

Introduction to R



R S Rajput

Assistant Professor Computer Science

Department of Mathematics, Statistics, Computer Science

G.B. Pant University of Agriculture & Technology

Pantnagar, Uttarakhand, INDIA



Introduction to R

- R is a powerful software environment for **data management** and **data analysis**.
- R is an **Open source Software**.
- Mostly R tools are working satisfactory on the low configured computer i.e. no advanced computing resources are required for all tools.
- Huge availability of references and Help.
- Well recognized in scientific computing.

Resource of R

- Installation set up software, Installation guideline, software manual and help available following website.

<https://www.r-project.org/>

Download installation setup file

(e.g. *.exe for windows based OS)

Simple to install: step-by-step wizard

Less installation time: Approx. 10-15 minutes required

Not required for a expertise person for installation

Web Resource for Learning R

- https://www.tutorialspoint.com/r/r_overview.htm

The screenshot displays the Tutorialspoint website interface. At the top, the browser address bar shows the URL https://www.tutorialspoint.com/r/r_overview.htm. Below the browser, the Tutorialspoint logo is visible on the left, and a navigation menu on the right includes links for Jobs, Examples, Whiteboard, Net Meeting, and a search icon. A secondary navigation bar contains links for HOME, Q/A, LIBRARY, VIDEOS, and TUTORS. The main content area is titled "R - Overview". Below the title is an "Advertisements" section featuring a banner for an "Artificial Intelligence Course". On the left side of the page, there is a sidebar with the R logo and the text "LEARN R PROGRAMMING programming language". Below this, a section titled "R Tutorial" lists three links: "R - Home", "R - Overview" (which is highlighted), and "R - Environment Setup". The main text area on the right begins with the sentence: "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 text is partially cut off at the bottom.

Web Resource for Learning R Cont..

- <http://www.r-tutor.com/r-introduction>

www.r-tutor.com/r-introduction

News Gmail

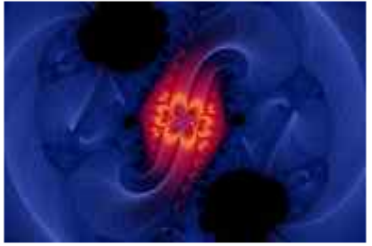
About | Contact | Resources | Terms of Use

R Tutorial

An R Introduction to Statistics

HOME DOWNLOAD SALES EBOOK SITE MAP

R Introduction




We offer here a couple of introductory tutorials on basic R concepts. It serves as background material for our main tutorial series *Elementary Statistics with R*.

The only hardware requirement for most of the R tutorials is a PC with the latest free open source R software installed. R has extensive documentation and active online community support. It is the perfect environment to get started in statistical computing.

Search this site:

