UNIT-1

Introduction to Statistics

Statistics is a mathematical science that includes methods for collecting, organizing, analyzing and visualizing data in such a way that meaningful conclusions can be drawn.

Statistics is also a field of study that summarizes the data, interpret the data making decisions based on the data.

Statistics is composed of two broad categories:

1. Descriptive Statistics
2. Inferential Statistics

1. Descriptive Statistics

Descriptive statistics describes the characteristics or properties of the data. It helps to summarize the data in a meaningful data in a meaningful way. It allows important patterns to emerge from the data. Data summarization techniques are used to identify the properties of data. It is helpful in understanding the distribution of data. They do not involve in generalizing beyond the data.

1.1 Two types of descriptive statistics

1. Measures of Central Tendency: (Mean , Median , Mode)
2. Measures of data spread or dispersion (range, quartiles, variance and standard deviation)

1.1.1 Measures of Central Tendency: (Mean , Median , Mode)

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. The mean, median and mode are all valid measures of central tendency.

Mean (Arithmetic)

The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data.

The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have values in a data set and they have values $x_1, x_2, \ldots x_n$, the sample mean, usually denoted by $\bar{x}$.

$\bar{x} = (x_1 x_2, \ldots x_n)/ n$.

An important property of the mean is that it includes every value in the data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

Median:

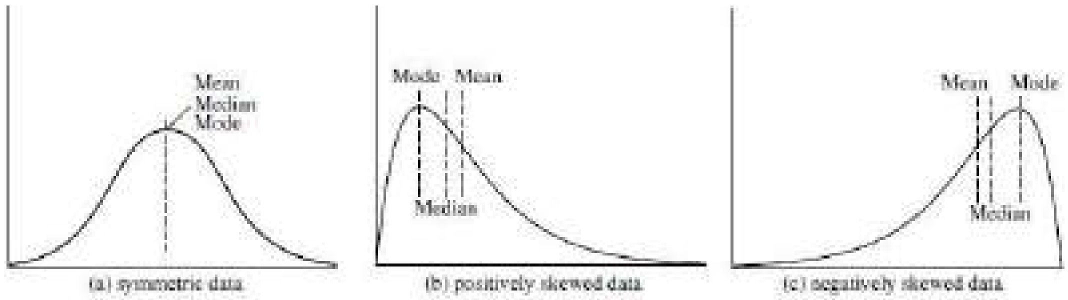
The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. It is a holistic measure. It is easy method of approximation of median value of a large data set.

Mode

The mode is the most frequent score in our data set. The mode is used for categorical data where we want to know which is the most common category occurring in the population. There are possibilities for the greatest frequency to correspond to different values. This results in more than one,two or more modes in a dataset. They are called as unimodal, bimodal and multimodal datasets. If each data occurs only once then the mode is equal to zero.

Unimodal frequency curve with symmetric data distribution , the mean median and mode are all the same.

In real applications the data is not symmetrical and they are asymmetric.It might be positively skewed or negatively skewed. If positively skewed then mode is smaller than median and in negatively skewed the mode occurs at a value greater than the median.



(a) symmetric data (b) positively skewed data (c) negatively skewed data

Mean, median, and mode of symmetric versus positively and negatively skewed data.

1.1.2 Measures of spread:

Measures of spread are the ways of summarizing a group of data by describing how scores are spread out. To describe this spread, a number of statistics are available to us, including the range, quartiles, absolute deviation, variance and standard deviation.

- The degree to which numerical data tend to spread is called the dispersion, or variance of the data. The common measures of data dispersion: Range, Quartiles, Outliers, and Boxplots.

Range : Range of the set is the difference between the largest (max()) and smallest (min()) values.
Ex: Step 1: Sort the numbers in order, from smallest to largest: 7, 10, 21, 33, 43, 45, 45, 65, 67, 87, 98, 99

Step 2: Subtract the smallest number in the set from the largest number in the set:
99 – 7 = 92

The range is 92

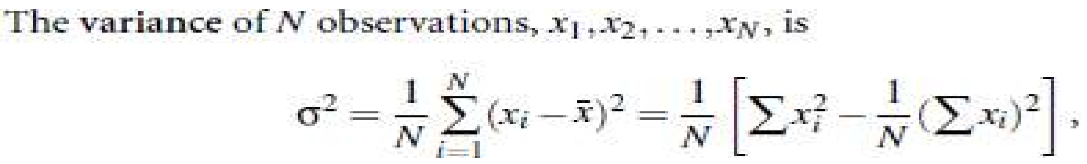
Quartiles : Percentile : kth percentile of a set of data in numerical order is the value xi having the property that k percent of the data entries lie at or below xi

- The first quartile (Q1) is the 25th percentile;
- The third quartile (Q3) is the 75th percentile
- The distance between the first and third quartiles is the range covered by the middle half of the data.
- Interquartile range (IQR) and is defined as IQR = Q3 - Q1.
- Outliers is to single out values falling at least 1.5 *IQR above the third quartile or below the first quartile.
- Five-number summary: median, the quartiles Q1 and Q3, and the smallest and largest individual observations comprise the five number summary: Minimum; Q1; Median; Q3; Maximum

  Example : Quartiles
- Start with the following data set:
- 1, 2, 2, 3, 4, 6, 6, 7, 7, 7, 8, 11, 12, 15, 15, 15, 17, 17, 18, 20
- There are a total of twenty data points in the set. There is an even number of data values, hence the median is the mean of the tenth and eleventh values.
- the median is: (7 + 8)/2 = 7.5.
- The median of the first half of the set is found between the fifth and sixth values of:
- 1, 2, 2, 3, 4, 6, 6, 7, 7, 7
- Thus the first quartile is found to equal Q1 = (4 + 6)/2 = 5
- To find the third quartile, examine the top half of the original data set. The median of
- 8, 11, 12, 15, 15, 15, 17, 17, 18, 20
- is (15 + 15)/2 = 15. Thus the third quartile Q3 = 15.

