print "Welcome to GFG!"
cat("Welcome to GFG!")
```



Output:

```
Welcome to GFG!
```

Advantages of R

- R is the most comprehensive statistical analysis package. As new technology and concepts often appear first in R.
- As R programming language is an open source. Thus, you can run R anywhere and at any time.
- R programming language is suitable for GNU/Linux and Windows operating systems.
- R programming is cross-platform and runs on any operating system.
- In R, everyone is welcome to provide new packages, bug fixes, and code enhancements.

Disadvantages of R

- In the R programming language, the standard of some packages is less than perfect.
- Although, R commands give little pressure on memory management. So R programming language may consume all available memory.
- In R basically, nobody to complain if something doesn't work.

- R programming language is much slower than other programming languages such as Python and MATLAB.

Applications of R

- We use R for Data Science. It gives us a broad variety of libraries related to statistics. It also provides the environment for statistical computing and design.
- R is used by many quantitative analysts as its programming tool. Thus, it helps in data importing and cleaning.
- R is the most prevalent language. So many data analysts and research programmers use it. Hence, it is used as a fundamental tool for finance.
- Tech giants like Google, Facebook, Bing, Twitter, Accenture, Wipro, and many more using R nowadays.

DATA SCIENCE

1st – Unit Notes

PRABHAKAR NAIDU .R
PRINCIPAL-MCA