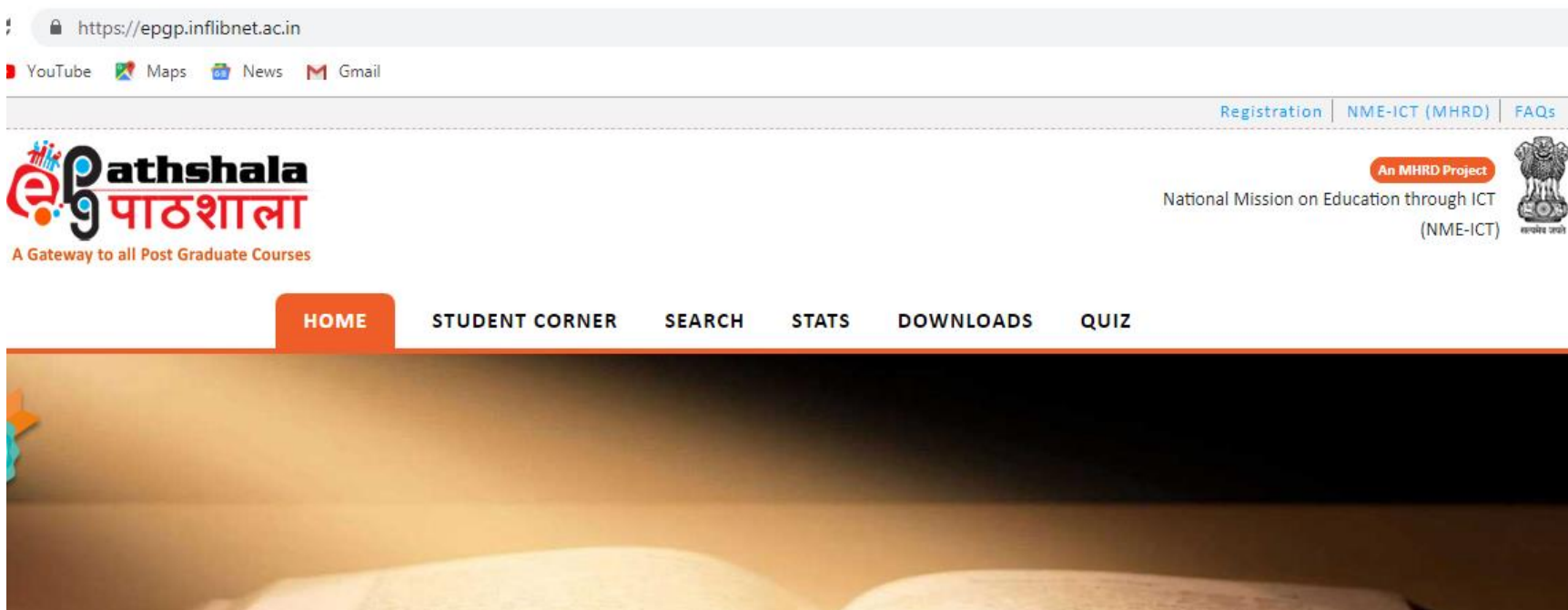
Search

R Tutorial eBook



Web Resource for Learning R Cont..

- UGC (MHRD) s' e-PG Pathshala
- <http://www.ugc.ac.in> (go to e-PG Pathshala Link)
- <http://epgp.inflibnet.ac.in/>



Starting R



- Start R by double clicking on R icon.
- A windows displayed, called R console with prompt “>”.
- The > is called the R prompt.
- It is used to indicate where you are to type.

```
R : Copyright 2001, The R Development Core Team
Version 1.4.0 (2001-12-19)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

>
```

Key Elements of R

- Data
- Vector
- Data Frame
- Operators
 - Mathematical, Relational, Logical, Assignment
- Function
 - Built in (Library) function, User defined function
- Statements
 - Assignment, Conditional, Looping, Data Import & Export

Data Management in R

- Data->Vector-> Data Frame

E_No	Name	Marks
1001	XXX	90.50
1002	YYY	80.25
1003	ZZZZ	95.00
1004	AAAA	87.50
1005	BBB	98.75

The diagram illustrates the relationship between Data, Vector, and Data Frame in R. The table shows five rows of data. A bracket on the right side of the table, spanning all rows, is labeled 'Vector', indicating that the 'Marks' column is a vector. Two callout boxes labeled 'Data' point to the 'E_No' and 'Name' columns, indicating that each column is a data vector. A large bracket at the bottom of the table is labeled 'Data Frame', indicating that the entire table structure is a data frame.