A small interquartile range indicates data that is clumped about the median. A larger interquartile range shows that the data is more spread out

Variance and Standard Deviation



The variance of N observations, $x_1, x_2, \ldots, x_N$, is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right],$$

Inferential Statistics – Definition and Types

Inferential statistics is generally used when the user needs to make a conclusion about the whole population at hand, and this is done using the various types of tests available. It is a technique which is used to understand trends and draw the required conclusions about a large population by taking and analyzing a sample from it. Descriptive statistics, on the other hand, is only about the smaller sized data set at hand – it usually does not involve large populations. Using variables and the relationships between them from the sample, we will be able to make generalizations and predict other relationships within the whole population, regardless of how large it is.

With inferential statistics, data is taken from samples and generalizations are made about a population. Inferential statistics use statistical models to compare sample data to other samples or to previous research.

There are two main areas of inferential statistics:

1. **Estimating parameters:**

This means taking a statistic from the sample data (for example the sample mean) and using it to infer about a population parameter (i.e. the population mean).There may be sampling variations because of chance fluctuations, variations in sampling techniques, and other sampling errors. Estimation about population characteristics may be influenced by such factors. Therefore, in estimation the important point is that to what extent our estimate is close to the true value.

Characteristics of Good Estimator: A good statistical estimator should have the following characteristics, (i) Unbiased (ii) Consistent (iii) Accuracy

i) Unbiased

An unbiased estimator is one in which, if we were to obtain an infinite number of random samples of a certain size, the mean of the statistic would be equal to the parameter. The sample mean, (x) is an unbiased estimate of population mean (μ)because if we look at possible random samples of size N from a population, then mean of the sample would be equal to μ.

ii) Consistent

A consistent estimator is one that as the sample size increased, the probability that estimate has a value close to the parameter also increased. Because it is a consistent estimator, a sample mean based on 20 scores has a greater probability of being closer to (μ) than does a sample mean based upon only 5 scores

iii) Accuracy

The sample mean is an unbiased and consistent estimator of population mean (μ).But we should not over look the fact that an estimate is just a rough or approximate calculation. It is unlikely in any estimate that (x) will be exactly equal to population mean (μ). Whether or not x is a good estimate of (μ) depends upon the representativeness of sample, the sample size, and the variability of scores in the population.

2. **<u>Hypothesis tests</u>.** This is where sample data can be used to answer research questions. For example, we might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

Inferential statistics is closely tied to the logic of hypothesis testing. We hypothesize that this value characterise the population of observations. The question is whether that hypothesis is reasonable evidence from the sample. Sometimes hypothesis testing is referred to as statistical decision-making process. In day-to-day situations we are required to take decisions about the population on the basis of sample information.

2.6.1 Statement of Hypothesis

A statistical hypothesis is defined as a statement, which may or may not be true about the population parameter or about the probability distribution of the parameter that we wish to validate on the basis of sample information. Most times, experiments are performed with random samples instead of the entire population and inferences drawn from the observed results are then generalised over to the entire population. But before drawing inferences about the population it should be always kept in mind that the observed results might have come due to chance factor. In order to have an accurate or more precise inference, the chance factor should be ruled out.

<u>Null Hypothesis</u>

The probability of chance occurrence of the observed results is examined by the null hypothesis (H0). Null hypothesis is a statement of no differences. The other way to state null hypothesis is that the two samples came from the same population. Here, we assume that population is normally distributed and both the groups have equal means and standard deviations.

Since the null hypothesis is a testable proposition, there is counter proposition to it known as alternative hypothesis and denoted by H1 . In contrast to null hypothesis, the alternative hypothesis (H1) proposes that

i) the two samples belong to two different populations,

ii) their means are estimates of two different parametric means of the respective population, and

iii) there is a significant difference between their sample means.

The alternative hypothesis (H1) is not directly tested statistically; rather its acceptance or rejection is determined by the rejection or retention of the null hypothesis. The probability ‘p’ of the null hypothesis being correct is assessed by a statistical test. If probability ‘p’ is too low, H0 is rejected and H1 is accepted.

It is inferred that the observed difference is significant. If probability ‘p’ is high, H0 is accepted and it is inferred that the difference is due to the chance factor and not due to the variable factor.

2.6.2 Level of Significance

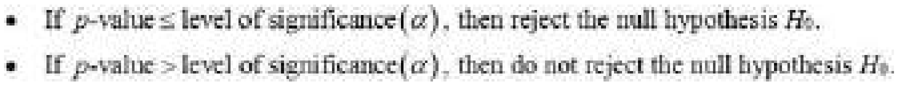
The level of significance is defined as the probability of rejecting a null hypothesis by the test when it is really true, which is denoted as α. That is, P (Type I error) = α.

Confidence level:

Confidence level refers to the possibility of a parameter that lies within a specified range of values, which is denoted as c. Moreover, the confidence level is connected with the level of significance. The relationship between level of significance and the confidence level is c=1−α. The common level of significance and the corresponding confidence level are given below:

- The level of significance 0.10 is related to the 90% confidence level.
- The level of significance 0.05 is related to the 95% confidence level.
- The level of significance 0.01 is related to the 99% confidence level.

The rejection rule is as follows:



- If p-value ≤ level of significance(α), then reject the null hypothesis H_0.
- If p-value > level of significance(α), then do not reject the null hypothesis H_0.