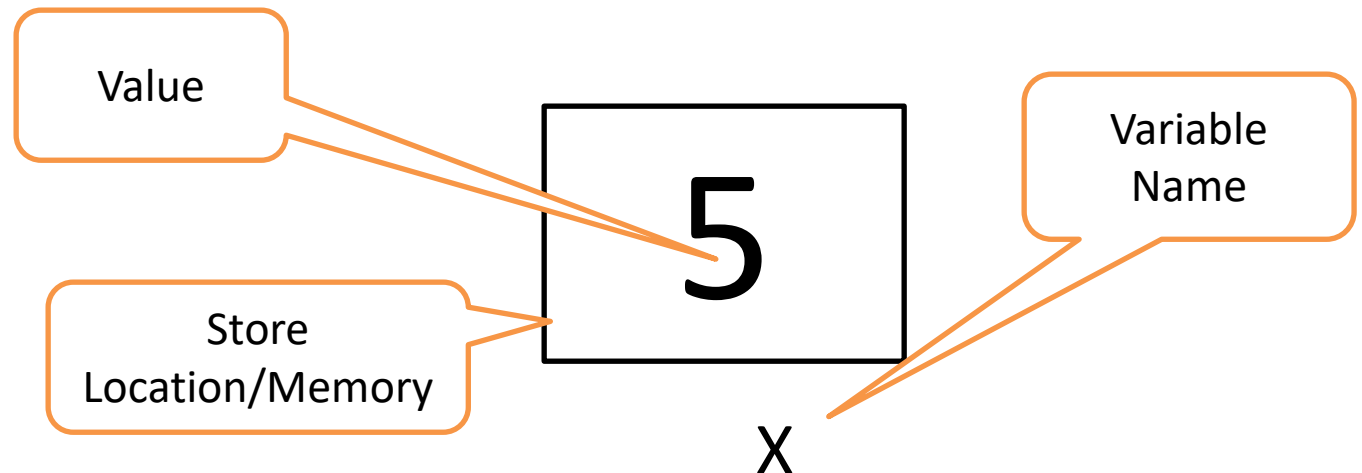
Variable/ Variable Naming in R

- A variable name can contain letters (a-z, A-Z), digits (0-9), dot (.) and under score (_)
- A variable name must be start with a letter or a dot.
- Example: if we want to store a value '5' into a variable X than R Code is.

>X<-5

OR

>X=5



Data types in R

- Numeric
- Integer
- Complex
- Logical
- Character

Basic Data Type

- Numeric Data
- Default Data type
- Decimal values are known as **numeric** in R
- Integer
- Numbers without decimal

```
>x=10.5
```

```
>x
```

```
[1] 10.5
```

```
>class(x)
```

```
[1]"numeric"
```

```
>y=as.integer(3)
```

```
>y
```

```
[1] 3
```

```
>class(y)
```

```
[1]"integer"
```

Basic Data Type Cont..

- Complex Data
- A Complex value in R is defined via the pure imaginary value i .
- Logical Data
- A logical value is created via comparison between variables

$$i = \sqrt{-1}$$

```
>z=1+2i
```

```
>z
```

```
[1]1+2i
```

```
>class (z)
```

```
[1]"complex"
```

```
>x=1; y=2
```

```
>z=x>y
```

```
>z
```

```
[1]FALSE
```

```
>class (z)
```

```
[1]"logical"
```

Basic Data Type Cont..

- Character Data
- A character object is used to represent string values in R.
- `class()` function used to print class of data type.
- `#` is used to comment.

```
>x="Technology"
```

```
>class(x)
```

```
[1]"character"
```

Vector

- A Vector is a container for data elements of the same basic data type. OR collection of similar data type.
- The container is known as **Vector**, and elements are known as components in vector.
- Generally a function 'c' is used to create vector.
- **'c' function**
`>x=c (5, 6, 7, 3, 5)`
- Here 'x' is a vector and 5, 6, 7, 3, 5 are elements, x is numeric vector
- **Some vector functions**
`>x, length(x), >x[5],>x[-2],>x[2:4]`

Data Frame

- A data frame is used for storing data in as a tables.
- It is a list of related vectors having equal length.
- For example, the following variable df is a data frame containing three vectors n, s, b.

```
>n = c(2,3,5)
```

```
>s = c("aa","bb","cc")
```

```
>b = c(TRUE,FALSE,TRUE)
```

```
>df = data.frame(n,s,b)
```

```
>df
```

- Data frame define: Using function **data.frame**
- Inbuilt dataframe: mtcars, Orange, trees

Mathematical Operators

- (+) Addition
- (-) Subtraction
- (*) Multiplication
- (/) Division
- (**) or ^) Power
- (%%) Reminder

Relational & Logical Operators

- $(==)$ Equal
- $(!=)$ Not equal
- $(<)$ Less than
- $(<=)$ Less than equal to
- $(>)$ Greater than
- $(>=)$ Greater than equal to
- $(!)$ Logical NOT
- $(&)$ Logical OR
- $(|)$ Logical AND

Assignment Operator

- (= or <-)

Example(s)

```
> X=10
```

```
>V1=c (1,2,3,4,5,6,7,8,9)
```

Date Import and Export from MS Excel

- Set working directory
`>setwd (" C:\data")`
- Data import from Excel file (as *.csv file format)
`>data =read.csv("input.csv")`
- Data Export to Excel file
`>data = write.csv("output.csv")`

Variate

- The quantified variable is known as variate.
- Quantification is the process of assigning numeral values to variable.
- The Statistical analysis involves variate analysis: Uni-variate, Bi-variate, Multi-variate analysis.