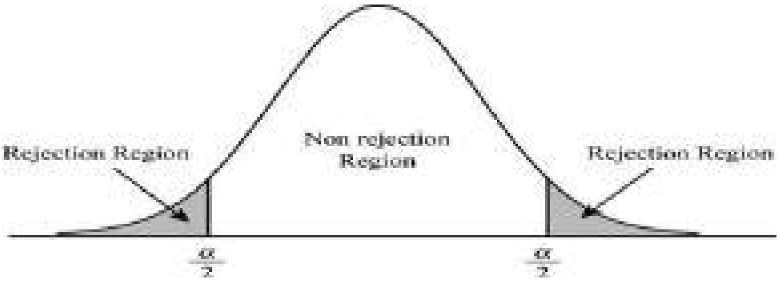
Rejection region:

The rejection region is the values of test statistic for which the null hypothesis is rejected.

Non rejection region:

The set of all possible values for which the null hypothesis is not rejected is called the rejection region.

The rejection region for two-tailed test is shown below:



The rejection region for one-tailed test is given below:

In the left-tailed test, the rejection region is shaded in left side.

In the right-tailed test, the rejection region is shaded in right side.

2.6.3 One-tail and Two-tail Test

Depending upon the statement in alternative hypothesis (H1), either a one-tail or two tail test is chosen for knowing the statistical significance. A one-tail test is a directional test. It is formulated to find the significance of both the magnitude and the direction (algebraic sign) of the observed difference between two statistics. Thus, in two-tailed tests researcher is interested in testing whether one sample mean is significantly higher (alternatively lower) than the other sample mean.

Types of Inferential Statistics Tests

There are many tests in this field, of which some of the most important are mentioned below.

1. Linear Regression Analysis

In this test, a linear algorithm is used to understand the relationship between two variables from the data set. One of those variables is the dependent variable, while there can be one or more independent variables used. In simpler terms, we try to predict the value of the dependent variable based on the available values of the independent variables. This is usually represented by using a scatter plot, although we can also use other types of graphs too.

2. Analysis of Variance

This is another statistical method which is extremely popular in data science. It is used to test and analyse the differences between two or more means from the data set. The significant differences between the means are obtained, using this test.

3. Analysis of Co-variance

This is only a development on the Analysis of Variance method and involves the inclusion of a continuous co-variance in the calculations. A co-variate is an independent variable which is continuous, and is used as regression variables. This method is used extensively in statistical modelling, in order to study the differences present between the average values of dependent variables.

4. Statistical Significance (T-Test)

A relatively simple test in inferential statistics, this is used to compare the means of two groups and understand if they are different from each other. The order of difference, or how significant the differences are can be obtained from this.

5. Correlation Analysis

Another extremely useful test, this is used to understand the extent to which two variables are dependent on each other. The strength of any relationship, if they exist, between the two variables can be obtained from this. You will be able to understand whether the variables have a strong correlation or a weak one. The correlation can also be negative or positive, depending upon the variables. A negative correlation means that the value of one variable decreases while the value of the other increases and positive correlation means that the value both variables decrease or increase simultaneously.

Differences between Descriptive and Inferential Statistics

Descriptive Statistics	Inferential Statistics
Concerned with describing the target population	Make inferences from the sample and generalize them to the population

Organise, analyse, present the data in a meaningful way	Compare, tests and predicts future outcomes
The analysed results are in the form of graphs, charts etc	The analysed results are the probability scores
Describes the data which is already known	Tries to make conclusions about the population beyond the data available
Tools: Measures of central tendency and measures of spread	Tools: Hypothesis tests, analysis of variance etc

Random Variables

A random variable, X, is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous.

Example of Random variable

- A person’s blood type
- Number of leaves on a tree
- Number of times a user visits LinkedIn in a day
- Length of a tweet.

Discrete Random Variables :

A discrete random variable is one which may take on only a countable number of distinct values such as 0,1,2,3,4,........ Discrete random variables are usually counts. If a random variable can take only a finite number of distinct values, then it must be discrete. Examples of discrete random variables include the number of children in a family, the Friday night attendance at a cinema, the number of patients in a doctor's surgery, the number of defective light bulbs in a box of ten.

The **probability distribution** of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function

Suppose a random variable X may take k different values, with the probability that X = x_i defined to be P(X = x_i) = p_i. The probabilities p_i must satisfy the following:

1: $0 \leq p_i \leq 1$ for each i

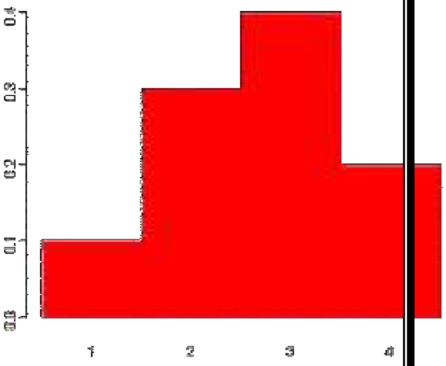
2: $p_1 + p_2 + ... + p_k = 1$.

Example

Suppose a variable X can take the values 1, 2, 3, or 4. The probabilities associated with each outcome are described by the following table:

Outcome	1	2	3	4
Probability	0.1	0.3	0.4	0.2

The probability that X is equal to 2 or 3 is the sum of the two probabilities: P(X = 2 or X = 3) = P(X = 2) + P(X = 3) = 0.3 + 0.4 = 0.7. Similarly, the probability that X is greater than 1 is equal to 1 - P(X = 1) = 1 - 0.1 = 0.9, by the complement rule.



Continuous Random Variables

A **continuous random variable** is one which takes an infinite number of possible values. Continuous random variables are usually measurements. Examples include height, weight, the amount of sugar in an orange, the time required to run a mile.

A continuous random variable is not defined at specific values. Instead, it is defined over an interval of values, and is represented by the **area under a curve** (known as an integral). The probability of observing any single value is equal to 0, since the number of values which may be assumed by the random variable is infinite.