Variate Examples

Univariate

Travel Time (minutes):

15, 29, 8, 42, 35, 21, 18, 42, 26

Bivariate

Ice Cream Sales vs Temperature

Temperature °C

Ice Cream Sales

14.2°	\$215
16.4°	\$325
11.9°	\$185
15.2°	\$332
18.5°	\$406
22.1°	\$522
19.4°	\$412
25.1°	\$614
23.4°	\$544
18.1°	\$421
22.6°	\$445
17.2°	\$408

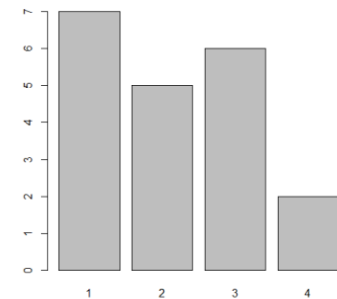
Data Visualization

- **Uni-variate and Qualitative** (*ordinal*)
 - Pie diagram, Bar plot
- **Uni-variate and Quantitative**
 - Box plot, Histogram
- **Bi-variate (Qualitative** (*ordinal*), **Qualitative** (*ordinal*))
 - Mosaic Plot
- **Bi-variate (Qualitative** (*ordinal*), **Quantitative**)
 - Plot of Boxes, Bar plot
- **Bi-variate (Quantitative, Quantitative)**
 - Scatter plot, Line diagram

Visualization Univariate Qualitative data

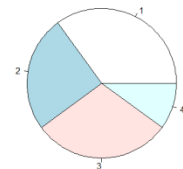
Bar Chart: Bar chart is a display frequency of a qualitative variable.

```
> grade=c(1,3,3,1,1,2,2,2,1,2,3,3,3,4,4,1,2,3,1,1)
> grade.freq=table(grade) # Frequency Vector
> grade.freq
> barplot(grade.freq)
```



Pie chart: Pie chart is a display percent frequency of a qualitative variable.

```
> pie(grade.freq)
```

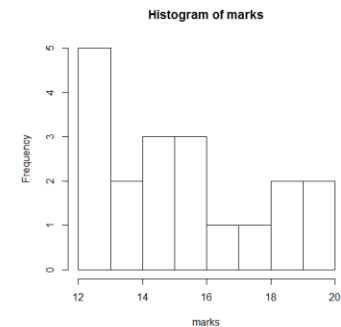
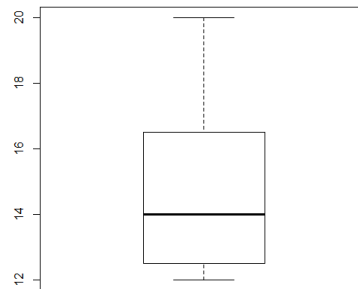


Visualization Univariate Quantitative data

Box Plot: Box plot is used to summarize the distribution of a numeric variable.

```
>marks=c(12,12,14,15,13,14,16,17,12,15,18,18,19,20,12,13,12,14,15)
```

```
>boxplot(marks)
```

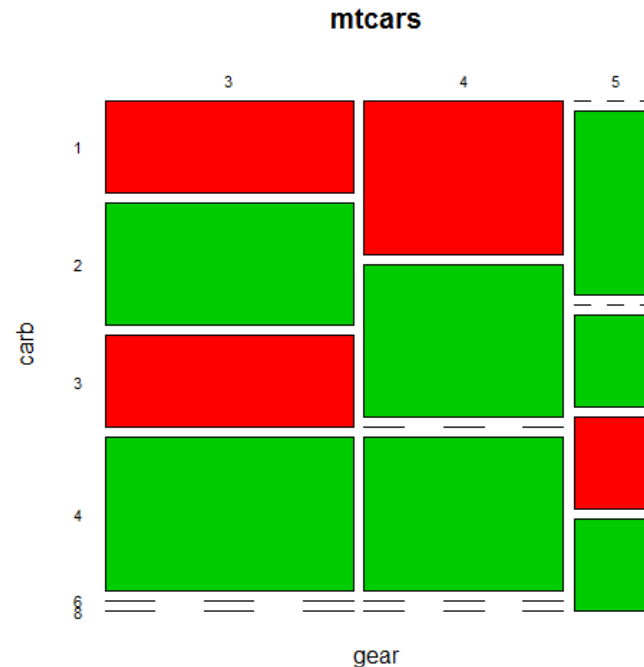


Histograms: Histogram is used to summarize the distribution of numeric variable.

```
>hist(marks, right=FALSE)
```

Visualization Bivariate data (Qualitative (*ordinal*), Qualitative (*ordinal*))