Suppose a random variable X may take all values over an interval of real numbers. Then the probability that X is in the set of outcomes A, P(A), is defined to be the area above A and under a curve. The curve, which represents a function p(x), must satisfy the following:

1: The curve has no negative values (p(x) $\geq$ 0 for all x)

2: The total area under the curve is equal to 1.

A curve meeting these requirements is known as a **density curve**.

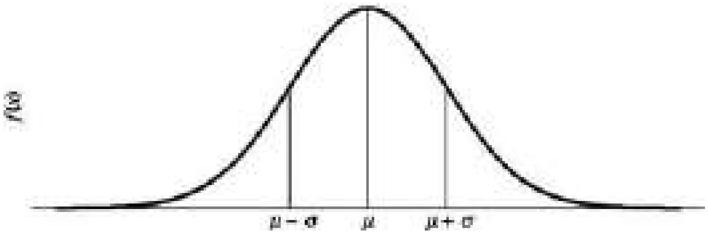
All random variables (discrete and continuous) have a **cumulative distribution function**. It is a function giving the probability that the random variable X is less than or equal to x, for every value x. For a discrete random variable, the cumulative distribution function is found by

summing up the probabilities.

Normal Probability Distribution

The Bell-Shaped Curve

The **Bell-shaped Curve** is commonly called the **normal curve** and is mathematically referred to as the Gaussian probability distribution. Unlike Bernoulli trials which are based on discrete counts, the **normal distribution** is used to determine the probability of a continuous random variable.



The **normal** or Gaussian Probability Distribution is most popular and important because of its unique mathematical properties which facilitate its application to practically any physical problem in the real world. The constants **μ** and $\boldsymbol{\sigma}^2$ are the parameters;

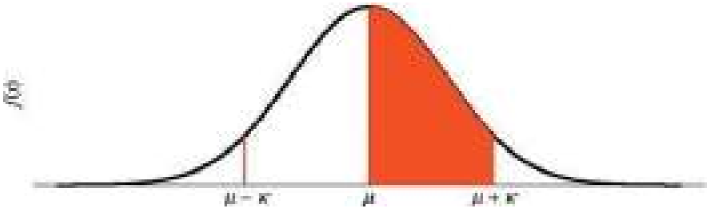
"**μ**" is the population true mean (or expected value) of the subject phenomenon characterized by the continuous random variable, **X**,

"$\boldsymbol{\sigma}^2$" is the population true variance characterized by the continuous random variable, **X**.

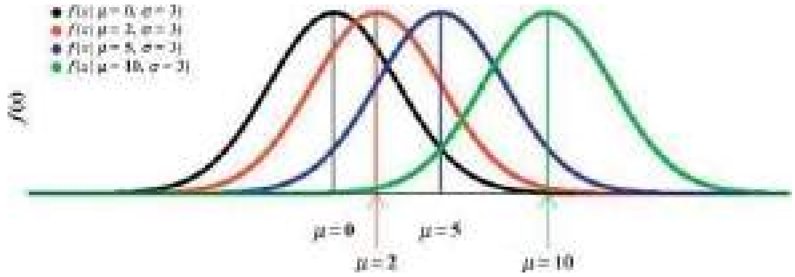
Hence, "**σ**" the population standard deviation characterized by the continuous random variable **X**;

the points located at μ−σ and μ+σ are the points of inflection; that is, where the graph changes from cupping up to cupping down

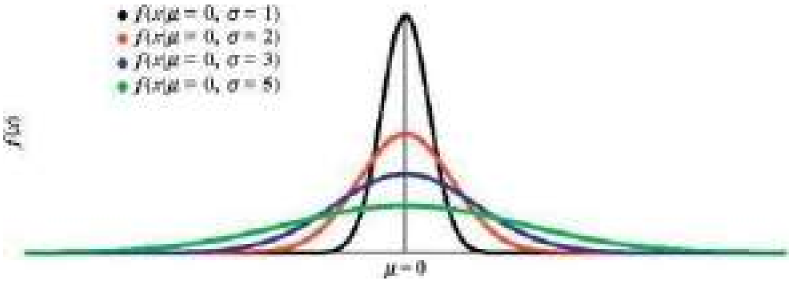
The **normal curve graph of the normal probability distribution**) is **symmetric** with respect to the mean **μ** as the **central position**. That is, the area between **μ** and **κ** units to the left of **μ** is equal to the area between **μ** and **κ** units to the right of **μ**.



There is not a unique **normal probability distribution.** The figure below is a graphical representation of the normal distribution for a fixed value of σ2 with μ varying.



The figure below is a graphical representation of the **normal distribution** for a fixed value of **μ** with varying σ^2.



SAMPLING and SAMPLING DISTRIBUTION

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. It helps us to make statistical inferences about the population. A population can be defined as a whole that includes all items and characteristics of the research taken into study. However, gathering all this information is time consuming and costly. We therefore make inferences about the population with the help of samples.

Random sampling:

In data collection, every individual observation has equal probability to be selected into a sample. In random sampling, there should be no pattern when drawing a sample.

Probability sampling:

It is the sampling technique in which every individual unit of the population has greater than zero probability of getting selected into a sample.

Non-probability sampling:

It is the sampling technique in which some elements of the population have no probability of getting selected into a sample.

Cluster samples:

It divides the population into groups (clusters). Then a random sample is chosen from the clusters.

Systematic sampling : select sample elements from an ordered frame. A sampling frame is just a list of participants that we want to get a sample from.

Stratified sampling : sample each subpopulation independently. First, divide the population into homogeneous (very similar) subgroups before getting the sample. Each population member only belongs to one group. Then apply simple random or a systematic method within each group to choose the sample.

Sampling Distribution

A sampling distribution is a probability distribution of a statistic. It is obtained through a large number of samples drawn from a specific population. It is the distribution of all possible values taken by the statistic when all possible samples of a fixed size n are taken from the population.

Sampling Distributions and Inferential Statistics

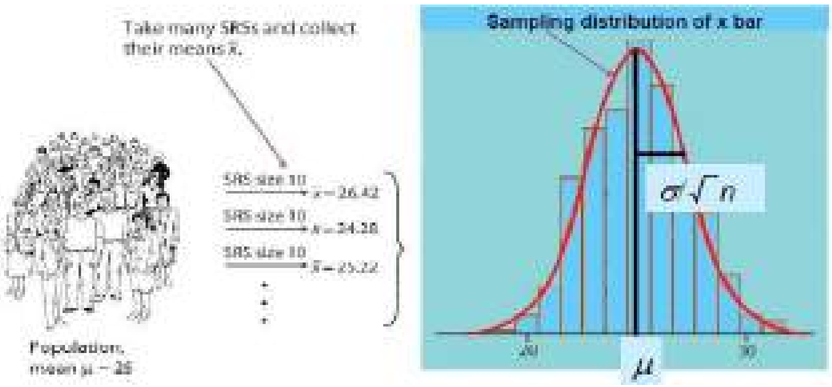
Sampling distributions are important for inferential statistics. A population is specified and the sampling distribution of the mean and the range were determined. In practice, the process proceeds the other way: the sample data is collected and from these data we estimate parameters of the sampling distribution. This knowledge of the sampling distribution can be very useful.

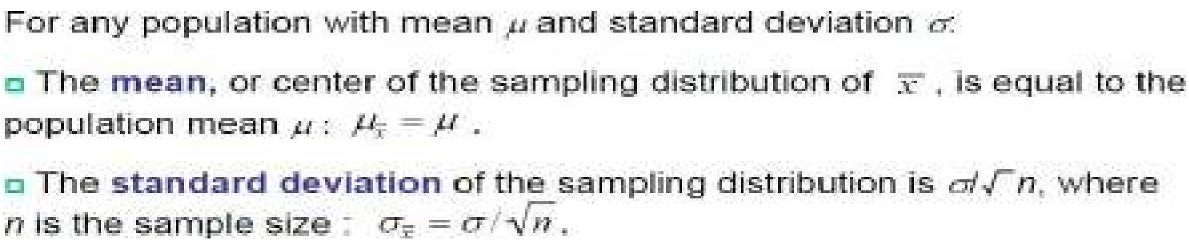
Knowing the degree to which means from different samples would differ from each other and from the population mean (this would give an idea of how close the particular sample mean is likely to be to the population mean)

The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean. This standard deviation is called the standard error of the mean.

If all the sample means were very close to the population mean, then the standard error of the mean would be small. On the other hand, if the sample means varied considerably, then the standard error of the mean would be large.

Sampling distribution of the sample mean

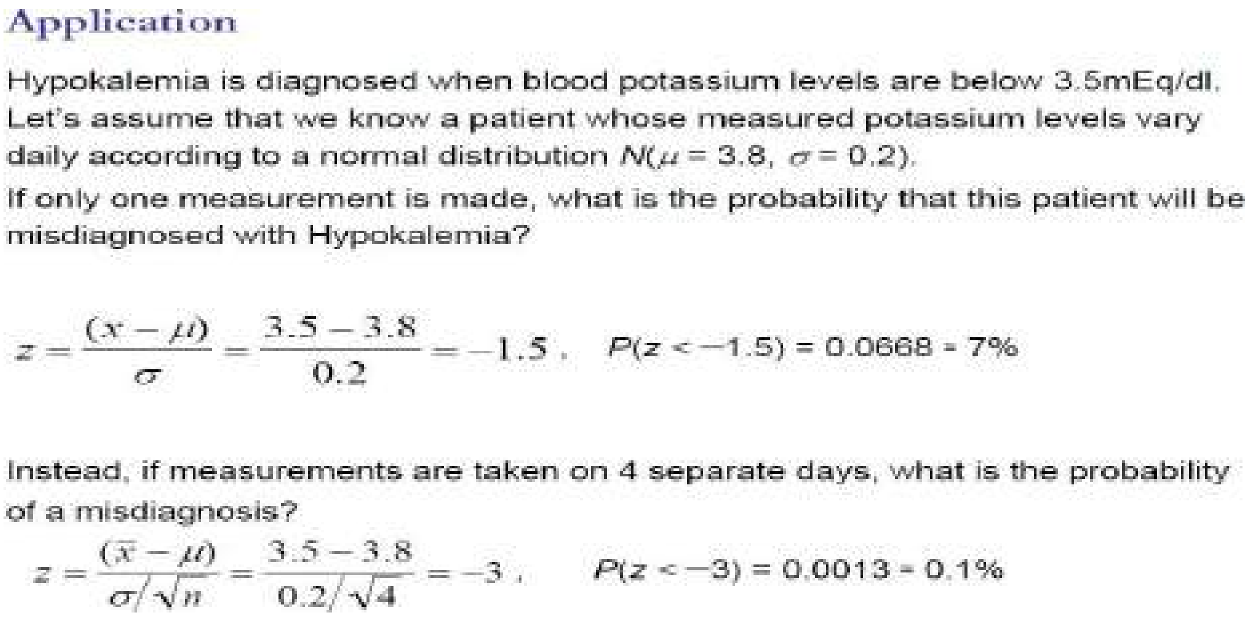
1. We take many random samples of a given size n from a population with mean μ and standard deviation σ. 2. Some sample means will be above the population mean μ and some will be below, making up the sampling distribution.	 Take many SRSs and collect their means $\bar{x}$. SRS size 10 $\bar{x} = 26.42$ SRS size 10 $\bar{x} = 24.28$ SRS size 10 $\bar{x} = 25.22$ Population, mean μ = 25 Sampling distribution of x bar σ √ n 20 30 μ



For any population with mean μ and standard deviation σ:

□ The **mean**, or center of the sampling distribution of $\bar{x}$, is equal to the population mean μ: $\mu_{\bar{x}} = \mu$.