- `> mosaicplot(~ gear + carb, data = mtcars, color = 2:3, las = 1)`

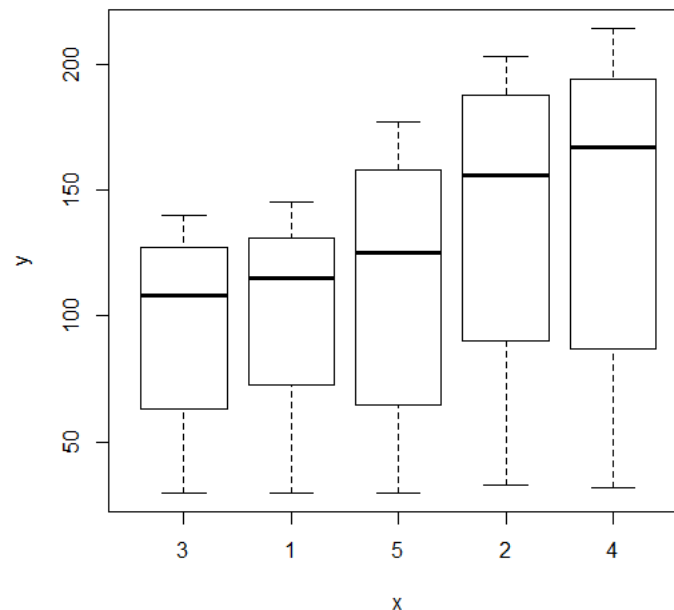


Visualization Bivariate data (Qualitative (*ordinal*), Quantitative)

Box Plot

```
>plot(Orange$Tree,Orange$circumference)
```

Tree is ordinal and, circumference is quantitative

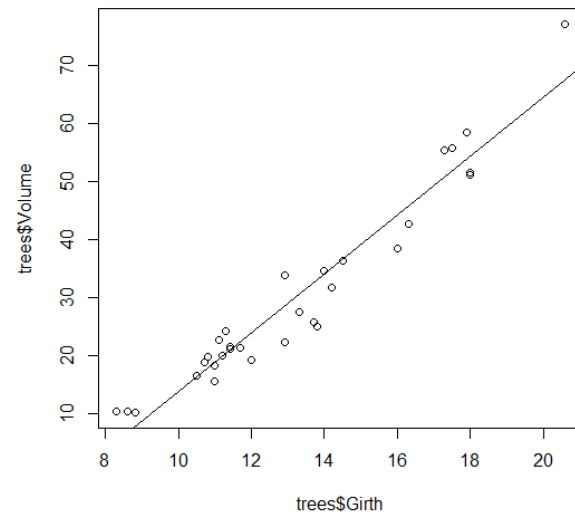
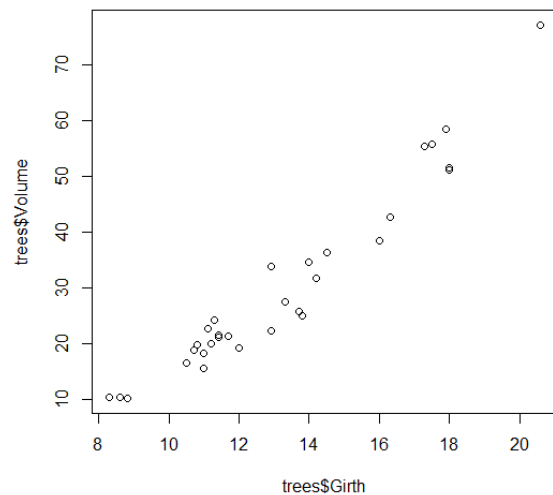


Visualization Bivariate data (Quantitative, Quantitative)

Scatter Plot (Girth and Volume both quantitative)

```
>plot(trees$Girth, trees$Volume)
```

```
>abline(lm(trees$Volume~trees$Girth))
```



Some Statistical Measures(Uni-variate)

- Mean
- Median
- Quartile
- Range
- Interquartile Range
- Variance
- Standard Deviation

Some Statistical Measures Cont..

Mean: The mean of an observation variable is a numerical measure of the central location of the data values. It is the sum of its data values divided by data count.

```
>marks=c(12,12,14,15,13,14,16,17,12,15,18,18,19,20,12,13,12,14,15)  
>mean(marks)
```

Median: The median of an observation variable is the value at the middle when the data is sorted in ascending order. It is an ordinal measure of the central location of the data values.

```
>median(marks)
```

Some Statistical Measures Cont..