□ The **standard deviation** of the sampling distribution is $\sigma/\sqrt{n}$, where n is the sample size : $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.



Application

Hypokalemia is diagnosed when blood potassium levels are below 3.5mEq/dl. Let's assume that we know a patient whose measured potassium levels vary daily according to a normal distribution $N(\mu = 3.8, \sigma = 0.2)$.

If only one measurement is made, what is the probability that this patient will be misdiagnosed with Hypokalemia?

$$z = \frac{(x-\mu)}{\sigma} = \frac{3.5-3.8}{0.2} = -1.5\ , \quad P(z < -1.5) = 0.0668 \approx 7\%$$

Instead, if measurements are taken on 4 separate days, what is the probability of a misdiagnosis?

$$z = \frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}} = \frac{3.5-3.8}{0.2/\sqrt{4}} = -3\ , \quad P(z < -3) = 0.0013 \approx 0.1\%$$

R overview and Installation

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.

R is free software distributed under a GNU-style copy left, and an official part of the GNU project called **GNUs**.

Features of R

R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.

R has an effective data handling and storage facility,

R provides a suite of operators for calculations on arrays, lists, vectors and matrices.

R provides a large, coherent and integrated collection of tools for data analysis.

R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

To Install R:

1. Open an internet browser and go to www.r-project.org.
2. Click the "download R" link in the middle of the page under "Getting Started."
3. Select a CRAN location (a mirror site) and click the corresponding link.
4. Click on the "Download R for Windows" link at the top of the page.
5. Click on the "install R for the first time" link at the top of the page.
6. Click "Download R for Windows" and save the executable file somewhere on computer. Run the .exe file and follow the installation instructions.
7. Now that R is installed, next step is to download and install RStudio.

To Install RStudio

1. Go to www.rstudio.com and click on the "Download RStudio" button.
2. Click on "Download RStudio Desktop."
3. Click on the version recommended for your system, or the latest Windows version, and save the executable file. Run the .exe file and follow the installation instructions.

R Command Prompt

Once R environment setup is done, then it’s easy to start R command prompt by just typing the following command at command prompt – “$ R”

This will launch R interpreter and will get a prompt > where we can start typing your program as follows −

```
> myString <- "Hello, World!"
> print ( myString)

[1] "Hello, World!"
```

Here first statement defines a string variable myString, where we assign a string "Hello, World!" and then next statement print() is being used to print the value stored in variable myString.

R Script File

execute scripts at command prompt with the help of R interpreter called **Rscript**.

```
# My first program in R Programming
myString <- "Hello, World!"
print ( myString)
```

Save the above code in a file test.R and execute it at command prompt as given below.

```
$ Rscript test.R
```

When we run the above program, it produces the following result.

```
"Hello, World!"
```

Comments

Comments are like helping text in your R program and they are ignored by the interpreter while executing actual program. Single comment is written using # in the beginning of the statement as follows −

```
# My first program in R Programming
```

R does not support multi-line comments but they can be written as follows:

```
  "This is a demo for multi-line comments and it should be put inside either a
     single OR double quote"

myString <- "Hello, World!"
print ( myString)
```

Result for above code is:

```
"Hello, World!"
```

R data types:

The variables are assigned with R-Objects and the data type of the R-object becomes the data type of the variable. There are many types of R-objects. The frequently used ones are −

- Vectors
- Lists
- Matrices
- Arrays
- Factors
- Data Frames

The simplest of these objects is the **vector object** and there are six data types of these atomic vectors, also termed as six classes of vectors. The other R-Objects are built upon the atomic vectors.

Data Type	Example	Verify

Logical	TRUE, FALSE	v <- TRUE
		print(class(v))
		[1] "logical"

Numeric	12.3, 5, 999	v <- 23.5
		print(class(v))
		[1] "numeric"

Integer	2L, 34L, 0L	v <- 2L

		print(class(v)) [1] "integer"

Complex	3 + 2i	v <- 2+5i
		print(class(v))
		[1] "complex"

Character	'a' , '"good", "TRUE", '23.4'	v <- "TRUE"
		print(class(v))
		[1] "character"

Raw	"Hello" is stored as 48 65 6c 6c 6f	v<-charToRaw("Hello")
		print(class(v))
		[1] "raw"

In R programming, the very basic data types are the R-objects called **vectors** which hold elements of different classes as shown above.

Vectors

When you want to create vector with more than one element, you should use **c()** function which means to combine the elements into a vector.

```
# Create a vector.
apple <- c('red','green',"yellow")
print(apple)
# Get the class of the vector.
print(class(apple))
```

When we execute the above code, it produces the following result −

```
"red"    "green" "yellow"
"character"
```

Lists

A list is an R-object which can contain many different types of elements inside it like vectors, functions and even another list inside it.

```
# Create a list.
list1 <- list(c(2,5,3),21.3,sin)
# Print the list.
print(list1)
```

When we execute the above code, it produces the following result −

```
[[1]]
[1] 2 5 3
[[2]]
[1] 21.3
 [[3]]
```

```
function (x) .Primitive("sin")
```

Matrices

A matrix is a two-dimensional rectangular data set. It can be created using a vector input to the matrix function.

```
# Create a matrix.
M = matrix( c('a','a','b','c','b','a'), nrow = 2, ncol = 3, byrow = TRUE)
print(M)
```

When we execute the above code, it produces the following result −

```
     [,1] [,2] [,3]
[1,] "a"  "a"  "b"
[2,] "c"  "b"  "a"
```

Arrays

While matrices are confined to two dimensions, arrays can be of any number of dimensions. The array function takes a dim attribute which creates the required number of dimension. In the below example we create an array with two elements which are 3x3 matrices each.

```
# Create an array.
a <- array(c('green','yellow'),dim = c(3,3,2))
print(a)
```

When we execute the above code, it produces the following result −

```
, , 1
     [,1]     [,2]     [,3]
[1,] "green" "yellow" "green"
[2,] "yellow" "green"  "yellow"
[3,] "green" "yellow" "green"
, , 2
     [,1]     [,2]     [,3]
[1,] "yellow" "green"  "yellow"
[2,] "green" "yellow" "green"
[3,] "yellow" "green"  "yellow"
```

Data Frames

Data frames are tabular data objects. Unlike a matrix in data frame each column can contain different modes of data. The first column can be numeric while the second column can be character and third column can be logical. It is a list of vectors of equal length. Data Frames are created using the **data.frame()** function.

```
# Create the data frame.
BMI <- data.frame(    gender = c("Male", "Male","Female"),    height = c(152, 171.5, 165),
   weight = c(81,93, 78),    Age = c(42,38,26) )
print(BMI)
```

Result −

```
  gender height weight Age
1   Male  152.0     81  42
2   Male  171.5     93  38
3 Female  165.0     78  26
```

R - Variables

A variable provides us with named storage that our programs can manipulate. A variable in R can store an atomic vector, group of atomic vectors or a combination of many R objects. A valid variable name consists of letters, numbers and the dot or underline characters. The variable name starts with a letter or the dot not followed by a number.

Variable Name	**Validity**	**Reason**
var_name2.	valid	Has letters, numbers, dot and underscore
var_name%	Invalid	Has the character '%'. Only dot(.) and underscore allowed.
2var_name	invalid	Starts with a number
.var_name, var.name	valid	Can start with a dot(.) but the dot(.)should not be followed by a number.
.2var_name	invalid	The starting dot is followed by a number making it invalid.
_var_name	invalid	Starts with _ which is not valid

R - Operators

An operator is a symbol that tells the compiler to perform specific mathematical or logical manipulations. R language is rich in built-in operators and provides following types of operators.

Types of Operators

types of operators in R programming −

Arithmetic Operators

Relational Operators

Logical Operators

Assignment Operators

Miscellaneous Operators

Descriptive Data analysis using R:

R provides a wide range of functions for obtaining summary statistics. One method of obtaining descriptive statistics is to use the **sapply( )** function with a specified summary statistic.

```
sapply(mydata, mean, na.rm=TRUE)
```

Possible functions used in sapply include **mean, sd, var, min, max, median, range, and quantile**.

Check your data

You can inspect your data using the functions **head**() and **tails**(), which will display the first and the last part of the data, respectively.

```
# Print the first 6 rows
head(my_data, 6)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1     5.1            3.5     1.4         0.2      setosa
2     4.9            3.0     1.4         0.2      setosa
3     4.7            3.2     1.3         0.2      setosa
4     4.6            3.1     1.5         0.2      setosa
5     5.0            3.6     1.4         0.2      setosa
6     5.4            3.9     1.7         0.4      setosa
```

R functions for computing descriptive statistics

Some R functions for computing descriptive statistics:

Description	R function
Mean	**mean**()
Standard deviation	**sd**()
Variance	**var**()
Minimum	**min**()
Maximum	**maximum**()
Median	**median**()
Range of values (minimum and maximum)	**range**()
Sample quantiles	**quantile**()
Generic function	**summary**()
Interquartile range	**IQR**()

Descriptive statistics for a single group

Measure of central tendency: mean, median, mode

Roughly speaking, the central tendency measures the “average” or the “middle” of your data. The most commonly used measures include:

the mean: the average value. It’s sensitive to outliers.
the median: the middle value. It’s a robust alternative to mean.
and the mode: the most frequent value

In R,

The function **mean**() and **median**() can be used to compute the mean and the median, respectively;
The function **mfv**() [in the **modeest** R package] can be used to compute the mode of a variable.

The R code below computes the mean, median and the mode of the variable Sepal.Length [in my_data data set]:

```
# Compute the mean value
mean(my_data$Sepal.Length)
[1] 5.843333

# Compute the median value
median(my_data$Sepal.Length)
[1] 5.8

# Compute the mode
# install.packages("modeest")
require(modeest)
mfv(my_data$Sepal.Length)
[1] 5
```

Measure of variability

Measures of variability gives how “spread out” the data are.

Range: minimum & maximum

Range corresponds to biggest value minus the smallest value. It gives you the full spread of the data.

```
# Compute the minimum value
min(my_data$Sepal.Length)
[1] 4.3
# Compute the maximum value
max(my_data$Sepal.Length)
[1] 7.9
# Range
range(my_data$Sepal.Length)
[1] 4.3 7.9
```

Interquartile range

The **interquartile range** (IQR) - corresponding to the difference between the first and third quartiles - is sometimes used as a robust alternative to the standard deviation.

R function:

```
quantile(x, probs = seq(0, 1, 0.25))
```

x: numeric vector whose sample quantiles are wanted.
probs: numeric vector of probabilities with values in [0,1].