Quartile: There are several quartiles of an observation variable. The first quartile, or lower quartile, is the value that cuts off the first 25% of the data when it is sorted in ascending order. The second quartile, or median, is the value that cuts off the first 50%. The third quartile, or upper quartile, is the value that cuts off the first 75%.

`>quantile(marks)`

Range: The range of an observation variable is the difference of its largest and smallest data values. It is a measure of how far apart the entire data spreads in value.

`>range(marks)`

Some Statistical Measures Cont..

Inter-quartile range: The inter-quartile range of an observation variable is the difference of its upper and lower quartiles. It is a measure of how far apart the middle portion of data spreads in value

`>IQR(marks)`

Variance: The variance is a numerical measure of how the data values is dispersed around the mean.

`>var(marks)`

Standard Deviation: The standard deviation of an observation variable is the square root of its variance.

`>sd(marks)`

T TEST

- Parametric Test: works on normally distributed scale data.
- Compares Two means
- There are different version for different design
 - One Sample i.e. One sample t-test
 - Dependent (related) Sample i.e. Paired t-test
 - Independent (unrelated) Sample i.e. Two sample t-test

HYPOTHESIS

- Null hypothesis: The sample come from the same population H_0
- Alternative Hypothesis: The samples come from different populations H_1

t.test() Function

General Form

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),  
mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

Example

```
t.test (x, alt="less", mu=10)
```

```
t.test (x, y, alt="two.sided")
```

```
t.test(x, y, alt="less", paired = T)
```

One Sample t-test

A sample of 24 people is taken. The length of time to prepare dinner is recorded in minutes, as given below

44.0 51.9 49.7 40.0. 55.5 43.4 41.3 45.2 40.7
41.1 49.1 30.9 45.2 55.3 52.1 55.1 38.8 43.1
39.2 58.6 49.8 43.2 47.9 46.6

Is there any evidence that the population mean time to prepare dinner is less than 48 minutes?
Use a level of significance of 0.05.

One Sample t-test

Step 1: One sample, Normally distributed ,one sample t test

Step 2

Null Hypothesis $H_0: \mu = 48$

Alternative Hypothesis $H_1: \mu < 48$

Step 3

`>p=c (44.0, 51.947.9, 46.6)`

`>t.test(p, alt="less", mu=48)`

One Sample t-test

data: p

$t = -1.7044$, $df = 24$,

alternative hypothesis: true mean is less than 48

95 percent confidence interval:

$-\text{Inf}$ 48.00909

sample estimates:

mean of x

45.628

INTERPRETATION ONE SAMPLE T TEST

The p-value is a number between 0 and 1 and interpreted in the following way

Right tailed or left tailed (95% confidence interval)

- A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis. (H_1 v)
- A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis. (H_0 v)

Two Tailed (95% confidence interval)

- A small p-value (typically ≤ 0.025) indicates strong evidence against the null hypothesis, so you reject the null hypothesis. (H_1 v)
- A large p-value (> 0.025) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis. (H_0 v)

Since P value 0.0506 is greater than 0.05 so H_0 is accepted

PAIRED T-TEST

Compare the means of two conditions in which the same (or closely matched) participants participated.

PAIRED T-TEST

A study was performed to test whether cars get better mileage on premium gas than on regular gas. Each of 10 cars was first filled with either regular or premium gas, decided by a coin toss, and the mileage for that tank was recorded. The mileage was recorded again for the same cars using the other kind of gasoline. Determine whether cars get significantly better mileage with premium gas.

Mileage

Regular 16,20,21,22,23,22,27,25,27,28

Premium 19,22,24,24,25,25,26,26,28,32

PAIRED T-TEST

Step 1 Data Normally distributed, Paired

Step 2

H0 : Difference of mean mileage of premium gas with regular gas is equal

H1 : Difference of mean mileage of premium gas with regular gas is greater

Step 3

```
> reg=c(16,20,21,22,23,22,27,25,27,28)
```

```
> pre=c(19,22,24,24,25,25,26,26,28,32)
```

```
> t.test (pre, reg, alt="greater", paired=T)
```

Paired t-test

data: pre and reg

t = 4.4721, df = 9, p-value = 0.0007749

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

1.180207 Inf

sample estimates:

mean of the differences

2

Two Sample t-test

- Compare the means of two groups of participants
- Groups are independent

Two Sample t-test

6 subjects were given a drug (treatment group) and an additional 6 subjects a placebo (control group). Their reaction time to a stimulus was measured (in ms).

Control : 91, 87, 99, 77, 88, 91

Treat : 101, 110, 103, 93, 99, 104

To perform a two sample t-test for comparing the means of the treatment and control groups.

Two Sample t-test

```
>Control = c(91, 87, 99, 77, 88, 91)
>Treat = c(101, 110, 103, 93, 99, 104)
> t.test(Control,Treat,alternative="less", var.equal=TRUE)
```

Two Sample t-test

data: Control and Treat

$t = -3.4456$, $df = 10$, $p\text{-value} = 0.003136$

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

$-\text{Inf}$ -6.082744

sample estimates:

mean of x mean of y

88.83333 101.66667

Since p value is 0.003136 is less than 0.05 so H_1 is accepted.

CHI-SQUARE TESTS

Goodness of fit

A goodness of fit test, checks to see if the data came from some specified population.

Test for Independence

Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of another.

CHI SQUARE GOODNESS OF FIT

Step 1

Prepare frequency vector

Prepare corresponding Probability vector

Step 2

H_0 : Fit is good

H_1 : Fit is not good

Step 3

Use R Function

`>chisq.test(frequency vector, p=probability vector)`

Step 4

Give result based on p-value

CHI SQUARE GOODNESS OF FIT

If we toss a die 150 times and, find that we have the following distribution of rolls is the die fair?

Face	1	2	3	4	5	6
Number of rolls	22	21	22	27	22	36

Solution

H0: Die is fair, H1: Die is not fair

```
>freq=c(22,21,22,27,22,36)
```

```
>prob=c(1/6,1/6,1/6,1/6,1/6,1/6)
```

```
>chisq.test(freq,p=prob)
```

Chi-squared test for given probabilities

data: freq

X-squared = 6.72, df = 5, p-value = 0.2423

P-value is greater than 0.05 so that H0 accepted.

CHI SQUARE TEST FOR INDEPENDENCE

Step 1

- Prepare vectors

- Prepare data frame

Step 2

- H0: Attributes are independents

- H1: Attributes are not independents

Step 3

- Use R Function

- `>chisq.test(data_frame)`

Step 4

- Give result based on p-value

CHI SQUARE TEST FOR INDEPENDENCE

From a survey conducted in different regions (rural and urban)
To know preference of persons to different television programs
(educational and entertainment). Following data was obtained.

	Education	Entertainment
Rural	45	35
Urban	20	50

Test whether preference to a program depends on region.

CHI SQUARE TEST FOR INDEPENDENCE

Step 1

```
>rural =c(45,35)
>urban=c(20,50)
>survey1=data.frame(rural,urban)
```

Step 2

H0: Performance of program independent on region

H1: Performance of program dependent on region

Step3

Use R Function

```
>chisq.test(survey1)
Pearson's Chi-squared test
data: survey1
X-squared = 11.648, df = 1, p-value = 0.0006429
```

Step4

Based on P ($0.00064 < 0.05$) value H_0 is rejected and H_1 accepted, i.e. Performance of program dependent on region.

Correlation

Correlation is used to test for a relationship between two numerical variables or two ranked (ordinal) variables.

Usually, in statistics, we measure three types of correlations:

1. Pearson correlation
2. Kendall rank correlation
3. Spearman correlation

Pearson r correlation-Parametric

Kendall Rank and Spearman correlation –Non Parametric

Correlation

A simplified format is

`cor(x, method=)`

where

x: data frame

Method: Specifies the type of correlation.

Options are pearson (default), spearman or kendall.

Correlation

Exercise

Protein intake X and fat intake Y (in gm) for ten old women given as

X 56,47,33,39,42,38,46,47,38,32

Y 56,83,49,52,65,52,56,48,59,70

Calculate correlation Coefficient (Pearson) , draw scatter plot matrix and scatter plot

Exercise

Find correlation coefficient (Pearson) between the sales and expenses from the data given below:

Firm: 1,2,3, 4,5,6,7,8,9,10

Sales (Rs Lakhs) 50,50,55,60,65,65,65,60,60,50

Expenses (Rs Lakhs): 11,13,14,16,16,15,15,,14,13,13

Draw scatter plot matrix, and scatter plot

THANKS