Example:

```
quantile(my_data$Sepal.Length)
  0% 25% 50% 75% 100%
 4.3 5.1 5.8 6.4  7.9
```

To compute deciles (0.1, 0.2, 0.3, …., 0.9), use this:

```
quantile(my_data$Sepal.Length, seq(0, 1, 0.1))
```

To compute the interquartile range, type this:

```
IQR(my_data$Sepal.Length)
[1] 1.3
```

Variance and standard deviation

The variance represents the average squared deviation from the mean. The standard deviation is the square root of the variance. It measures the average deviation of the values, in the data, from the mean value.

```
# Compute the variance
var(my_data$Sepal.Length)

# Compute the standard deviation =
# square root of th variance
sd(my_data$Sepal.Length)
```

Computing an overall summary of a variable and an entire data frame summary() function

Summary of a single variable. Five values are returned: the mean, median, 25th and 75th quartiles, min and max in one single line call:

```
summary(my_data$Sepal.Length)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300 5.100 5.800 5.843 6.400  7.900
```

Summary of a data frame. In this case, the function **summary**() is automatically applied to each column. The format of the result depends on the type of the data contained in the column. For example:

- If the column is a numeric variable, mean, median, min, max and quartiles are returned.
- If the column is a factor variable, the number of observations in each group is returned.

```
summary(my_data, digits = 1)
  Sepal.Length  Sepal.Width  Petal.Length    Petal.Width    Species
```

```
Min. :4 Min. :2 Min. :1 Min.  :0.1 setosa :50
1st  Qu.:5    1st  Qu.:3    1st  Qu.:2      1st  Qu.:0.3      versicolor:50
Median :6 Median :3 Median :4 Median :1.3 virginica :50
Mean   :6   Mean   :3   Mean   :4   Mean   :1.2
3rd  Qu.:6  3rd  Qu.:3    3rd  Qu.:5    3rd  Qu.:1.8
Max. :8 Max. :4 Max. :7 Max. :2.5
```

sapply() function

```
# Compute the mean of each column
sapply(my_data[, -5], mean)
Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
    5.843333     3.057333     3.758000     1.199333
#          Compute          quartiles
sapply(my_data[, -5], quantile)
     Sepal.Length Sepal.Width Petal.Length Petal.Width
0%           4.3         2.0         1.00         0.1
25%           5.1         2.8         1.60    0.3
50%           5.8         3.0         4.35    1.3
75%           6.4         3.3         5.10    1.8
100%          7.9         4.4         6.90    2.5
```

Descriptive Data Analysis using R > Description of Basic Functions used to Describe Data in R

builtins()	# List all built-in functions
help() or ? or ??	#i.e. help(boxplot)
getwd() and setwd()	# working with a file directory
q()	#To close R
ls()	#Lists all user defined objects.
rm()	#Removes objects from an environment.
demo()	#Lists the demonstrations in the packages that are loaded.
demo(package = .packages(all.available = TRUE))	#Lists the demonstrations in all installed packages.
?NA	# Help page on handling of missing data values
abs(x)	# The absolute value of "x"
append()	# Add elements to a vector
cat(x)	# Prints the arguments
cbind()	# Combine vectors by row/column (cf. "paste" in Unix)
grep()	# Pattern matching
identical()	# Test if 2 objects are *exactly* equal
length(x)	# Return no. of elements in vector x
ls()	# List objects in current environment
mat.or.vec()	# Create a matrix or vector
paste(x)	# Concatenate vectors after converting to character
range(x)	# Returns the minimum and maximum of x
rep(1,5)	# Repeat the number 1 five times

rev(x)	# List the elements of "x" in reverse order
seq(1,10,0.4)	# Generate a sequence (1 -> 10, spaced by 0.4)
sequence()	# Create a vector of sequences
sign(x)	# Returns the signs of the elements of x
sort(x)	# Sort the vector x
order(x)	# list sorted element numbers of x
tolower(),toupper()	# Convert string to lower/upper case letters
unique(x)	# Remove duplicate entries from vector
vector()	# Produces a vector of given length and mode
formatC(x)	# Format x using 'C' style formatting specifications
floor(x), ceiling(x), round(x), signif(x), trunc(x)	# rounding functions
Sys.time()	# Return system time
Sys.Date()	# Return system date
getwd()	# Return working directory
setwd()	# Set working directory

Inferential statistics using R

Simple linear regression analysis

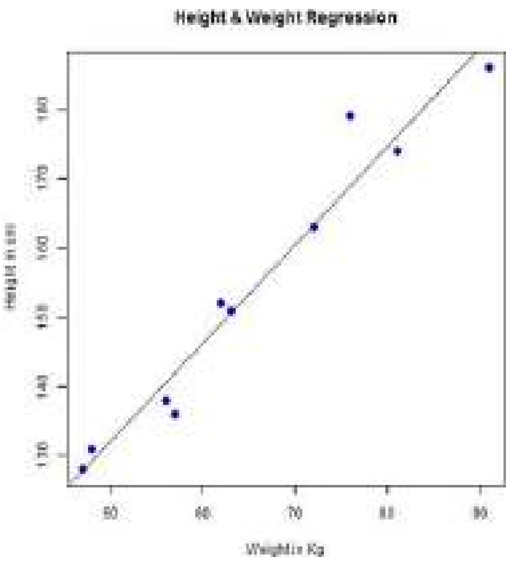
- Regression analysis is a very widely used statistical tool to establish a relationship model between two variables
- One of these variable is called predictor variable
- The other variable is called response variable
- The general mathematical equation for a linear regression is y = mx + b

	Register_no	Name	Dept	CGPA	Height	Weight
1	18N312001	JOHN	IT	8.5	151	63
2	18N312005	SIM	CSE	9.2	174	81
3	18N312011	TIM	IT	9.5	138	56
4	18N312061	LILLY	IT	9.34	186	91
5	18N312099	CARL	MECH	8.12	128	47

- lm() Function
- This function creates the relationship model between the predictor and the response variable.
- The basic syntax for lm() function in linear regression is −
- lm(formula,data)
- # Apply the lm() function.
- relation <- lm(stud.data$weight ~ stud.data$height)
- print(relation)

Output

```
Coefficients:
(Intercept (m))          x
       -38.4551        0.6746
```


Height & Weight Regression
Height in cm
Weight in Kg